ORIGINAL RESEARCH PAPER

Quality evaluation of ultrasound images for fetal crown–rump length measurement at 11 to 14 weeks' gestation: A fusion model can be interpreted using SHAP method

LU LIU[1], TING WANG[1], YANPING LI[1], HONGYAN TIAN[1], HAIDONG ZHANG[1], CHENYANG ZHOU[2], WENJING ZHU[3], WENJUN CAI[4]

[1]Department of Ultrasound Medicine, South China Hospital, Medical School, Shenzhen University, Shenzhen, P. R. China; [2]Department of Information, South China Hospital, Medical School, Shenzhen University, Shenzhen, P. R. China; [3]Medical Research Department, Qingdao Hospital, University of Health and Rehabilitation Sciences (Qingdao Municipal Hospital), Qingdao, P. R. China; [4]Department of Ultrasound, Shenzhen University General Hospital, Medical School, Shenzhen University, Shenzhen, P. R. China

ABSTRACT

Objectives: To explore a fusion model designed for the quality evaluation of ultrasound images utilized in fetal crown–rump length (CRL) measurement, and to use SHapley Additive exPlanations (SHAP) method to elucidate the model's decision-making processes.

Methods: We retrospectively collected 1149 images of midsagittal planes of the entire fetus during early pregnancy from two hospitals. Two senior radiologists categorized the images into standard and non-standard planes. Seven image segmentation models were trained to select the best model for automatically segmenting the region of interest. The radiomics features and deep transfer learning (DTL) features were extracted and selected to establish radiomics models and DTL models. We also constructed fusion models to enhance the classification performance and the optimal one underwent comparison with radiologists. The SHAP method was employed to interpret and visualize the model.

Received: 12 October 2025 | Accepted: 12 December 2025

Correspondence: Wenjun Cai / Academic Affiliation: Department of Ultrasound, Shenzhen University General Hospital, Medical School, Shenzhen University. / E-mail: 2977673731@qq.com
Wenjing Zhu / Academic Affiliation: Medical Research Department, Qingdao Hospital, University of Health and Rehabilitation Sciences (Qingdao Municipal Hospital), Qingdao, P. R. China. / E-mail: zhuwj@uor.edu.cn
Chenyang Zhou / Academic Affiliation: Department of Information, South China Hospital, Medical School, Shenzhen University. / E-mail: 1518712207@qq.com

Results: The DeepLabV3 ResNet101 segmentation model demonstrated the highest performance (DSC: 97.15%). The early fusion model exhibited superior classification performance in validation set (AUC: 0.947, 95% CI: 0.924-0.970, accuracy: 88.4%, sensitivity: 83.0%, specificity: 92.7%, PPV: 90.1%, NPV: 87.3%, precision: 90.1%). The model demonstrated performance commensurate with that of senior radiologists while surpassing junior radiologists. Notably, when leveraging the model's support, there was a substantial improvement in their overall performance.

Conclusions: The early fusion model demonstrated satisfactory performance in the intelligent quality evaluation of ultrasound images for CRL measurement. It has the potential to enhance the professional skills of junior radiologists.

Key words: Deep learning, Quality evaluation, Ultrasound standard plane, Crown–rump length, SHAP

Introduction

Crown–rump length (CRL) measurement is an essential part of the routine first-trimester ultrasound scan, conducted at 11 to 14 weeks' gestation [1]. The measurement of CRL is used to estimate gestational age (GA) [2]. As the date of conception cannot be precisely determined except in pregnancies resulting from assisted reproductive technology, most pregnancies are dated by the last menstrual period, although this may sometimes be uncertain or unreliable. Therefore, determining GA by ultrasound scan at the first-trimester of pregnancy, based on the measurement of CRL, appears to be the most reliable method. Accurate estimation of GA is crucial for assessing whether fetal size is appropriate-for-GA [3, 4]. The accurate measurement of CRL is based on a standard plane—a midline sagittal section of the entire fetus with optimum magnification [5]. To improve the consistency and reproducibility of measurement, the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG) has outlined criteria for the standard plane [1, 2]. However, due to variations in fetal position and the sonographer's skill, CRL measurements are sometimes taken on non-standard planes. To ensure the accuracy of GA estimation, rigorous quality control of the image for CRL measurement is essential. However, manual quality control increases the workload for radiologists and is challenging to implement in primary hospitals.

Recently, artificial intelligence (AI) technology has achieved significant advancements within the domain of obstetric ultrasound [6, 7]. It has been demonstrably established that deep learning, especially via convolutional neural network (CNN), ranks among the leading AI methodologies in the domain of image analysis [8]. To enhance clinical work efficiency and confront limited medical resources, this research introduced an innovative fusion model that ingeniously combined the strengths of radiomics and CNN for automatic quality evaluation of ultrasound images used in CRL measurement. The SHapley Additive exPlanations (SHAP) method was employed to interpret and visualize the decision-making process of this model [9, 10]. Moreover, an automated deep learning framework was designed for image segmentation to accurately identify the regions of interest (ROIs).

Methods

Our study was granted formal approval from the Institutional Review Board (IRB) of South China Hospital of Shenzhen University (No: HNLS20240326001-A). Given the retrospective design of this study, informed consent requirements were subsequently waived by the ethics committees. The process of model development is illustrated in Figure 1.

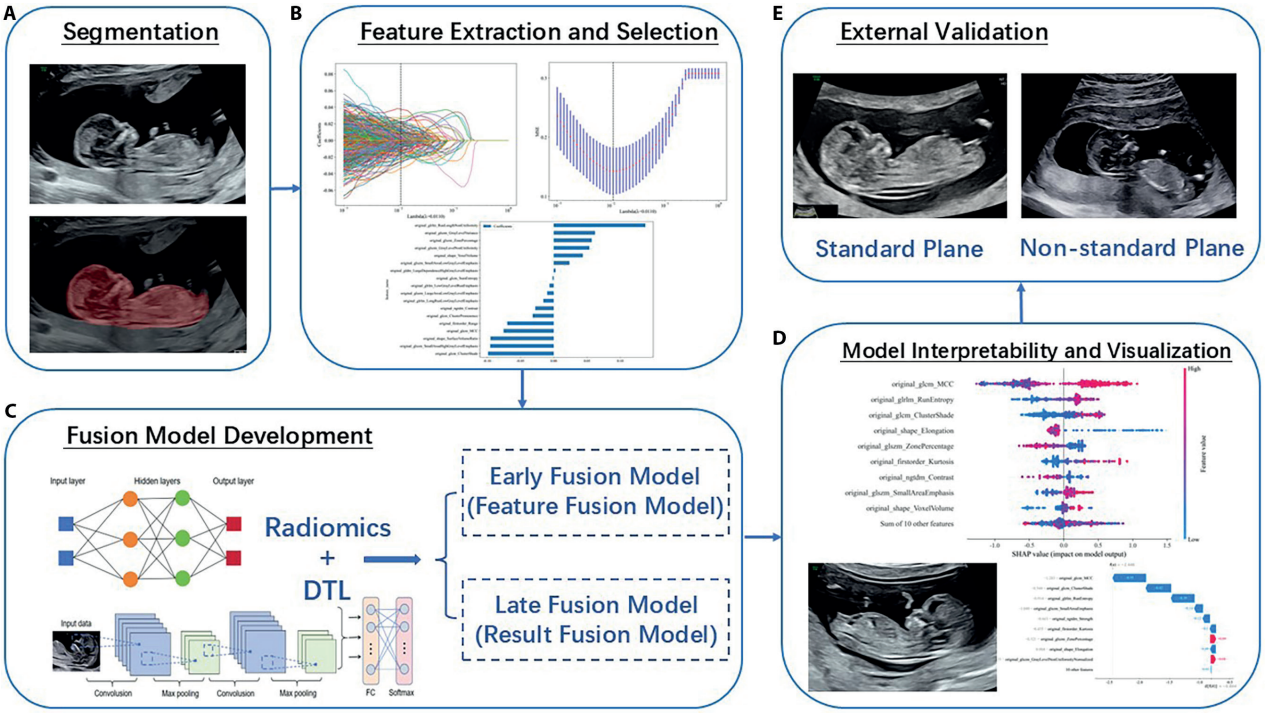

Figure 1. The workflow for developing the AI model.

Participants and images acquisition

Ultrasound images of fetal midsagittal planes, captured between 11 and 14 weeks of gestation, were collected from two separate hospitals, serving as the training set and validation set, respectively. Between January 2022 and October 2024, ultrasound examinations were carried out using a range of ultrasonographic devices, including the GE Voluson E8, GE Voluson E10, and Samsung HERA XW10. Following the specifications for CRL measurement outlined in the ISUOG practice guidelines [1, 2], two senior radiologists (H.Z. and L.L.) categorized each ultrasound image as a standard plane or non-standard plane. The radiologists' concordant classifications provided the reference standard for assessing model performance. A third senior radiologist (W.C.) was consulted to reach a consensus in instances where their classifications differed. The inclusion criteria: (1) Pregnant women at 11 to 14 weeks' gestation received ultrasound scans at the two designated hospitals; (2) CRL measurements ranging from 45 to 84 mm; (3) Two experienced radiologists (L.L. and H.T.) evaluated all images for quality assurance and selected images that met quality standards; (4) Only one image per fetus, the one that most closely adhered to the standard plane requirements, was included. The exclusion criteria: (1) Fetuses diagnosed with severe malformations (such as gastroschisis, open neural tube defects, etc.); (2) Images with measurement lines. The standard criteria of ultrasound image for CRL measurement [1, 2]:

1. A midsagittal section of the whole fetus should be obtained. The ultrasound examination enables visualization of multiple fetal anatomical structures: the facial profile, nasal bone, nuchal translucency, posterior brain, intracranial translucency, cardiac activity, spinal column, abdominal wall, and diaphragmatic structures.
2. The fetus was positioned horizontally on the ultrasound display, with the crown-rump measurement axis oriented perpendicular (90°) to the ultrasound beam.
3. For accurate measurement, the fetus must maintain a neutral position without flexion or hyperextension. This is confirmed by

visualizing a clear amniotic fluid space between the fetal chin and chest wall.
4. The ultrasound image should be appropriately magnified such that the fetal structures occupy approximately 75% of the screen width.
5. Both the end points of the crown and the rump must be clearly visualized for accurate measurement.

Image segmentation model

In the training set, we randomly chose 230 images for segmentation model training, while the remaining 919 images were used for validation. Two experienced radiologists (L.L. and W.C.) manually delineated the ROIs in the selected images using Labelme software (V5.4.0). Their manual annotations provided the reference standard for evaluating the accuracy of these models. To evaluate inter-observer and intra-observer concordance, the intraclass correlation coefficient (ICC) was utilized. Specifically, an ICC threshold of ≥0.75 was established as evidence for commendable levels of agreement. We trained an array of cutting-edge deep learning-based architectures for the task of image segmentation. These models encompassed various prominent variants such as FCN ResNet50, FCN ResNet101, DeepLabV3 ResNet50, DeepLabV3 ResNet101, DeepLabV3 MobileNetV3-Large, LR-ASPP MobileNetV3-Large, and U-Net. The Dice similarity coefficient (DSC) was utilized to evaluate how well the automated segmentations aligned with expert manual markings, thereby assessing segmentation performance. The model demonstrating the highest accuracy was selected for implementing automated segmentation on the leftover images. To uphold quality standards, instances of imprecise segmentations underwent manual refinement by domain experts.

Feature extraction

Within the research workflow, extraction of radiomic features was performed using the validated Pyradiomics software application. These extracted features fall into three primary categories: geometric, intensity-based, and textural features. Geometric features provide insights into the morphological properties of the ROI. Intensity-based features represent the first-order statistical distribution of voxel intensities within the ROI. Textural features, on the other hand, capture patterns or higher-order spatial relationships among intensities, encompassing metrics such as gray-level co-occurrence matrix (GLCM), gray-level dependence matrix (GLDM), gray-level size zone matrix (GLSZM), gray-level run length matrix (GLRLM), and neighboring gray tone difference matrix (NGTDM). The extraction of deep learning features originated from a comprehensive set of 30 models, each stemming from one of nine prominent CNN families: ResNet, VGG, DenseNet, MobileNet, Inception, SqueezeNet, ShuffleNetV2, MNASNet, and AlexNet. To address the overfitting issue commonly encountered by deep learning models due to limited training images, we employed transfer learning to pre-train the model utilizing image datasets curated from ImageNet. Thereafter, leveraging the parameters cultivated during this pre-training phase within the deep transfer learning (DTL) framework, we proceeded to extract deep learning features—commonly derived from the output of the penultimate network layer.

Feature selection

The extracted features underwent Z-score normalization to conform to a standard normal distribution. Subsequently, for the purpose of feature screening, either the *t*-test or the Mann-Whitney *U*-test was utilized based on data characteristics. For the identification of features demonstrating notable distinctions between standard versus non-standard planes, we exclusively preserved those characterized by a *P*-value inferior to 0.05. Addressing the challenge of substantial feature redundancy, we utilized Spearman's rank correlation coefficient as an analytical tool to ascertain inter-feature correlations and subsequently evaluate potential multicollinear effects. From each pair of features exhibiting a correlation coefficient surpassing 0.9, we selectively preserved just one representative feature, effectively removing the highly redundant counterparts. Within the Python environment (V3.70), the least absolute shrinkage and selection operator (LASSO) regression model was applied to carry out feature selection tasks and subsequently

reduce the dimensionality of the dataset. Based on the regulation parameter λ, the LASSO method exerted a shrinkage effect on all regression coefficients toward 0, while rigorously assigning the coefficients of irrelevant features to 0. Employing a 10-fold cross-validation framework under the minimal criteria paradigm, we identified the optimal value of λ as that which yielded the lowest cross-validation error. The Rad score was formulated based on a curated set of highly robust and minimally redundant features. Its computation involved a linear synthesis of these features, wherein each feature's contribution was proportional to its associated model coefficient.

Model development and assessment

We input the selected radiomics or DTL features into different machine learning algorithms to establish classification models using Python (V3.70), including Logistic Regression (LR), Support Vector Machine (SVM), k-nearest neighbor (KNN), Random Forest, Extra Trees, XGBoost, LightGBM, and Multi-Layer Perception (MLP). Subsequently, a 5-fold cross-validation procedure was implemented with the aim of identifying the most suitable hyperparameters for model fitting. Receiver operating characteristic (ROC) curves were generated for the visual evaluation of each model's classification efficacy. Concurrently, key performance metrics—including the area under the ROC curve (AUC), specificity, sensitivity, accuracy, positive predictive value (PPV), negative predictive value (NPV), and precision—were computed to quantitatively characterize model performance.

Fusion models development

Fusion models combining the strengths of radiomics and CNN were developed to enhance classification performance. The fusion strategies included early fusion (feature-level integration) and late fusion (result-level integration). In general, early fusion denotes the procedure of integrating heterogeneous features originating from diverse sources into a cohesive dataset right at the initial input phase of the model. In this study, the implementation of early fusion involved combining extracted radiomics features with DTL features via concatenation. Subsequently, these integrated features were subjected to feature screening and subsequent model construction processes, as detailed earlier, ultimately resulting in the formulation of the early fusion model. Under the paradigm of late fusion, individual predictive results for each modality are first generated by distinct models, followed by their integration to produce the ultimate output. Late fusion admits multiple implementation modalities, including but not limited to voting protocols, weighted averaging techniques, stacking ensemble approaches, or employing an additional machine learning model for the purpose of synthesizing outcomes generated by distinct modalities. In this study, late fusion was realized via the independent development of distinct radiomics and DTL models. Subsequently, their respective outputs were consolidated employing a stacking ensemble approach, selected for its superior classification performance in constructing the late fusion model [11].

Comparison of radiologists and AI model

Two seasoned senior radiologists (H.Z. and L.L.) conducted independent categorizations of all images into standard versus non-standard planes following the established protocols. Two senior radiologists (H.T. and W.C.) and two junior radiologists (T.W. and Y. L.) independently classified the images in the validation set. A comparative analysis was conducted between the classification capabilities of human radiologists and those of the AI model. Subsequently, junior radiologists were tasked with reclassifying each image under the guidance of the AI model, followed by an assessment of their enhanced performance metrics.

Model interpretability and visualization

To visualize the decision-making mechanism of the model and elucidate the importance of features along with their influence on it, we utilized the SHAP approach—originating from game theory—to generate visual representations of machine learning model outputs. This technique helps in identifying and prioritizing the features that play a crucial role in compound classification.

Statistical analysis

For continuous data, descriptive statistics including mean ± standard deviation (SD) were provided, followed by comparative analyses conducted with either parametric *t*-tests or nonparametric Mann–Whitney *U*-tests implemented in IBM SPSS (V21.0). A two-tailed *P*-value threshold below 0.05 was adopted to define statistical significance. Furthermore, the calculation of a 95% confidence interval (CI) accompanied the determination of AUC values. Additionally, Python (V3.70) was used to perform Z-score normalization, ICCs, Spearman rank correlation tests, and LASSO regression analysis. The DeLong test was conducted to compare the AUCs of the models. The calibration efficacy of the models was evaluated through Hosmer-Lemeshow testing, which examines the concordance between predicted and observed classification probabilities. Additionally, decision curve analysis (DCA) was conducted to gauge the clinical practicality and utility of these predictive tools.

Results

Clinical characteristics

A total of 1149 ultrasound images of fetal midsagittal planes were collected in our study. Of these, 804 images originated from one medical institution, constituting the training set, while 345 images sourced from another hospital formed the validation set, maintaining a ratio of approximately 7:3. According to the radiologists' categorization, 561 images (48.83%) were identified as standard planes, while 588 images (51.17%) were considered non-standard planes. Within the training set, 408 images (50.75%) were categorized as standard sections, whereas 396 images (49.25%) were classified as non-standard sections. Within the validation set, 153 images (44.35%) were classified as standard sections, while 192 images (55.65%) were designated as non-standard sections (Figure 2). The average age of pregnant women was 31.89 ± 7.23 years in the training set and 32.75 ± 6.84 years in the validation set. The average fetal CRL was 65.32 ± 10.54 mm and 64.57 ± 9.89 mm in the training set and validation set, respectively.

Image segmentation models

The DeepLabV3 ResNet101 segmentation model demonstrating the highest performance (DSC: 97.15 ± 0.20%) was employed for automated processing of the remaining image dataset (Table 1). On average, the automated segmentation process for an individual image required 3.9 seconds to complete, compared to an average of 15.6 seconds required for manual annotation per image. The hyperparameters for various models are detailed in Appendix 1.

Feature extraction and feature selection

A set of 107 radiomics features was extracted from the regions of interest, which included 18 first-order features, 75 texture features, and 14 shape features. Furthermore, 30 CNN models were pre-trained to extract DTL features, with the number of features ranging from 512 to 4096 depending on the specific model. After feature selection, the most robust features (18 radiomics features and varying from 63 to 356 DTL features depending on the specific model) were retained to construct the models. Figure 3 presents the selection processes of radiomics features and DTL features.

Radiomics model

A series of machine learning models were established and the SVM model exhibited the optimal performance (AUC: 0.871, 95% CI: 0.833-0.908, accuracy: 78.9%, sensitivity: 85.0%, specificity: 72.5%, PPV: 76.1%, NPV: 82.5%, precision: 76.1%) in the validation set (Table 2, Figure 4A).

DTL models

The selected DTL features were utilized to construct DTL models. The pre-trained DenseNet201 model, when combined with the SVM algorithm, demonstrated superior performance in the validation set (AUC: 0.929, 95% CI: 0.901-0.956, accuracy: 87.5%, sensitivity: 83.0%, specificity: 91.1%, PPV: 88.2%, NPV: 87.1%, precision: 88.2%) (Table 3, Figure 4B).

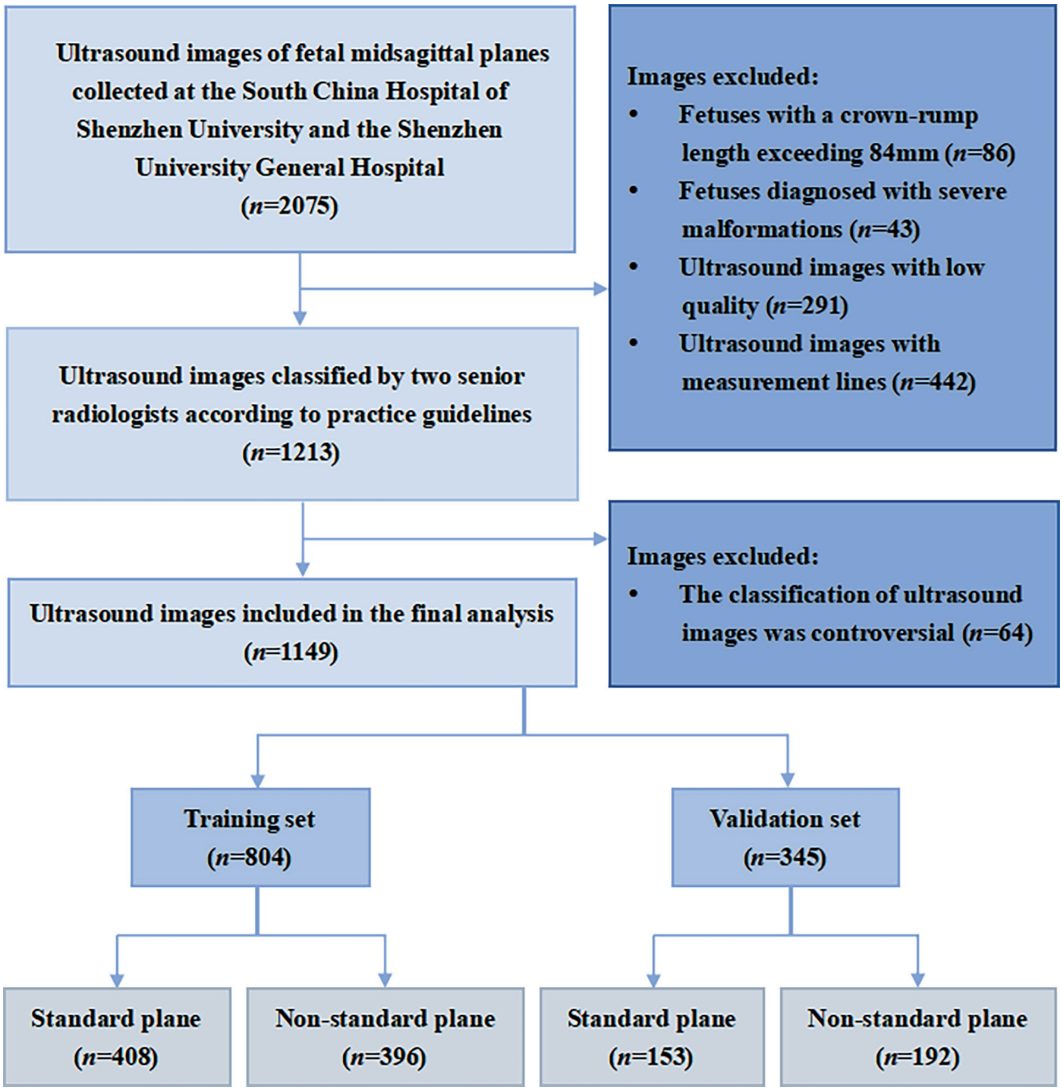

Figure 2. Flowchart of the study subjects screening based on inclusion and exclusion criteria.

Table 1. DSC (%) of the deep learning-based segmentation models compared with the radiologists' annotations.

Model	FCN ResNet50	FCN ResNet101	DeepLabV3 ResNet50	DeepLabV3 ResNet101	DeepLabV3 MobileNetV3-Large	LR-ASPP MobileNetV3-Large	U-Net
Radiologist 1 (Mean ± SD, %)	97.07 ± 0.21	97.14 ± 0.18	96.55 ± 0.23	**97.15 ± 0.20**	94.21 ± 0.22	94.37 ± 0.18	90.53 ± 0.18
Radiologist 2 (Mean ± SD, %)	97.00 ± 0.35	97.08 ± 0.26	96.73 ± 0.19	**97.09 ± 0.31**	94.18 ± 0.27	94.46 ± 0.21	90.53 ± 0.16

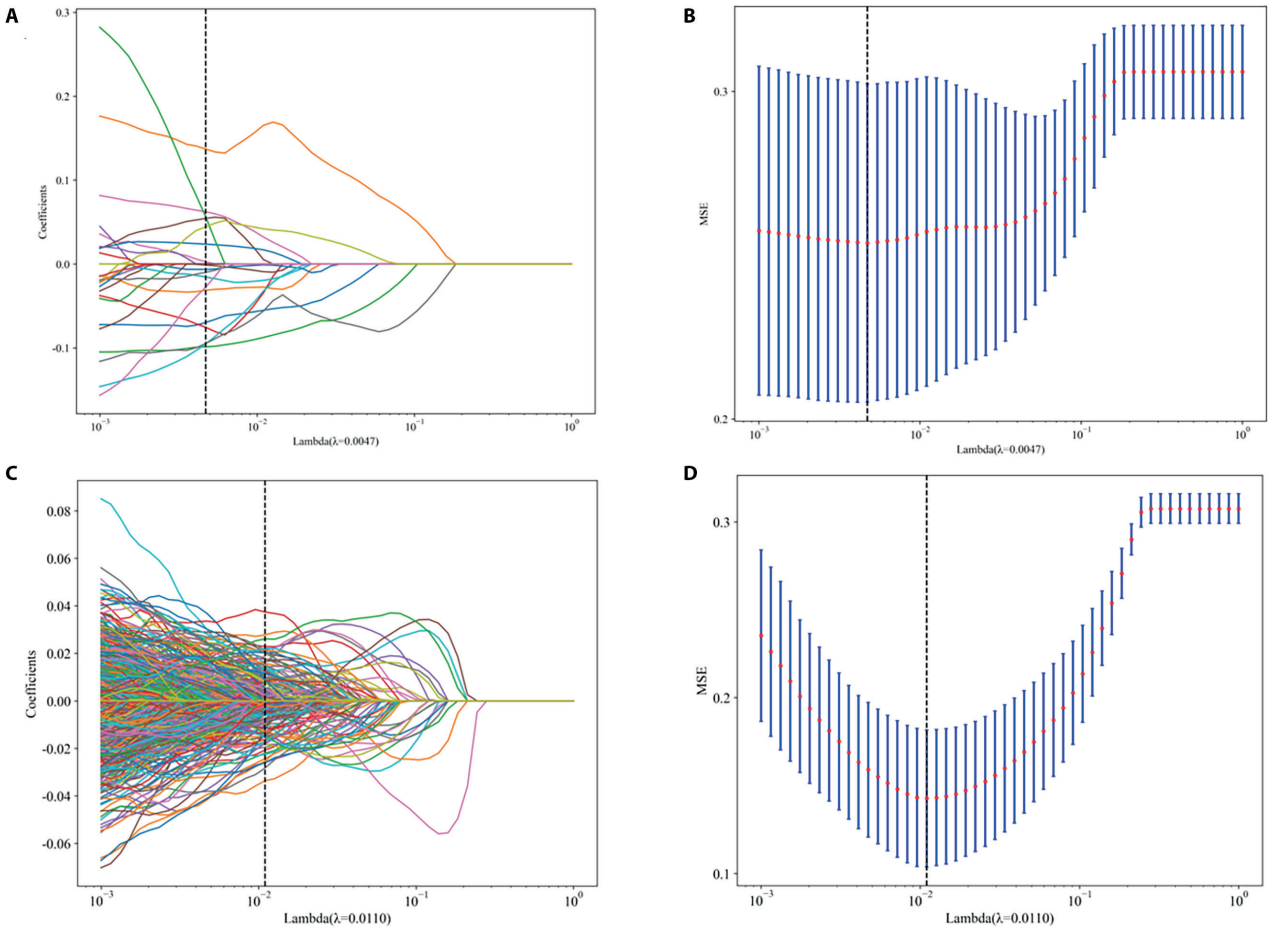

Figure 3. The selection processes of radiomics features and DTL features using LASSO logistic regression. (A) Radiomics feature selection: Coefficients of 10-fold cross-validation based on LASSO. (B) Radiomics feature selection: MSE of 10-fold cross-validation based on LASSO. (C) DTL (DenseNet201 model) feature selection: Coefficients of 10-fold cross-validation based on LASSO. (D) DTL (DenseNet201 model) feature selection: MSE of 10-fold cross-validation based on LASSO. MSE: mean square error.

Fusion models

The radiomics features and DTL features (DenseNet201) were concatenated to develop early fusion models, with the SVM algorithm demonstrating the highest performance in the validation set (AUC: 0.947, 95% CI: 0.924-0.970, accuracy: 88.4%, sensitivity: 83.0%, specificity: 92.7%, PPV: 90.1%, NPV: 87.3%, precision: 90.1%) (Table 4, Figure 4C). The results of the radiomics model and DTL model (DenseNet201) were combined to establish late fusion models, with the MLP algorithm exhibiting the best classification performance in the validation set (AUC: 0.940, 95% CI: 0.916-0.964, accuracy: 87.2%, sensitivity: 85.0%, specificity: 89.1%, PPV: 86.1%, NPV: 88.1%, precision: 86.1%) (Table 4).

Comparison of radiomics, DTL, and fusion models

Compared to the radiomics, DTL, and late fusion models, the early fusion model exhibited markedly better performance metrics (Table 4, Figure 5A). Through DeLong testing, significant intergroup variations were identified in AUC metrics when comparing both the radiomics versus DTL models and the radiomics against various fusion models within the validation set ($P < 0.05$). However, the DeLong test indicated

Table 2. Classification performance of the radiomics models.

Algorithm	Set	AUC (95% CI)	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Precision (%)
LR	Training	0.803 (0.773-0.833)	72.5	71.8	73.2	73.4	71.6	73.4
LR	Validation	0.841 (0.801-0.882)	73.9	85.6	64.6	65.8	84.9	65.8
SVM	**Training**	**0.863 (0.838-0.888)**	**78.9**	**85.0**	**72.5**	**76.1**	**82.5**	**76.1**
SVM	**Validation**	**0.871 (0.833-0.908)**	**80.3**	**73.9**	**85.4**	**80.1**	**80.4**	**80.1**
KNN	Training	0.888 (0.867-0.909)	76.7	60.3	93.7	90.8	69.6	90.8
KNN	Validation	0.798 (0.752-0.844)	72.8	57.5	84.9	75.2	71.5	75.2
Random Forest	Training	0.813 (0.784-0.842)	73.4	81.6	64.9	70.6	77.4	70.6
Random Forest	Validation	0.831 (0.788-0.873)	76.8	71.2	81.2	75.2	78.0	75.2
Extra Trees	Training	0.742 (0.709-0.776)	68.4	71.8	64.9	67.8	69.1	67.8
Extra Trees	Validation	0.743 (0.691-0.794)	67.5	73.2	63.0	61.2	74.7	61.2
XGBoost	Training	0.886 (0.864-0.908)	80.2	78.7	81.8	81.7	78.8	81.7
XGBoost	Validation	0.834 (0.792-0.876)	74.5	79.1	70.8	68.4	81.0	68.4
LightGBM	Training	0.847 (0.821-0.874)	76.6	74.8	78.5	78.2	75.1	78.2
LightGBM	Validation	0.827 (0.784-0.871)	75.7	77.1	74.5	70.7	80.3	70.7
MLP	Training	0.848 (0.822-0.874)	76.9	81.1	72.5	75.2	78.8	75.2
MLP	Validation	0.854 (0.816-0.893)	77.7	77.1	78.1	73.7	81.1	73.7

no statistically significant difference in AUC between the DTL and early fusion model ($P > 0.05$) (Figure 5B). The results indicated that both the DTL and fusion models surpassed the radiomics model in classification performance. The DCA revealed that each evaluated model substantially improved classification performance relative to scenarios devoid of predictive algorithms, with the DTL and fusion models showing particularly strong enhancements (Figure 5C). The Hosmer-Lemeshow test indicated strong agreement between the predicted probabilities of the late fusion model and the actual classification outcomes in the validation set ($P > 0.05$) (Figure 5D).

Comparison with radiologists

In the validation set, the early fusion model achieved classification performance metrics on par with those demonstrated by senior radiologists (average AUC: 0.950, accuracy: 93.0%, sensitivity: 92.8%, specificity: 93.2%, PPV: 91.6%, NPV: 94.2%, precision: 91.6%). In comparison, the junior radiologists

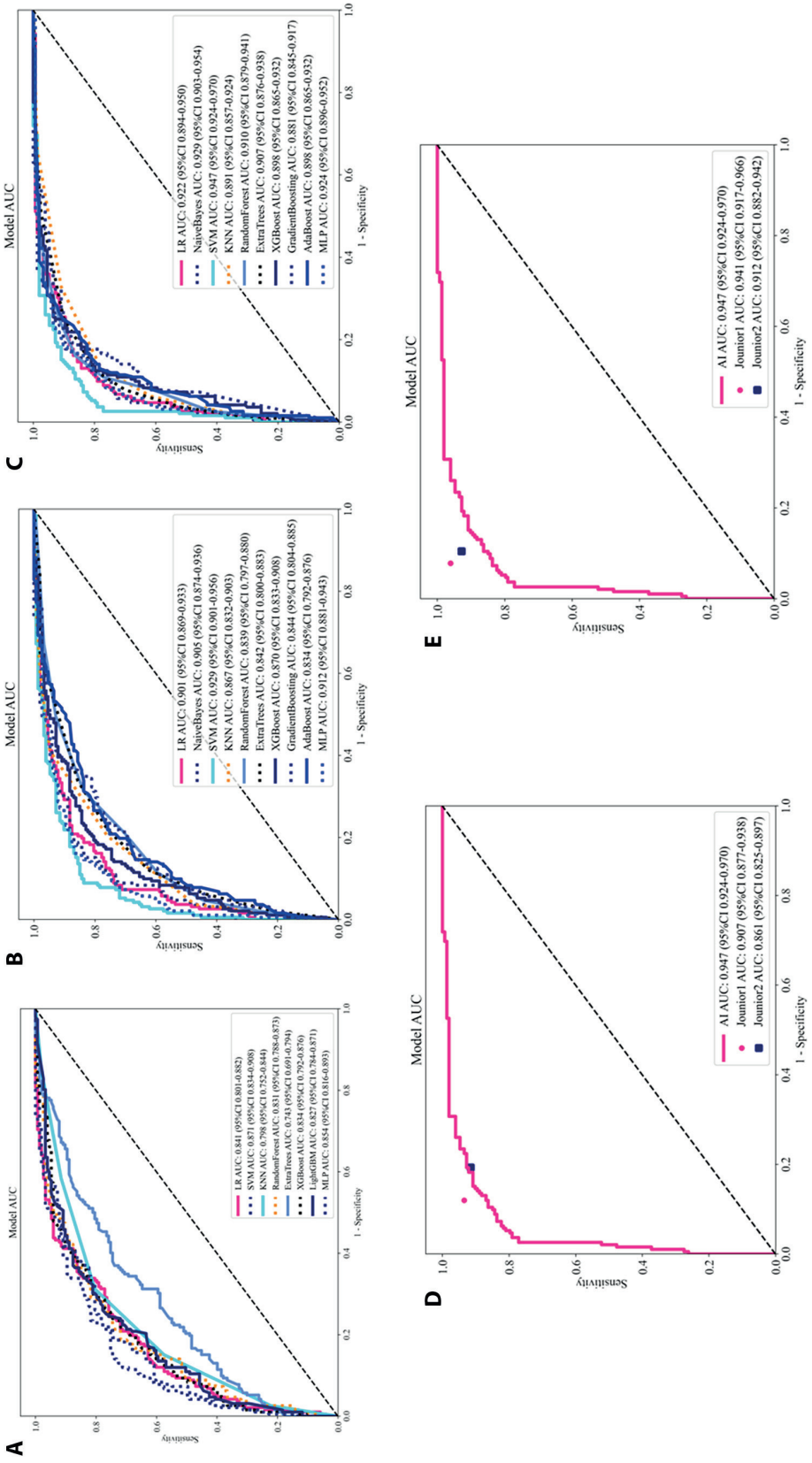

Figure 4. The ROC curves of the models in the validation set. (A) The ROC curves of radiomics model. (B) The ROC curves of DTL model (DenseNet201). (C) The ROC curves of early fusion model. (D) The ROC curves for junior radiologists without the assistance of AI model. (E) The ROC curves for junior radiologists with the assistance of AI model.

Table 3. Classification performance of the DTL models.

Model Series	Model Name	Set	AUC (95% CI)	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Precision (%)
ResNet	ResNet18	Training	0.996 (0.993-0.999)	97.4	97.3	97.5	97.5	97.2	97.5
	ResNet18	Validation	0.892 (0.858-0.926)	82.9	77.1	87.5	83.1	82.8	83.1
	ResNet50	Training	1.000 (1.000-1.000)	99.9	99.8	100.0	100.0	99.7	100.0
	ResNet50	Validation	0.911 (0.881-0.941)	84.1	88.2	80.7	78.5	89.6	78.5
	ResNet101	Training	0.999 (0.999-1.000)	99.0	98.8	99.2	99.3	98.7	99.3
	ResNet101	Validation	0.890 (0.856-0.925)	82.6	85.0	80.7	77.8	87.1	77.8
	ResNeXt50_32x4d	Training	0.997 (0.995-1.000)	97.8	97.1	98.5	98.5	97.0	98.5
	ResNeXt50_32x4d	Validation	0.907 (0.875-0.938)	83.2	88.2	79.2	77.1	89.4	77.1
	ResNeXt101_32x8d	Training	0.980 (0.973-0.987)	92.7	94.4	90.9	91.4	94.0	91.4
	ResNeXt101_32x8d	Validation	0.914 (0.885-0.943)	84.9	86.3	83.9	81.0	88.5	81.0
	Wide_ResNet50_2	Training	0.994 (0.991-0.998)	97.4	98.0	96.7	96.9	98.0	96.9
	Wide_ResNet50_2	Validation	0.909 (0.878-0.939)	85.8	79.1	91.1	87.7	84.5	87.7
	Wide_ResNet101_2	Training	0.967 (0.956-0.978)	91.4	89.5	93.4	93.4	89.6	93.4
	Wide_ResNet101_2	Validation	0.908 (0.877-0.938)	82.3	86.3	79.2	76.7	87.9	76.7
VGG	VGG11	Training	0.979 (0.970-0.988)	93.2	93.6	92.7	92.9	93.4	92.9
	VGG11	Validation	0.924 (0.898-0.951)	84.3	88.9	80.7	78.6	90.1	78.6
	VGG11_bn	Training	1.000 (1.000-1.000)	99.8	99.5	100.0	100.0	99.5	100.0
	VGG11_bn	Validation	0.908 (0.878-0.939)	83.5	84.3	82.8	79.6	86.9	79.6
	VGG13	Training	0.983 (0.974-0.993)	95.3	94.9	95.7	95.8	94.7	95.8
	VGG13	Validation	0.914 (0.885-0.943)	84.1	81.0	86.5	82.7	85.1	82.7
	VGG13_bn	Training	0.965 (0.952-0.978)	91.7	87.3	96.2	96.0	88.0	96.0
	VGG13_bn	Validation	0.894 (0.861-0.927)	82.0	82.4	81.8	78.3	85.3	78.3
	VGG16	Training	1.000 (1.000-1.000)	99.8	99.5	100.0	100.0	99.5	100.0
	VGG16	Validation	0.892 (0.856-0.928)	83.8	79.7	87.0	83.0	84.3	83.0
	VGG16_bn	Training	0.985 (0.978-0.992)	93.9	91.7	96.2	96.1	91.8	96.1
	VGG16_bn	Validation	0.870 (0.833-0.908)	80.6	82.4	79.2	75.9	84.9	75.9
	VGG19	Training	0.973 (0.963-0.983)	91.8	90.7	92.9	93.0	90.6	93.0
	VGG19	Validation	0.881 (0.846-0.917)	82.0	81.0	82.8	79.0	84.6	79.0
	VGG19_bn	Training	0.970 (0.959-0.982)	92.9	90.9	94.9	94.9	91.0	94.9
	VGG19_bn	Validation	0.925 (0.899-0.952)	85.5	83.0	87.5	84.1	86.6	84.1

Model Series	Model Name	Set	AUC (95% CI)	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Precision (%)
DenseNet	DenseNet121	Training	0.999 (0.999-1.000)	98.6	98.0	99.2	99.3	98.0	99.3
	DenseNet121	Validation	0.907 (0.876-0.938)	83.2	81.7	84.4	80.6	85.3	80.6
	DenseNet169	Training	0.994 (0.991-0.997)	96.4	96.6	96.2	96.3	96.5	96.3
	DenseNet169	Validation	0.907 (0.876-0.937)	81.7	88.9	76.0	74.7	89.6	74.7
	DenseNet201	**Training**	**1.000 (1.000-1.000)**	**99.4**	**99.0**	**99.7**	**99.8**	**99.0**	**99.8**
	DenseNet201	**Validation**	**0.929 (0.901-0.956)**	**87.5**	**83.0**	**91.1**	**88.2**	**87.1**	**88.2**
MobileNet	MobileNet_v2	Training	0.992 (0.988-0.996)	95.9	96.6	95.2	95.4	96.4	95.4
	MobileNet_v2	Validation	0.910 (0.879-0.941)	83.2	79.1	86.5	82.3	83.8	82.3
	MobileNet_v3_large	Training	0.991 (0.987-0.995)	95.3	94.4	96.2	96.2	94.3	96.2
	MobileNet_v3_large	Validation	0.845 (0.804-0.886)	77.4	83.7	72.4	70.7	84.8	70.7
	MobileNet_v3_small	Training	0.989 (0.983-0.994)	95.5	96.6	94.4	94.7	96.4	94.7
	MobileNet_v3_small	Validation	0.856 (0.816-0.895)	78.6	75.2	81.2	76.2	80.4	76.2
Inception	GoogleNet	Training	0.968 (0.957-0.978)	90.7	90.4	90.9	91.1	90.2	91.1
	GoogleNet	Validation	0.816 (0.771-0.861)	75.4	81.0	70.8	68.9	82.4	68.9
	Inception_v3	Training	0.993 (0.989-0.997)	95.9	94.6	97.2	97.2	94.6	97.2
	Inception_v3	Validation	0.919 (0.890-0.948)	84.6	83.7	85.4	82.1	86.8	82.1
SqueezeNet	SqueezeNet1_0	Training	0.967 (0.954-0.980)	93.0	91.7	94.4	94.4	91.7	94.4
	SqueezeNet1_0	Validation	0.848 (0.807-0.888)	77.1	64.7	87.0	79.8	75.6	79.8
	SqueezeNet1_1	Training	0.973 (0.962-0.985)	93.5	92.9	94.2	94.3	92.8	94.3
	SqueezeNet1_1	Validation	0.800 (0.753-0.847)	75.4	75.2	75.5	71.0	79.2	71.0
ShuffleNetV2	ShuffleNet_v2_x0_5	Training	0.983 (0.975-0.991)	94.4	91.9	97.0	96.9	92.1	96.9
	ShuffleNet_v2_x0_5	Validation	0.905 (0.873-0.937)	83.2	79.1	86.5	82.3	83.8	82.3
	ShuffleNet_v2_x1_0	Training	0.974 (0.962-0.985)	93.2	94.9	91.4	91.9	94.5	91.9
	ShuffleNet_v2_x1_0	Validation	0.889 (0.853-0.925)	84.3	81.7	86.5	82.8	85.6	82.8
MNASNet	MNASNet0_5	Training	0.997 (0.993-1.000)	98.0	98.5	97.5	97.6	98.5	97.6
	MNASNet0_5	Validation	0.858 (0.817-0.899)	80.3	83.0	78.1	75.1	85.2	75.1
	MNASNet1_0	Training	0.997 (0.995-0.999)	97.5	97.3	97.7	97.8	97.2	97.8
	MNASNet1_0	Validation	0.863 (0.825-0.901)	78.8	79.1	78.6	74.7	82.5	74.7
AlexNet	AlexNet	Training	0.995 (0.991-0.999)	97.3	97.5	97.0	97.1	97.5	97.1
	AlexNet	Validation	0.878 (0.841-0.914)	81.2	87.6	76.0	74.4	88.5	74.4

Table 4. Classification performance of radiomics, DTL, and fusion models.

Model	Set	AUC (95% CI)	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Precision (%)
Radiomics	Training	0.863 (0.838-0.888)	78.9	85.0	72.5	76.1	82.5	76.1
Radiomics	Validation	0.871 (0.833-0.908)	80.3	73.9	85.4	80.1	80.4	80.1
DTL	Training	1.000 (1.000-1.000)	99.4	99.0	99.7	99.8	99.0	99.8
DTL	Validation	0.929 (0.901-0.956)	87.5	83.0	91.1	88.2	87.1	88.2
Early fusion	**Training**	**0.982 (0.974-0.991)**	**94.4**	**94.1**	**94.7**	**94.8**	**94.0**	**94.8**
Early fusion	**Validation**	**0.947 (0.924-0.970)**	**88.4**	**83.0**	**92.7**	**90.1**	**87.3**	**90.1**
Late fusion	Training	0.976 (0.966-0.985)	93.4	92.2	94.7	94.7	92.1	94.7
Late fusion	Validation	0.940 (0.916-0.964)	87.2	85.0	89.1	86.1	88.1	86.1

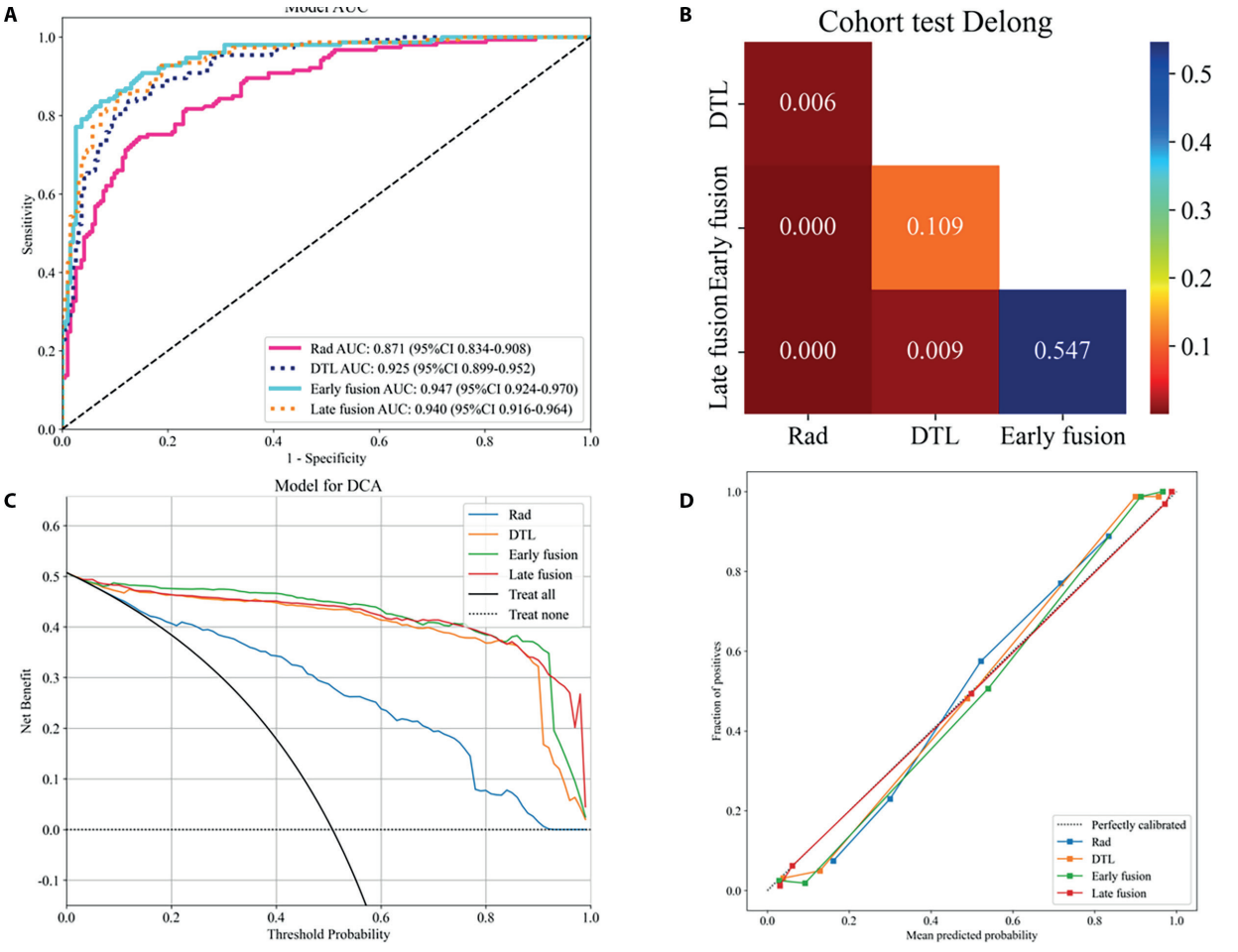

Figure 5. Comparison of radiomics model, DTL model (DenseNet201), early fusion model, and late fusion model in the validation set. (A) The ROC curves of different model. (B) The *P*-value of DeLong test between different models. (C) The DCA curves for different models. (D) The calibration curves for different models.

Table 5. Classification performance of radiologists without and with the assistance of the AI model.

Group	Set	AUC (95% CI)	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Precision (%)
AI	Early fusion	0.947 (0.924-0.970)	88.4	83.0	92.7	90.1	87.3	90.1
Junior radiologist 1	Without AI	0.907 (0.877-0.938)	90.4	93.5	88.0	86.1	94.4	86.1
Junior radiologist 2	Without AI	0.861 (0.825-0.897)	85.5	91.5	80.7	79.1	92.3	79.1
Junior radiologist (Average)	Without AI	0.884	88.0	92.5	84.4	82.6	93.3	82.6
Junior radiologist 1	**With AI**	**0.941 (0.917-0.966)**	**93.9**	**96.1**	**92.2**	**90.7**	**96.7**	**90.7**
Junior radiologist 2	**With AI**	**0.912 (0.882-0.942)**	**91.0**	**92.8**	**89.6**	**87.7**	**94.0**	**87.7**
Junior radiologist (Average)	With AI	0.927	92.5	94.4	90.9	89.2	95.4	89.2
Senior radiologist (Average)	Without AI	0.950	93.0	92.8	93.2	91.6	94.2	91.6

demonstrated inferior classification performance compared to the AI model (early fusion model) (average AUC: 0.884, accuracy: 88.0%, sensitivity: 92.5%, specificity: 84.4%, PPV: 82.6%, NPV: 93.3%, precision: 82.6%) (Table 5, Figure 4D). Nevertheless, junior radiologists significantly improved their performance with the aid of the AI model (average AUC: 0.927, accuracy: 92.5%, sensitivity: 94.4%, specificity: 90.9%, PPV: 89.2%, NPV: 95.4%, precision: 89.2%) (Table 5, Figure 4E).

Interpretability and visualization of the early fusion model

Within the SHAP summary plot, feature significance is quantitatively represented by their respective SHAP values, where higher absolute magnitudes correspond to stronger influence on model predictive results (Figure 6A). The summary plot ranks features based on their importance in descending order, highlighting that the texture feature original_glcm_MCC is the most important contributor to the early fusion model. Employing colored directional arrows, the SHAP force plot graphically represents feature-specific SHAP values to elucidate their respective impacts on predictive outcomes for single image instances (Figures 6B and 6C). Arrow length quantitatively signifies a feature's impact magnitude on predicted values, whereas hue differentiation—utilizing red for positive and blue for negative contributions—explicitly communicates the directional polarity of each feature's effect. Ultimately, the definitive SHAP score corresponds to the image's predicted outcome, calculated as the aggregate of its foundational baseline value and the summated influence contributions from all constituent features.

Discussion

In this work, we innovatively introduced two fusion strategies—feature fusion and result fusion—that combine the strengths of radiomics and DTL to improve the classification accuracy for quality evaluation of ultrasound images used in CRL measurements. The early fusion strategy concatenating the features of radiomics and DTL outperformed single-modality architectures and late fusion model in terms of classification accuracy, and external validation also showed commendable generalizability. It can automatically categorize ultrasound images into standard and non-standard planes, achieving intelligent quality control

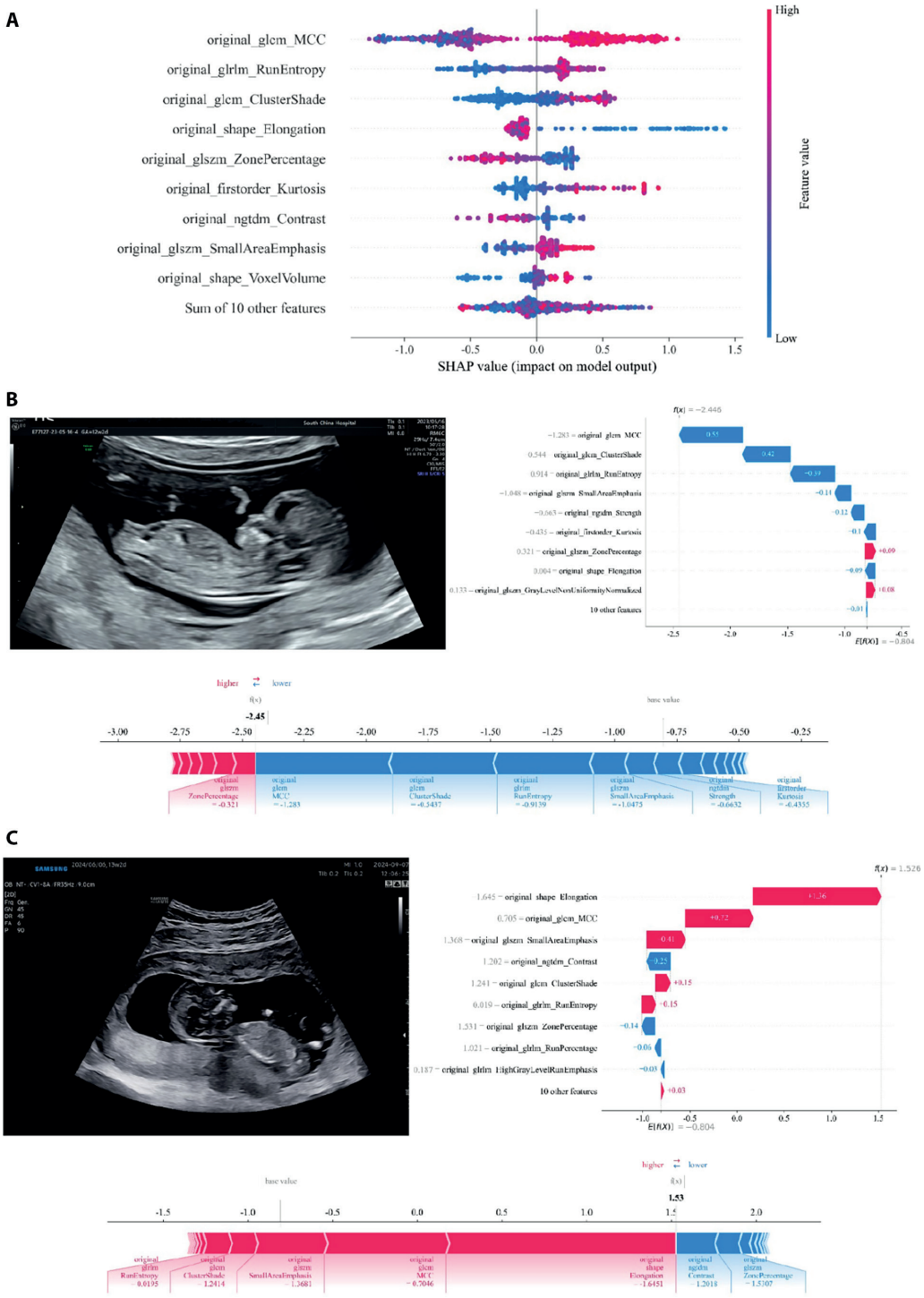

Figure 6. Interpretability and visualization of the early fusion model using SHAP method. (A) The SHAP summary plot. (B) The SHAP force plot of a standard plane: the final SHAP value for this image was -2.446, which was lower than the base value of -0.804, thereby indicating that this image was classified as a standard plane. (C) The SHAP force plot of a non-standard plane: the final SHAP value for this image was 1.526, which was higher than the base value of -0.804, thereby indicating that this image was classified as a non-standard plane.

for obstetric ultrasound. Approximately one-third of the pregnant women enrolled in this study exhibit a body mass index (BMI) exceeding 24 or have a prior history of cesarean section surgery. The classification performance of the model when applied to images of these pregnant women did not show significant differences. Only two images were obtained via transvaginal scan (TVS) and the rest of the images were acquired through transabdominal scan (TAS). The model correctly classified the two TVS images. As the sample size of TVS images was too small, we were unable to assess whether a TVS image / TAS image can affect the final classification. Furthermore, the AI model matched the expertise level of senior staff radiologists and outperformed entry-level radiologists, and the assistance provided by the AI model can significantly enhance the classification ability of junior radiologists. Moreover, the image segmentation model exhibited satisfactory segmentation accuracy, facilitating a substantial 75% decrease in processing time allocated for ROI delineation and significantly reducing the workload of manual annotation. Prior investigations have engineered diverse AI architectures for interpreting first-trimester fetal midsagittal ultrasound scans [12-23]. Nevertheless, extant literature has predominantly directed attention toward the midsagittal planes of the fetal face for nuchal translucency measurement. Primarily, such studies prioritized the identification of key anatomical formations, precise localization of reference sections [13-16], performing automated measurement for fetal biometry [13, 17], or screening for developmental anomalies in fetal specimens [18-20]. In comparison, limited scholarly attention has been devoted to the quality evaluation of standard midsagittal planes of the whole fetus for CRL measurement. Zhen et al. [14] established a system for the automatic detection of nine standard planes in first-trimester ultrasound scans, including the midsagittal plane for CRL measurement in ultrasound videos. Evaluation results indicated that the system achieved a high level of accuracy in standard plane detection, with quality equivalent to the assessments made by senior sonographers. A primary limitation of AI models is the lack of interpretability, making it difficult to comprehend the rationale behind their decisions and hindering their clinical application [10]. Explainable AI aims to uncover the decision-making processes of black-box AI models by interpreting their internal mechanisms and identifying the critical factors that drive computational outcomes [8]. SHAP method is increasingly being utilized for interpreting and visualizing the prediction process of machine learning models by assigning attribution values to each feature of the model [24]. In this research, we utilized SHAP to demystify the black-box design of the early fusion model, providing an intuitive explanation of the contributions of both radiomics and DTL features and their influence on model predictions [9]. In the SHAP summary plot, each dot corresponds to an individual instance, while their associated SHAP values quantitatively reflect the feature's contribution to the model (with a larger SHAP value signifying a greater contribution) [25]. Figure 6A illustrates that the pivotal features determining the prediction outcome of early fusion model are the texture features. Moreover, the SHAP force plot demonstrates which features the model mainly relies on to predict the classification outcome for each image. The proposed AI models in this study could potentially integrate into routine clinical practice to optimize the workflow of ultrasound scan during early pregnancy. Performing rigorous quality evaluation for ultrasound images is both time-consuming and labor-intensive. This model can realize intelligent quality control and accurately classify images into standard and non-standard planes, alleviating the workload of radiologists. It also addresses some of the inherent problems of ultrasound examination, such as subjectivity and interobserver variability [26-28]. Moreover, using the SHAP method for model interpretability is valuable in enhancing the professional skills of junior radiologists. When encountering diagnostic challenges or ambiguities, radiologists may consult the analytical insights derived from SHAP force plots to gain an understanding of the model's decision-making mechanism and identify which features—such as shape, texture, or intensity—played a predominant role in the classification procedure. The present study suffered from multiple limitations: (1) Its retrospective design combined with a limited sample volume increases susceptibility to sampling bias and raises concerns about the representativeness due to possible distributional changes [29, 30]. (2) This was a two-center study, and

subsequent multi-centric prospective studies involving larger cohorts are imperative for further validation. (3) Within this research, image categorization was confined to distinguishing between standard and non-standard anatomical planes, and the next step of our research will involve further scoring of images according to guidelines. (4) Considering that the ultrasonographic machines used for image acquisition are pretty much the best in their class, whether the model also works as well when applied to images collected from older or inferior quality machines remains to be further validated. We proactively acquired a cohort of 105 images generated by older or inferior-quality ultrasound machines from other hospitals. The result revealed that the model's classification performance, while slightly inferior to the results presented in this study (AUC: 0.921, 95% CI: 0.906-0.936), still surpasses that of junior radiologists. (5) We also assessed the classification performance of experienced radiologists when assisted by the AI model. However, it was unfortunate to find that their classification accuracy did not yield further improvement.

Conclusions

We introduced an innovative early fusion architecture designed for intelligent quality assessment of ultrasound images in CRL measurements. To elucidate the decision-making mechanism within the AI model, the SHAP method was employed to enhance its interpretability and facilitate visual representation. We also established an accurate image segmentation model to achieve automatic delineation of ROI. These models are anticipated to serve as an effective tool for optimizing clinical workflow of ultrasound examination during first-trimester of pregnancy, enhancing image quality control, enhancing the professional skills of junior radiologists, and alleviating the workload of radiologists.

List of abbreviations:

CRL: crown–rump length
ROI: region of interest
DTL: deep transfer learning
SHAP: SHapley Additive exPlanations
ISUOG: International Society of Ultrasound in Obstetrics and Gynecology
AI: Artificial intelligence
CNN: convolutional neural network
ICC: interclass correlation coefficient
DSC: Dice similarity coefficient
LASSO: least absolute shrinkage and selection operator
LR: Logistic Regression
SVM: Support Vector Machine
KNN: k-nearest neighbor
MLP: Multi-Layer Perception
ROC: receiver operating characteristic
AUC: area under the curve
PPV: positive predictive value
NPV: negative predictive value
DCA: decision curve analysis
SD: standard deviation
CI: confidence interval
MSE: mean square error

Declarations

Relevance statement: The model is anticipated to serve as an effective tool for optimizing clinical workflow of ultrasound examination, enhancing image quality control, and alleviating the workload of radiologists.

Ethics approval and consent to participate: The study was conducted in accordance with the Declaration of Helsinki and approved by the ethical committees of the South China Hospital of Shenzhen University (approval number: HNLS20240326001-A). Patient consent was waived by the ethical committees due to the retrospective nature of the study.

Consent for publication: Not applicable.

Availability of data and material: The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request (W.C.). The data are not publicly available due to hospital regulations.

Competing interests: The authors declare that they have no competing interests.

Funding: This research was funded by Shenzhen Engineering Research Center of Multi-modal Fusion medical intelligent Diagnosis Technology (XMHT20220104016), Shandong Province Natural Science Foundation (No. ZR202111120048), 2022 Shinan District Science and Technology Plan Project (2023-2-015-YY), Development of innovative medical devices for pediatric ophthalmology based on machine vision and eye tracking technology (24-1-5-yqpy-23-qy), and National Natural Science Foundation of China (No. 82473113).

Authors' contributions: Conceptualization, L.L. and W.C.; data curation, L.L., W.C., Y.L., and T.W.; methodology, L.L., C.Z., and W.Z.; formal analysis, L.L., H.T. and H.Z.; supervision, C.Z., W.C., and W.Z.; validation, T.W. and Y.L.; investigation, L.L. and T.W.; writing—original draft, L.L.; writing—review and editing, C.Z. and W.C. All authors have read and agreed to the published version of the manuscript.

Acknowledgements: We sincerely thank Onekey AI platform for code consultation.

Declaration of AI and AI-assisted technologies in the writing process: During the preparation of this work, the authors did not use AI and AI-assisted technologies in the writing process.

References

1. Salomon LJ, Alfirevic Z, Bilardo CM, Chalouhi GE, Ghi T, Kagan KO, et al. ISUOG practice guidelines: performance of first-trimester fetal ultrasound scan. Ultrasound Obstet Gynecol. 2013;41(1):102-13. doi: 10.1002/uog.12342. PMID: 23280739.
2. Bilardo C M, Chaoui R, Hyett J A, Kagan KO, Karim JN, Papageorghiou AT, et al. ISUOG Practice Guidelines (updated): performance of 11-14-week ultrasound scan. Ultrasound Obstet Gynecol. 2023;61(1):127-143. doi: 10.1002/uog.26106. PMID: 36594739.
3. Salomon L J, Alfirevic Z, Da Silva Costa F, Deter RL, Figueras F, Ghi T, et al. ISUOG Practice Guidelines: ultrasound assessment of fetal biometry and growth. Ultrasound Obstet Gynecol. 2019;53(6):715-723. doi: 10.1002/uog.20272. PMID: 31169958.
4. Papageorghiou A T, Kennedy S H, Salomon L J, Ohuma EO, Cheikh Ismail L, Barros FC, et al. International standards for early fetal size and pregnancy dating based on ultrasound measurement of crown-rump length in the first trimester of pregnancy. Ultrasound Obstet Gynecol. 2014;44(6):641-8. doi: 10.1002/uog.13448. PMID: 25044000; PMCID: PMC4286014.
5. Wanyonyi S Z, Napolitano R, Ohuma E O, Salomon LJ, Papageorghiou AT. Image-scoring system for crown-rump length measurement. Ultrasound Obstet Gynecol. 2014;44(6):649-54. doi: 10.1002/uog.13376. PMID: 24677327.
6. Ramirez Zegarra R, Ghi T. Use of artificial intelligence and deep learning in fetal ultrasound imaging. Ultrasound Obstet Gynecol. 2023;62(2):185-194. doi: 10.1002/uog.26130. PMID: 36436205.
7. Fiorentino M C, Villani F P, Di Cosmo M, Frontoni E, Moccia S. A review on deep-learning algorithms for fetal ultrasound -image analysis. Med Image Anal. 2023;83:102629. doi: 10.1016/j.media.2022.102629. PMID: 36308861.
8. Shazly S A, Trabuco E C, Ngufor C G, Famuyide AO. Introduction to Machine Learning in Obstetrics and Gynecology. Obstet Gynecol. 2022;139(4):669-679. doi: 10.1097/AOG.0000000000004706. PMID: 35272300.
9. Wang Y, Lang J, Zuo J Z, Dong Y, Hu Z, Xu X, et al. The radiomic-clinical model using the SHAP method for assessing the treatment response of whole-brain radiotherapy: a multicentric study. Eur Radiol. 2022;32(12):8737-8747. doi: 10.1007/s00330-022-08887-0. PMID: 35678859.
10. Rodríguez-Pérez R, Bajorath J. Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. J Med Chem. 2020;63(16):8761-8777. doi: 10.1021/acs.jmedchem.9b01101. PMID: 31512867.
11. Zhao S, Wang J, Jin C, Zhang X, Xue C, Zhou R, et al. Stacking Ensemble Learning-Based [^{18}F]FDG PET Radiomics for Outcome Prediction in Diffuse Large B-Cell Lymphoma. J Nucl Med. 2023;64(10):1603-1609. doi: 10.2967/jnumed.122.265244. PMID: 37500261.
12. Umans E, Dewilde K, Williams H, Deprest J, Van den Bosch T. Artificial Intelligence in Imaging in the First Trimester of Pregnancy: A Systematic Review. Fetal Diagn Ther. 2024;51(4):343-356. doi: 10.1159/000538243. PMID: 38493764; PMCID: PMC11318576.
13. Sciortino G, Tegolo D, Valenti C. Automatic detection and measurement of nuchal translucency. Comput Biol Med. 2017;82:12-20. doi: 10.1016/j.compbiomed.2017.01.008. PMID: 28126630.
14. Zhen C, Wang H, Cheng J, Yang X, Chen C, Hu X, et al. Locating Multiple Standard Planes in First-Trimester Ultrasound Videos via the Detection and Scoring of Key Anatomical Structures. Ultrasound Med Biol. 2023;49(9):2006-2016. doi: 10.1016/j.ultrasmedbio.2023.05.005. PMID: 37291008.
15. Tsai P Y, Hung C H, Chen C Y, Sun YN. Automatic Fetal Middle Sagittal Plane Detection in Ultrasound Using Generative Adversarial Network. Diagnostics (Basel). 2020;11(1):21. doi: 10.3390/diagnostics11010021. PMID: 33374307; PMCID: PMC7824131.
16. Nie S, Yu J, Chen P, et al. Automatic Detection of Standard Sagittal Plane in the First Trimester of Pregnancy Using 3-D Ultrasound Data. Ultrasound Med Biol, 2017,43(1):286-300. doi: 10.1016/j.ultrasmedbio.2016.08.034. PMID: 27810260.
17. Ji C, Liu K, Yang X, et al. A novel artificial intelligence model for fetal facial profile marker measurement during the first trimester. BMC Pregnancy Childbirth, 2023,

23(1):718. doi: 10.1186/s12884-023-06046-x. PMID: 37817098; PMCID: PMC10563312.
18. Sun Y, Zhang L, Dong D, et al. Application of an individualized nomogram in first-trimester screening for trisomy 21. Ultrasound Obstet Gynecol, 2021,58(1):56-66. doi: 10.1002/uog.22087. PMID: 32438493; PMCID: PMC8362158.
19. Zhang L, Dong D, Sun Y, et al. Development and Validation of a Deep Learning Model to Screen for Trisomy 21 During the First Trimester From Nuchal Ultrasonographic Images. JAMA Netw Open, 2022,5(6):e2217854. doi: 10.1001/jamanetworkopen.2022.17854. PMID: 35727579; PMCID: PMC9214589
20. Walker M C, Willner I, Miguel O X, et al. Using deep-learning in fetal ultrasound analysis for diagnosis of cystic hygroma in the first trimester. PLoS One, 2022,17(6):e269323. doi: 10.1371/journal.pone.0269323. PMID: 35731736; PMCID: PMC9216531.
21. Lin Q, Zhou Y, Shi S, et al. 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. Comput Methods Programs Biomed, 2022,226:107170. doi: 10.1016/j.cmpb.2022.107170. PMID: 36272307.
22. Xue H, Yu W, Liu Z, et al. Early Pregnancy Fetal Facial Ultrasound Standard Plane-Assisted Recognition Algorithm. J Ultrasound Med, 2023,42(8):1859-1880. doi: 10.1002/jum.16209. PMID: 36896480.
23. Drukker L, Noble J A, Papageorghiou A T. Introduction to artificial intelligence in ultrasound imaging in obstetrics and gynecology. Ultrasound Obstet Gynecol, 2020,56(4):498-505. doi: 10.1002/uog.22122. PMID: 32530098; PMCID: PMC7702141.
24. Xu J, Chen T, Fang X, et al. Prediction model of pressure injury occurrence in diabetic patients during ICU hospitalization--XGBoost machine learning model can be interpreted based on SHAP. Intensive Crit Care Nurs. 2024 Aug; 83:103715. doi: 10.1016/j.iccn.2024.103715. PMID: 38701634.
25. Fu Q, Wu Y, Zhu M, et al. Identifying cardiovascular disease risk in the U.S. population using environmental volatile organic compounds exposure: A machine learning predictive model based on the SHAP methodology. Ecotoxicol Environ Saf, 2024 Nov 1;286:117210. doi: 10.1016/j.ecoenv.2024.117210. PMID: 39447292.
26. He F, Wang Y, Xiu Y, et al. Artificial Intelligence in Prenatal Ultrasound Diagnosis. Front Med (Lausanne). 2021 Dec 16; 8:729978. doi: 10.3389/fmed.2021.729978. PMID: 34977053; PMCID: PMC8716504.
27. Espinoza J, Good S, Russell E, et al. Does the use of automated fetal biometry improve clinical work flow efficiency? J Ultrasound Med. 2013 May;32(5):847-50. doi: 10.7863/ultra.32.5.847. PMID: 23620327.
28. Yazdi B, Zanker P, Wanger P, et al. Optimal caliper placement: manual vs automated methods. Ultrasound Obstet Gynecol. 2014 Feb;43(2):170-5. doi: 10.1002/uog.12509. PMID: 23671025.
29. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics. 2020 Apr 1;21(2):345-352. doi: 10.1093/biostatistics/kxz041. PMID: 31742354.
30. Challen R, Denny J, Pitt M, et al. Artificial intelligence, bias and clinical safety. BMJ Qual Saf. 2019 Mar;28(3):231-237. doi: 10.1136/bmjqs-2018-008370. PMID: 30636200; PMCID: PMC6560460.