scientific reports

Check for updates

OPEN

Clinical validation of explainable AI for fetal growth scans through multi-level, cross-institutional prospective end-user evaluation

Zahra Bashir[1,2,3]✉, Manxi Lin[4], Aasa Feragen[4], Kamil Mikolaj[4], Caroline Taksøe-Vester[1,3,5], Anders Nymark Christensen[4], Morten B. S. Svendsen[3], Mette Hvilshøj Fabricius[2], Lisbeth Andreasen[6], Mads Nielsen[7] & Martin Grønnebæk Tolsgaard[1,3,5]

We aimed to develop and evaluate Explainable Artificial Intelligence (XAI) for fetal ultrasound using actionable concepts as feedback to end-users, using a prospective cross-center, multi-level approach. We developed, implemented, and tested a deep-learning model for fetal growth scans using both retrospective and prospective data. We used a modified Progressive Concept Bottleneck Model with pre-established clinical concepts as explanations (feedback on image optimization and presence of anatomical landmarks) as well as segmentations (outlining anatomical landmarks). The model was evaluated prospectively by assessing the following: the model's ability to assess standard plane quality, the correctness of explanations, the clinical usefulness of explanations, and the model's ability to discriminate between different levels of expertise among clinicians. We used 9352 annotated images for model development and 100 videos for prospective evaluation. Overall classification accuracy was 96.3%. The model's performance in assessing standard plane quality was on par with that of clinicians. Agreement between model segmentations and explanations provided by expert clinicians was found in 83.3% and 74.2% of cases, respectively. A panel of clinicians evaluated segmentations as useful in 72.4% of cases and explanations as useful in 75.0% of cases. Finally, the model reliably discriminated between the performances of clinicians with different levels of experience (p- values < 0.01 for all measures) Our study has successfully developed an Explainable AI model for real-time feedback to clinicians performing fetal growth scans. This work contributes to the existing literature by addressing the gap in the clinical validation of Explainable AI models within fetal medicine, emphasizing the importance of multi-level, cross-institutional, and prospective evaluation with clinician end-users. The prospective clinical validation uncovered challenges and opportunities that could not have been anticipated if we had only focused on retrospective development and validation, such as leveraging AI to gauge operator competence in fetal ultrasound.

Keyword Artificial intelligence, Fetal growth scans, Explainable AI, Human-AI collaboration

Fetal growth assessment is a crucial aspect of prenatal care, and ultrasound is the primary tool used for its evaluation. However, the accuracy of ultrasound examinations is highly operator dependent. Insufficient operator skills result in poor ultrasound images, which again impacts diagnostic accuracy when predicting birth weight[1].

Artificial intelligence (AI) techniques such as deep learning are now increasingly being used in other areas of medical imaging for automated image analysis and feedback to clinicians[2]. Within the field of obstetric imaging, deep learning has been used for a variety of applications, including anomaly prediction, standard plane detection, and automated caliper placement[3–6] Existing approaches have described how to automate tasks

[1]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [2]Department of Obstetrics and Gynecology, Slagelse Hospital, Fælledvej 11, 4200 Slagelse, Denmark. [3]Copenhagen Academy for Medical Education and Simulation (CAMES), Rigshospitalet, Denmark. [4]Technical University of Denmark (DTU), Lyngby, Denmark. [5]Center of Fetal Medicine, Dept. of Obstetrics, Copenhagen University Hospital, Rigshospitalet, Denmark. [6]Department of Obstetrics and Gynecology, Hvidovre Hospital, Hvidovre, Denmark. [7]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark. ✉email: zab@regionsjaelland.dk

normally performed by clinicians; however, relatively little attention has been invested in how deep learning can *augment*clinicians' performances through actionable feedback. This may be a consequence of the limited explainability offered by existing 'black box' approaches that provide a prediction but fail to explain why or how this decision was made. Lack of explainability makes it difficult for clinician end-users to understand and trust deep learning models when they cannot evaluate the information that was used to arrive at a certain decision[7–10] .This motivates using Explainable AI (XAI) to provide actionable and trustworthy feedback to improve clinicians' performances during fetal ultrasound scans.

Current approaches to validation of XAI models are often confined to retrospective datasets, within a limited number of hospitals, and often without any clinical evaluation of usefulness, correctness, or utility for patient care[11] Hence, there is an unmet need to validate and evaluate XAI models across multiple hospitals, with several groups of clinicians prospectively, to gain insights into opportunities and barriers when developing and implementing XAI models for fetal ultrasound in the future.

This study aimed to develop an explainable deep learning model for fetal growth scans using actionable concepts as feedback to clinician end-users. The model was developed using large amounts of retrospective ultrasound data, implemented in a prospective dataset, and evaluated by clinician end-users.

Methods

Study design and data source

We used a combination of retrospective and prospective data for developing, implementing, and testing a deep-learning model for fetal growth scans. The study was reported according to the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) guidelines[12].

Retrospective dataset

The retrospective dataset, used for developing and testing the model, included DICOM images from 17 Danish hospitals obtained during standard clinical care. We extracted images from 3rd-trimester ultrasound (US) that included the following image classes (Standard and nonstandard planes): transthalamic planes, transabdominal planes, and femur planes, with gestational age (GA) > = 28 weeks (> = 196 days). Regarding the ultrasound device, no exclusions were made based on device type; all data were included regardless of the ultrasound machine used. No exclusions were made for pathological cases or twin pregnancies. To improve the model's ability to distinguish between background and targeted plane images, an "Other" class was added to the training dataset. This class included 6010 two-dimensional background images from retrospective 2nd-trimester scans, such as spine, facial profile, placenta, feet, umbilical cord, amniotic fluid, aorta, bladder, and the 4-chamber view of the heart.

These images were paired with their corresponding maternal–fetal characteristics through the Danish Fetal Medicine Database (FØTO)[13]. The FØTO database contains information from all pregnant women scanned as part of standard antenatal care for pregnant women in Denmark.

Prospective data

In the prospective validation, we employed the Voluson E10 ultrasound machine for all scans.

The study involved 122 singleton pregnant women who were offered a growth scan from gestational week 28. Informed consent, which included an agreement to access their medical records and a full-length video recording of the ultrasound screen, was obtained before the US was performed.

The clinicians performing the growth scans included non-experts (specialist obstetricians and obstetrician trainees) and experts (fetal medicine specialists and sonographers) who performed obstetric ultrasounds at Slagelse Hospital and Copenhagen University Hospital Rigshospitalet, Denmark. All clinicians received a questionnaire containing general information about the operators' US experience, clinical title, and workplace name using REDCap electronic data capture tools[14].

Prospective US videos

A data collection unit was constructed for recording videos of ultrasound examinations. The recording was started when the operators initiated the examination, and the operator ended the recording when the scan was completed.

Data management

The recordings were continuously uploaded to a secure server at DTU Compute in an anonymized form. Information about the video recordings, such as the video name, patient gestational age, body mass index, and estimated fetal weight, were collected using REDCap. All ultrasound images were stripped of calipers used to perform measurements using the Telea inpainting method as detailed in[15].

Model development

The deep learning model used in this study was a modified Progressive Concept Bottleneck Model (PCBM)[16]. The PCBM is an intrinsically explainable model that provides two consecutive layers of explanation: the first layer consists of segmentations that give visual feedback to the clinicians on the anatomical structures visible in the image. The second layer consists of property concepts (explanations) such as symmetry, magnification, and visual quality of different anatomical concepts. The property concepts are scalar concepts (as known from the original concept bottleneck paper)[17] that quantify the degree to which the image satisfies the ISUOG criteria with respect to symmetry, magnification, and visual quality of the relevant anatomical concepts (e.g. femur bone endpoints). More detail can be found in the technical paper by Lin et al.[16].

We divided the retrospective dataset into training (n = 7527), validation (n = 662), and testing (n = 1163) sets on the subject level. The PCBM was trained to classify images into seven categories: femur standard plane, femur non-standard plane, standard transabdominal plane, transabdominal non-standard plane, transthalamic standard plane, transthalamic non-standard plane, and other class (not belonging to any of the considered anatomical planes).

Please see Supplementary Material 1 for a detailed description of the segmentation and explanations.

Appendix 1 shows the details of the PCBM architecture. Imitating the clinician's decision process during fetal image quality assessment, the model consisted of three stages.

In the first stage, the observer block, a DTU-Net[18] was used for image segmentation. The predicted segmentation from the observer was then concatenated and put into the second stage of the PCBM, the conceiver block.

In the second stage, the conceiver block consisted of three ResNet18 networks to predict the quality of structures, symmetry properties, and whether it is possible to place the calipers correctly. Caliper placement assesses whether it is possible to accurately position the calipers to measure key anatomical dimensions, such as biparietal diameter (BPD) and occipitofrontal diameter (OFD) in the transthalamic plane, ensuring that the measurements meet clinical standards. The predicted properties were then concatenated along with two additional properties, namely occupancy and angle, derived directly from the organ segmentations.

At the third stage of the PCBM, the concatenated property concepts were fed into a multi-layer perceptron, which we refer to as the predictor. The predictor outputs seven scores, corresponding to the seven categories defined above, which were converted to class probabilities using a softmax. The random seed was set to 42 throughout the experiments for reproducibility. The conceiver was initialized with pre-trained weights from ImageNet, while the remaining PCBM blocks were initialized with a uniform distribution. The models were implemented in Python 3.10.8, PyTorch 1.13.0, and CUDA 11.7. All experiments were done on a server with AlmaLinux 8.7 system and two Nvidia RTX A6000 GPUs.

The observer, conceiver, and predictor were trained with the stochastic gradient descent optimizer. The initial learning rate was set to 0.1 and was reduced by half once the evaluation loss stopped improving. The L2 regularization was set to 1e-4 to avoid over-fitting. The batch size was set to 32 for all three stages. The observer, conceiver and predictor were trained for 300, 500, and 100 epochs, respectively, where in each epoch, the model was trained on the training set and evaluated on the validation set. The model with the best performance on the validation set was selected as the best model. The observer was trained on the annotated segmentation ground truth with the target function described by Lin et al.[18] The conceiver was optimized on a combination of binary cross-entropy loss, Huber loss, and Hinge loss. The binary cross entropy loss was applied to the non-quality properties, while the Huber loss worked on the quality grades. The Hinge loss was included for optimizing the ranking of the predicted quality grades within a batch. A cross-entropy loss was employed for the predictor. We evaluated the three stages by IoU, classification accuracy as well as mean squared error, and classification accuracy, respectively.

Prospective validation of model performance and explanations

During inference on the prospective video recordings, classification confidence was used to rank frames. In our experimental validation, we used AI-based autocapture as a validation of the model's ability to quantify standard plane quality. Specifically, for each prospective video recording, we selected the frame with the highest confidence for the femur plane, transthalamic plane, and transabdominal plane as the AI-selected optimal standard plane. This approach allowed us to assess the model's ability to rank images by their standard plane quality compared to a clinician.

To make the quality assessment and auto-capture robust, the model classification confidence for each frame was smoothed with a mean filter of width 10. The frame with the highest smoothed confidence was then selected for each anatomy plane.

The model was validated prospectively in four steps, using 100 prospective videos that were not accessed during the development and training of the model.

Step 1. Assessing the model's ability to assess standard plane quality

In the first step of the validation, the AI's ability to assess standard plane quality was tested. A range of expert clinicians compared the quality of operator-picked images with that of AI auto-capture images for each ultrasound examination video. For each recorded video, manual and auto-capture images of the femur plane, transthalamic plane, and transabdominal planes (300 comparisons in total) were rated by groups of expert clinicians (sonographers and fetal medicine specialists) at Slagelse Hospital and Copenhagen University Hospital, Rigshospitalet (RH). The study involved 36 different expert clinicians from two hospitals, consisting of eight fetal medicine specialists and 28 sonographers. On average, 8–12 raters participated each day.

The expert clinicians were blinded to the test and asked to choose the best image among options A, B, or to indicate if both were of equal quality. Additionally, each participant noted whether Image A and B met the standard plane criteria as prescribed by the ISUOG guidelines[19]. Expert clinicians from Slagelse Hospital validated the images collected at RH (n = 84), while expert clinicians at RH validated those collected at Slagelse Hospital (n = 216). The ratings were completed over a period of 14 days, taking a total of 199 min.

To eliminate any indication of which images were captured by a clinician, calipers were removed from the images, similar to the process used for the training data. Additionally, we ensured that the color tone of the manual capture images and the AI auto-capture images from the video were matched. If necessary, color tones were adjusted using histogram matching[20].

Step 2. Assessing the correctness of the explanations
In the second step, the accuracy of both explanations and segmentation (see Fig. 2) provided by the model was validated by two expert clinicians. A total of 120 images , representing low, medium, and high quality categories, were presented for evaluation, with 40 images for each of the three planes (transthalamic, transabdominal, and femur). The two expert clinicians evaluated the correctness of the explanations scores and segmentation concepts in duplicates. Disagreements were resolved by discussion until consensus. The ISUOG standard plane criteria were used as a guiding reference[19].

Step 3. Assessing the clinical usefulness of model explanations
The third step, the clinical usefulness of the AI-auto-capture was evaluated by a diverse group of clinicians (sonographers, obstetricians, and trainees) from six hospitals in East Denmark. Each participant reviewed examples of high-quality and poor-quality images, along with their corresponding explanations and segmentations (4 images for each plane: transthalamic, transabdominal, and femur). Participants were then asked to indicate whether they found the model's explanations and segmentations useful for each image (See Appendix 2).

Step 4. Assessing the model's ability to discriminate between expert clinicians and non-expert clinicians
In the fourth step, we evaluated whether the model's confidence could be used to assess the competence level of the ultrasound clinicians. We compared expert clinicians (fetal medicine specialists and sonographers) to non-experts clinicians (obstetricians and trainees) by assessing the AI's confidence in the standard plane quality of the three best images chosen by the operator from each recorded video.

Statistics
The model performance was evaluated using precision, recall and the F1 score. For each class, we denoted true positive, true negative, false positive, and false negative with TP, TN, FP, and FN respectively. Precision was defined as TP / (TP + FP), and recall was defined as TP / (TP + FN). F1 score was defined as 2*(precision*recall)/(precision + recall). All distributions were inspected for normality and equality of variance, and in case of equality, the t-test was used, if not the Welch t-test was used. In case of non-normal distributions Wilcoxon or Mann–Whitney-U tests were used. For the statistical comparison of auto-captured and manual-captured images, we used mean and standard deviation (SD). We used Bonferroni correction of p-values to correct for multiple testing.

Statistical analyses were carried out using IBM SPSS Statistics (Statistical Package for Social Sciences, SPSS, USA) version 28 and SciPy version 1.11.1[21]..

Ethics
The retrospective data has been approved by The Danish Data Protection Agency (Protocol No. P-2019–310) and The Danish Patient Safety Authority (Protocol No. 3–3031-2915/1), exempting the need for informed consent. Approvals from the Danish Data Protection Agency (Protocol No. P-2021–570) have been obtained for the prospective validation and implementation. Informed consent was obtained from the included pregnant women for the collection of prospective data. The project was submitted to the Regional Ethics Committee, which has assessed that the study is exempt from The Scientific Ethical Treatment Act (jr. nr. 21,024,445). All methods were performed in accordance with the relevant guidelines and regulations.

Results
A total of 3342 images were manually labeled from the retrospective data, and we added 6010 unlabeled 'other' class images to eliminate false-positive segmentation. For prospective validation of our AI model, we recorded 122 full-length US growth scanning videos from n = 122 pregnant women at two different hospitals. Videos were between n = 2907 to n = 52,238 frames. See Fig. 1 for a flowchart. The demographic details are displayed in Table 1.

Fetal anomalies and twin pregnancies were not included in the prospective evaluation of the model; only singleton pregnancies were included. Out of the 100 included pregnancies, 9% of fetuses were classified as large for gestational age (LGA), and 17% were classified as small for gestational age (SGA), while the remaining 74% were of normal weight.

For the retrospective dataset, no exclusions were made for pathological children or twin pregnancies. Consequently, this dataset included 8.5% twins, 0.3% triplets, and 0.02% quadruplets, with the rest being singleton pregnancies. In terms of pathology, 17.9% were SGA, 7.7% were LGA, and the rest were within the normal weight range. Additionally, 10.6% of fetuses had a registered anomaly.

Out of the 100 US growth scans used for prospective validation, 22 were performed by non-expert clinicians, while the remaining 78 were performed by expert clinicians. For the prospective validation, Voluson E10 ultrasound machines were used. In the retrospective data, the distribution of devices used was as follows: 84.2% V830, 9.8% Voluson E10, 0.7% V730, 0.2% iU22, 0.2% Voluson S10, 0.09% Voluson S, 0.06% LOGIQ7 and 4.75% were unknown machines.

Model performance
The model performance is presented in Table 2. The overall classification accuracy achieved by the model was 96.29%.

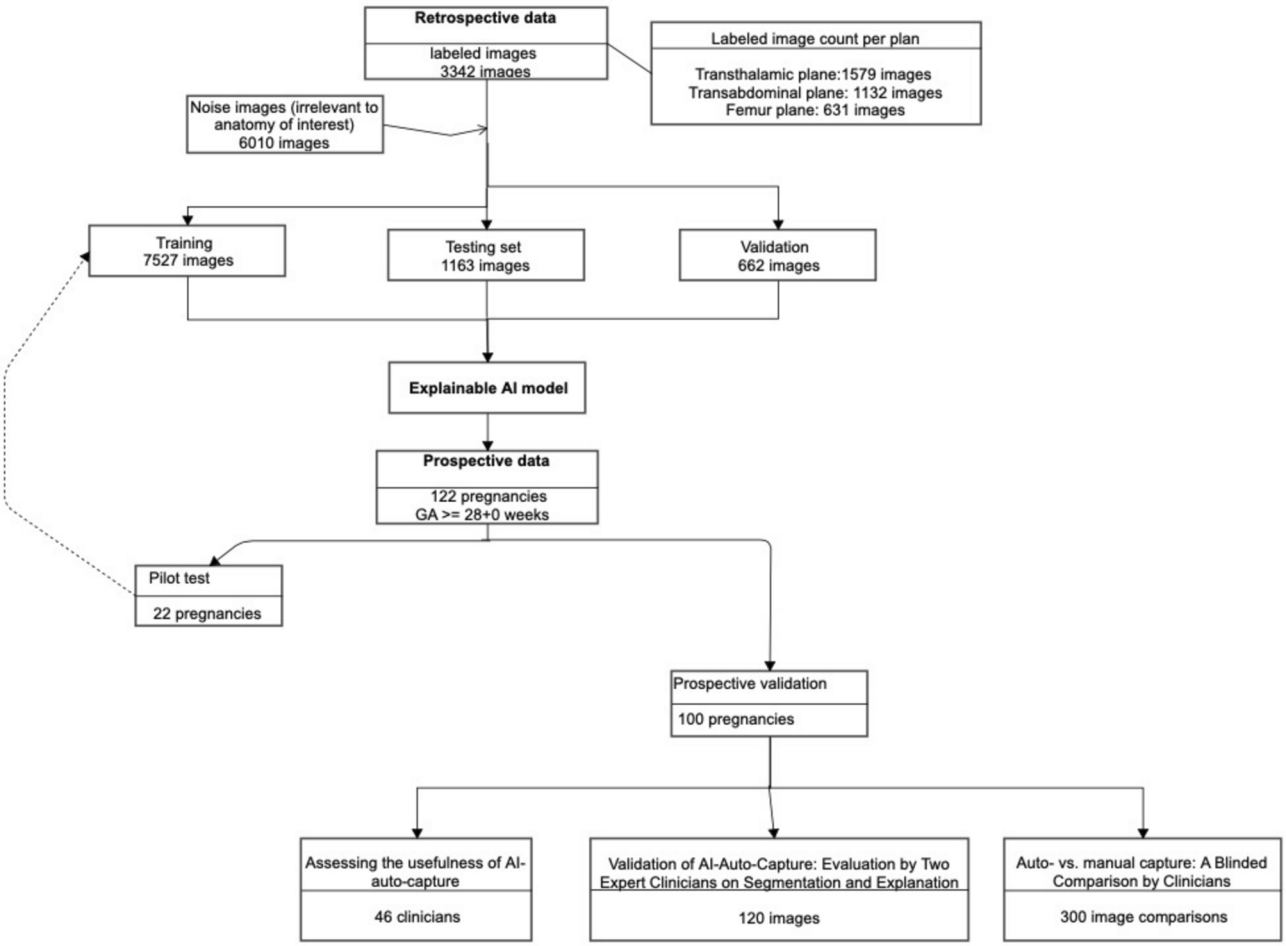

Fig. 1. A Flowchart Overview: Visualizing the model development and prospective validation. HC = Head circumference, AC = Abdominal circumference, FL = Femur length and GA = Gestational age. "Other" class = images not belonging to the anatomy of interest.

	Retrospective data		
	Labeled images	Other images	Prospective data
Patient, N	1831	3063	100
Image/Video number, N	3342	6010	100*
Age (Mean, SD), years	31.58 (5.26)	31.61 (5.21)	32.28 (5.35)
BMI (Mean, SD)	23.43 (4.74)	22.93 (4.21)	28.61 (7.25)
GA (Mean, SD), weeks	31.07 (4.87)	24·08 (2.62)	32.98 (3.30)

Table 1. It presents information about the retrospective and prospective data. The retrospective data includes image numbers, while the prospective data includes video numbers. (*) It should be noted that each video in the prospective data contains between 2907-52,238 frames. The BMI of women in the prospective data set was higher than for the women in the retrospective dataset.

Prospective validation

Step 1. Assessing the model's ability to assess standard plane quality

Preferences for auto-capture or manual-capture images The mean preference scores for both auto-capture and manual-capture images were calculated for all three planes, resulting in 22.8% (SD 2.5) preference for auto-capture, 24.5% (SD 2.7) for manual capture, and 52.6% (SD 3.1) for the equal quality.

To compare auto-capture and manual capture images, we summarized the raters' preferences (panel of expert clinicians) by assigning two scores to each pair of images, reflecting the proportion of votes preferring the auto-capture vs manual-capture images, respectively. When the raters considered both types of images to be of equal quality, a vote of 0.5 was added to both scores. A statistical comparison of the two modes of capture across the three anatomical planes (femur plane, transthalamic plane, transabdominal plane) is presented in Table

	Precision	Recall	F1 score
Femur SP	0.92	0.97	0.95
Transabdominal SP	0.73	0.87	0.79
Transthalamic SP	0.57	0.96	0.72
Femur NSP	0.97	0.91	0.94
Transabdominal NSP	0.95	0.90	0.92
Transthalamic NSP	0.99	0.88	0.93
Other	1	1	1

Table 2. Model performance. Precision, recall and F1 scores are standard validation metrics for classification, which are here computed as one class versus all. SP = Standard plane, NSP = non-standard plane.

	Auto-capture Mean (SD)	Manual captured Mean (SD)	P-values (Wilcoxon signed-rank test)
Femur plane	0.47 (0.20)	0.53 (0.20)	0.13
Transabdominal plane	0.55 (0.21)	0.45 (0.21)	0.04
Transthalamic plane	0.46 (0.19)	0.54 (0.19)	0.05

Table 3. Comparison of Best Auto-Capture Images and Manual-Captured Images for Each Plane, Presented as Mean and Standard Deviation (SD). The P-value ≤ 0.017 is considered statistically significant following the application of the Bonferroni correction.

	Auto-capture Mean (SD)	Manual captured Mean (SD)	P-values (Wilcoxon signed-rank test)
Femur plane	0.55 (0.35)	0.63 (0.31)	0.00067*
Transabdominal plane	0.40 (0.33)	0.35 (0.32)	0.018
Transthalamic plane	0.20 (0.24)	0.23 (0.24)	0.080

Table 4. Comparison of auto-capture and manual-captured standard plane images according to ISUOG standard plane criteria. The results were reported as mean and standard deviation (SD). The P-value ≤ 0.017 is considered statistically significant.

3. Using Bonferroni correction for multiple testing, we obtained a threshold of 0.017 to achieve a significance level of 0.05. Our analyses revealed no significant differences between auto-capture and manual-capture images. Furthermore, no significant differences were observed when we stratified the groups by the performance of expert and non-expert clinicians during the full-length growth scans.

Fulfillment of standard plane criteria for auto-capture and manual-captured images: As detailed in Table 4, manual-captured images were found to be standard planes significantly more often than auto-captured images for the femur plane. However, no significant differences were found for the other two anatomical planes, as presented in Table 4.

The agreement between two expert clinicians on the individual segmentation and explanations provided by the model is shown in Appendix 3. The expert clinicians agreed on all model segmentations and explanations in 83.3% (SD 0.37) and 74.2% (SD 0.44) of cases, respectively.

Step 3. Assessing the clinical usefulness of model explanations

A total of 46 clinicians, including 14 expert clinicians, 18 obstetricians, and 14 trainees provided feedback on the model's segmentation and explanation usefulness. According to the results presented in Table 5, the clinicians found the segmentation useful in 72.4% of cases, while the explanation was deemed useful in 75.0% of cases. A statistically significant difference was observed in the evaluation of segmentation and explanation usefulness between clinician groups, with obstetricians demonstrating a greater likelihood of perceiving the model's segmentation and explanation as useful compared to experts and trainees ($P < 0.001$ and $P < 0.004$, respectively).

See Fig. 2 for an example of model output.

Step 4. Assessing the model's ability to discriminate between expert clinicians and non-expert clinicians

The AI model's confidence level in manual capture images was assessed across groups of clinicians categorized by their expertise. The results revealed that expert clinicians consistently achieved significantly superior image quality in each plane compared to non-expert clinicians. The results are presented in Table 6. A P-value of ≤ 0.017 is considered statistically significant following the application of the Bonferroni correction. Significant differences were observed between expert and non-expert clinicians for the femur plane, transabdominal plane,

n = 12 images	Segmentation				Explanation			
	Useful	Neither nor	Not useful	Total N (%)	Useful	Neither nor	Not useful	Total N (%)
Clinicians (n = 46)	Observed votes N (%)	Observed votes N (%)	Observed votes N (%)		Observed votes N (%)	Observed votes N (%)	Observed votes N (%)	
Expert clinicians (n = 14)	115 (68.9)	18 (10.8)	34 (20.4)	167 (100)	133 (80.1)	17 (10.2)	16 (9.6)	166 (100)
Obstetricians (n = 18)	168 (77.8)	16 (7.4)	32 (14.8)	216 (100)	167 (77.7)	23 (10.7)	25 (11.6)	215 (100)
Trainees (n = 14)	116 (69.0)	22 (13.1)	30 (17.9)	168 (100)	112 (66.7)	24 (14.3)	32 (19.0)	168 (100)
Total N (%)	399 (72.4)	56 (10.2)	96 (17.4)	551 (100)	412 (75.0)	64 (12.0)	73 (13.0)	549 (100)

Table 5. Evaluations of the usefulness of model segmentations and explanations across different groups of clinicians. Experts clinicans are fetal medicine specialists and sonographers.

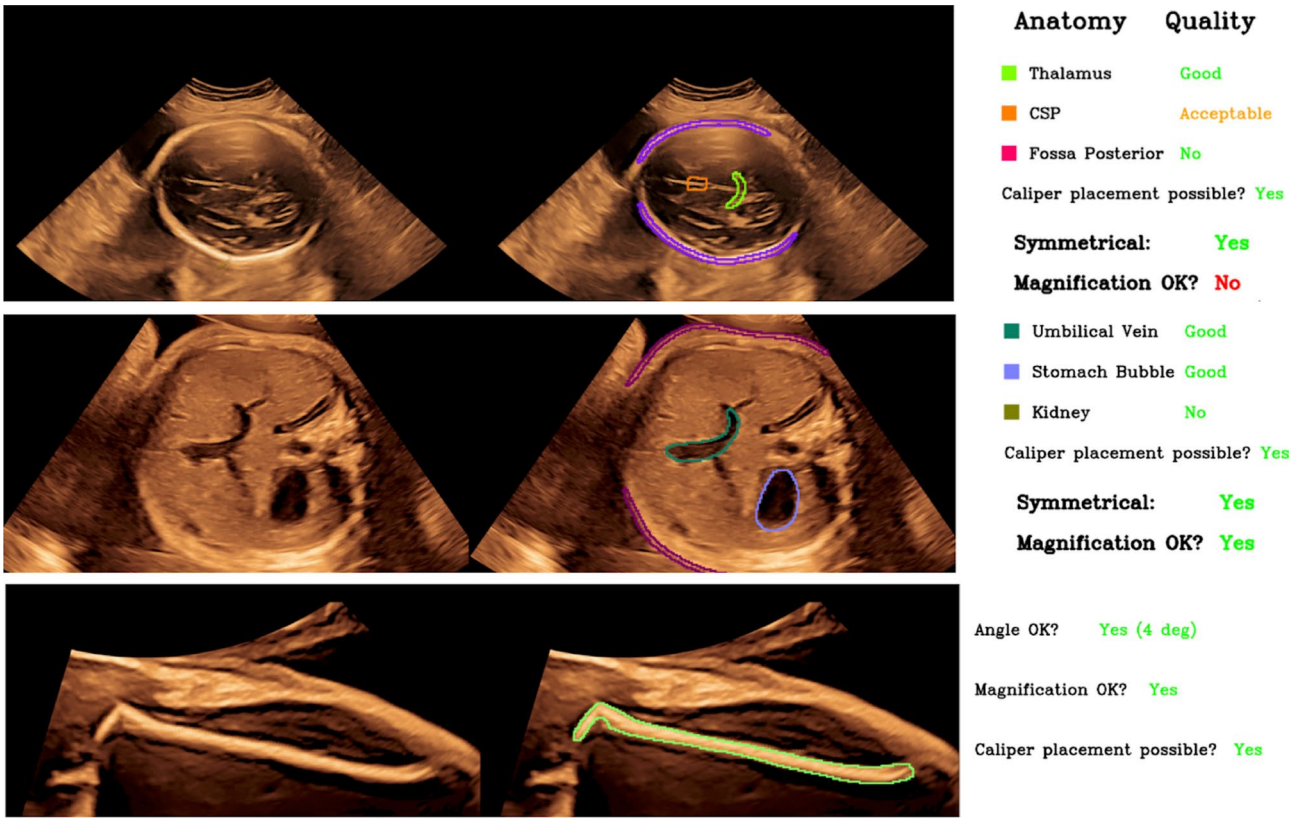

Fig. 2. Example of model output. The left image is the raw ultrasound image with segmentations in the middle and concept explanations to the right.

	Expert clinicians (n = 78)	Non-expert clinicians (n = 22)	
	Manual-capture Mean (SD)	Manual-capture Mean (SD)	P-value (Mann–Whitney U-test)
Femur plane	0.75 (0.30)	0.44 (0.41)	0.002
Transabdominal plane	0.34 (0.32)	0.13 (0.22)	0.007
Transthalamic plane	0.32 (0.31)	0.19 (0.24)	0.004

Table 6. Model confidence for scans performed by expert and non-expert clinicians. Expert clinicians are fetal medicine specialists and sonographers, while non-expert clinicians refer to specialist obstetricians and obstetrician trainees.

and transthalamic plane. In Appendix 4, we compare the auto-capture with manual capture images for clinicians with varying levels of experience. For Femur plane images, the expert clinicians selected better images than the auto-capture.

Discussion

In this study, we developed an AI model capable of assessing the quality of and providing feedback on standard plane images of the fetal transthalamic plane, transabdominal plane, and femur plane in real-time videos. Our study addresses the important gap in the existing literature[22–25], where few studies have included clinical validation of XAI models[26,27].While the majority of existing studies on deep learning have been retrospective[22], with limited efforts to apply these findings to the dynamic and unpredictable clinical environment using 'black-box' AI models, our study makes a substantial contribution. We developed our AI model using retrospective data from multiple hospitals and validated it on prospective full-length growth scan videos. Our study adds quality to the existing literature in this area[28,29] by involving a larger group of expert clinicians to validate the image quality preferences of the AI algorithms compared to those of clinicians. Further, it assesses the usefulness of XAI by end-users, evaluating explanations and feedback on relevant anatomical landmarks across multiple hospitals and among different health professionals with varying expertise (sonographers, fetal medicine experts, board-certified obstetricians, and trainees).

In the broader context of AI explainability, heatmap XAI techniques such as Grad-CAM (Gradient-weighted Class Activation Mapping)[30]have been popular for explaining the decisions of convolutional neural networks (CNNs), also in fetal ultrasound[31–33]. While these are useful for identifying important regions, it is generally recognized in the literature that they may face limitations in precision and contextual detail, which can affect their clinical relevance[9–11,31,33,34]. This issue is compounded by the fact that interpreting Grad-CAM outputs typically requires a certain level of technical expertise that many clinicians may not have. This gap between the complexity of the tool and the user's expertise can limit its practical application in a clinical setting[9,10,34]. These limitations underscore the need for XAI to be more closely aligned with the clinical process.

In addition to developing an XAI model that allowed insights into the model's predictions, our study is the first to use an AI model to measure clinician competence in the context of fetal ultrasound. Importantly, this offers new ways to assess and improve ultrasound competence by offering actionable feedback without the presence of clinician expert supervisors, which can be time-consuming and often infeasible beyond initial training[35,36]. The finding that image quality is both a meaningful and useful measure of feedback for clinicians and reflects differences in their competencies, supports the potential for collaboration between XAI and clinicians' underlying reasoning. Yet, the model's segmentations and explanations were perceived as useful in 74.5% and 74.3% of cases, respectively. We also noted that participants were more likely to find feedback and segmentations less useful when image quality was poor, whereas high-quality images led to more favorable evaluations. This observation aligns with recent research, such as the study by Taksoee-Vester et al., which demonstrated that clinician agreement with AI segmentation improves with better image quality and that clinicians generally outperformed the AI in low-quality image scenarios[37].

Obstetricians were more likely to find the model useful. Surprisingly, the usability score was lower among trainee clinicians compared to obstetricians and expert clinicians. However, this perception of usability does not necessarily reflect who actually benefits from the model feedback[38,39]. For example, a recent study in radiology demonstrated that experience, specialization, and prior use of AI tools are not reliable indicators of who benefits most from AI assistance[40].

Our study has several strengths, including testing the model on a large retrospective dataset before validating it on prospective full-length videos. We assessed the model's usability, the correctness of its segmentation and explanation, and its ability to measure clinicians' competence. Additionally, despite the majority of patients in the retrospective dataset being of normal weight, our AI model performed well on patients with a high BMI in the prospective data.

There are some limitations to our study. We did not assess the impact of our model on improving the quality of clinicians' scans or on clinical outcomes. Participation in the model's usability testing was voluntary, which may have introduced selection bias. Importantly, in our study design, the clinicians were unable to interact with the model explanations if they, for instance, wished to change one of the concept assumptions. This lack of interactivity with an XAI model may limit true human-AI collaboration. Future studies should explore how interaction effects between clinicians and AI expertise could enhance collaboration.

Conclusion

Our study has successfully developed an Explainable AI model for providing real-time feedback to clinicians performing fetal growth scans. This work contributes to the existing literature by addressing the gap in the clinical validation of Explainable AI models. It highlights the importance of multi-level, cross-institutional, and prospective evaluation with clinician end-users. The prospective clinical validation revealed challenges and opportunities that might not have been apparent through retrospective development alone, such as using AI to assess clinicians' competence in fetal ultrasound. Future research on Explainable AI in fetal medicine should include validation across a broad selection of clinician end-users to fully understand potential challenges and opportunities, thereby optimizing clinician-AI collaboration.

Data availability

We will provide data to support our findings upon reasonable request. However, we cannot share the entire dataset as this is protected by institutional regulations. However, the code can be shared after the publication of the

paper upon request under a non-commercial license. For inquiries regarding data availability and code , please write prof. Martin GrønnebækTolsgaard: martin.groennebaek.tolsgaard@regionh.dk.

Received: 3 March 2024; Accepted: 13 January 2025
Published online: 15 January 2025

References

1. Andreasen, L. A. et al. Why we succeed and fail in detecting fetal growth restriction: A population-based study. *Acta Obstet. Gynecol. Scand.* **100**, 893–899 (2021).
2. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
3. Bano, S. *et al.* (2021) AutoFB: Automating Fetal Biometry Estimation from Standard Ultrasound Planes. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **12907 LNCS** 228–238.
4. Chen, H. et al. Ultrasound standard plane detection using a composite neural network framework. *IEEE Trans. Cybern.* **47**, 1576–1586 (2017).
5. Płotka, S. et al. Deep learning fetal ultrasound video model match human observers in biometric measurements. *Phys. Med. Biol.* **67**(4), 045013 (2022).
6. Lin, M. et al. Use of real-time artificial intelligence in detection of abnormal image patterns in standard sonographic reference planes in screening for fetal intracranial malformations. *Ultrasound Obstet. Gynecol.* **59**, 304–316 (2022).
7. Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy* **23**, 1–45 (2021).
8. Petch, J., Di, S. & Nelson, W. Opening the black box: The promise and limitations of explainable machine learning in cardiology. *Can. J. Cardiol.* **38**, 204–213 (2022).
9. Pahud de Mortanges, A. et al. Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. *npj Digit Med.* **7**, 1–10 (2024).
10. Jung, J., Lee, H., Jung, H. & Kim, H. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. *Heliyon* **9**, e16110 (2023).
11. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet. Digit. Heal.* **3**, e745–e750 (2021).
12. Mongan, J., Moy, L. & Kahn, C. E. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers. *Radiol. Artif. Intell.* **2**, e200029 (2020).
13. Føto-databasen — DFMS. https://www.dfms.dk/new-page-33
14. Harris, P. A. et al. The REDCap consortium: Building an international community of software platform partners. *J. Biomed. Inform.* **95**, 103208 (2019).
15. Telea, A. An image inpainting technique based on the fast marching method. *J. Graph. Tools* **9**, 23–34 (2004).
16. Lin, M., Feragen, A., Bashir, Z., Tolsgaard, M. G. & Christensen, A. N. I saw, I conceived, I concluded: Progressive Concepts as Bottlenecks. *eprint arXiv* https://doi.org/10.48550/arXiv.2211.10630 (2022).
17. Koh, P. W. *et al.* Concept Bottleneck Models. *37th Int. Conf. Mach. Learn. ICML 2020* **PartF16814** 5294–5304 (2020)
18. Lin, M. et al. DTU-Net: Learning Topological Similarity for Curvilinear Structure Segmentation. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) **13939** LNCS, 654–666 (2023).
19. Salomon, L. J. et al. ISUOG practice guidelines: ultrasound assessment of fetal biometry and growth. *Ultrasound Obstet. Gynecol.* **53**, 715–723 (2019).
20. Tu, L. Dong, C. Histogram equalization and image feature matching. *Proc. 2013 6th Int. Congr. Image Signal Process. CISP 2013* **1** 443–447 (2013)
21. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
22. van den Heuvel, T. L. A., de Bruijn, D., de Korte, C. L. & van Ginneken, B. Automated measurement of fetal head circumference using 2D ultrasound images. *PLoS One* **4**, 1–20 (2018).
23. Li, J. et al. Automatic Fetal Head Circumference Measurement in Ultrasound Using Random Forest and Fast Ellipse Fitting. *IEEE J. Biomed. Heal. Inform.* **22**, 215–223 (2018).
24. Kim, B. et al. Machine-learning-based automatic identification of fetal abdominal circumference from ultrasound images. *Physiol. Meas.* **39**(10), 105007 (2018).
25. Kim, H. P. et al. Automatic evaluation of fetal head biometry from ultrasound images using machine learning. *Physiol. Meas.* **40**(6), 065009 (2019).
26. He, F., Wang, Y., Xiu, Y., Zhang, Y. & Chen, L. Artificial intelligence in prenatal ultrasound diagnosis. *Front. Med.* **8**, 1–9 (2021).
27. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 1–9 (2019).
28. Salim, I. et al. Evaluation of automated tool for two-dimensional fetal biometry. *Ultrasound Obstet. Gynecol.* **54**, 650–654 (2019).
29. Sarno, L. et al. Use of artificial intelligence in obstetrics: not quite ready for prime time. *Am. J. Obstet. Gynecol. MFM* **5**(2), 100792 (2023).
30. Selvaraju, R. R. et al. Grad-CAM: Why did you say that? *ArXiv* 1–4 (2016).
31. Lasala, A., Fiorentino, M. C., Bandini, A. & Moccia, S. FetalBrainAwareNet: Bridging GANs with anatomical insight for fetal ultrasound brain plane synthesis. *Comput. Med. Imaging Graph.* **116**, 102405 (2024).
32. Baumgartner, C. F. et al. SonoNet: Real-Time detection and localisation of fetal standard scan planes in freehand ultrasound. *IEEE Trans. Med. Imaging* **36**, 2204–2215 (2017).
33. Migliorelli, G. et al. On the use of contrastive learning for standard-plane classification in fetal ultrasound imaging. *Comput. Biol. Med.* **174**, 108430 (2024).
34. Jin, W., Li, X., Fatehi, M. & Hamarneh, G. Guidelines and evaluation of clinical explainable AI in medical image analysis. *Med. Image Anal.* https://doi.org/10.1016/j.media.2022.102684 (2023).
35. Tolsgaard, M. G. et al. Reliable and valid assessment of ultrasound operator competence in obstetrics and gynecology. *Ultrasound Obstet. Gynecol.* **43**, 437–443 (2014).
36. Tolsgaard, M. et al. When are trainees ready to perform transvaginal ultrasound?. *An Observational Study. Ultraschall Med.* **40**, 366–373 (2019).
37. Taksoee-Vester, C. A. et al. AI supported fetal echocardiography with quality assessment. *Sci. Rep.* **14**, 1–9 (2024).
38. Eva, K. W., Regehr, G. Gruppen, L. D. 2018 6. Blinded By "Insight": Self-Assessment and Its Role in Performance Improvement. in *Reconsidering Medical Education in the Twenty-First Century* (eds. Hodges, B. D. Lingard, L.) 131–154 (Cornell University Press). https://doi.org/10.7591/9780801465802-010
39. Taksøe-Vester, C. et al. Up or down? A randomized trial comparing image orientations during transvaginal ultrasound training. *Acta Obstet. Gynecol. Scand.* **97**, 1455–1462 (2018).
40. Yu, F. et al. Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nat. Med.* **30**, 837–849 (2024).

Acknowledgments
The project received support from The Local Research Fund for Næstved, Slagelse, and Ringsted Hospitals, as well as the Danish Regions' AI Signature Project. The funding organizations had no role in the project's execution, and the authors bear full responsibility for the project's content.

Author contributions
ZB, MT, LA, AF, AN and ML initiated and organized the research project. ZB and LA conducted literature searches. ZB, ML, MBS and CT were involved in the collection and management of data. ZB, ML, MT, AF and AN took part in data analysis and confirmed the integrity of the raw data. MT, MHF were involved in Funding acquisition. MT, AF, AN, LA, MN and MBS were involved in supervision. ZB, ML, MT, AF and LA wrote the manuscript. All authors participated in interpreting and analyzing the data, each reviewing, providing feedback on, and endorsing the final manuscript. All authors' collective responsibility extends to the manuscript's content and the choice to submit it for publication.

Funding
The Local Research Fund for Næstved, Slagelse, and Ringsted Hospitals, A1152 Danish Regions' AI Signature Project

Declarations

Competing interests
The authors declare no competing interests.

Disclosure
We used AI-assisted technology (GPT, Microsoft Word, Grammarly) to enhance the linguistic clarity of the paper. Our research group bears full responsibility for the content and ensures its accuracy in this paper.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-86536-4.

Correspondence and requests for materials should be addressed to Z.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.