npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01704-0

Predicting abnormal fetal growth using deep learning

Check for updates

Kamil Wojciech Mikołaj[1], Anders Nymark Christensen[1], Caroline Amalie Taksøe-Vester[2,3,4], Aasa Feragen[1], Olav Bjørn Petersen[3,4], Manxi Lin[1], Mads Nielsen[5], Morten Bo Søndergaard Svendsen[6] & Martin Grønnebæk Tolsgaard[2,3,4] ✉

Ultrasound assessment of fetal size and growth is the mainstay of monitoring fetal well-being during pregnancy, as being small for gestational age (SGA) or large for gestational age (LGA) poses significant risks for both the fetus and the mother. This study aimed to enhance the prediction accuracy of abnormal fetal growth. We developed a deep learning model, trained on a dataset of 433,096 ultrasound images derived from 94,538 examinations conducted on 65,752 patients. The deep learning model performed significantly better in detecting both SGA (58% vs 70%) and LGA compared with the current clinical standard, the Hadlock formula (41% vs 55%), $p < 0.001$. Additionally, the model estimates were significantly less biased across all demographic and technical variables compared to the Hadlock formula. Incorporating key anatomical features such as cortical structures, liver texture, and skin thickness was likely to be responsible for the improved prediction accuracy observed.

Ultrasound assessment of fetal size and growth is the mainstay of monitoring fetal well-being during pregnancy, as being too small (Small for Gestational Age, below the 10th percentile) or too large (Large for Gestational Age, above the 90th percentile) poses significant risks for both the fetus and the mother[1,2]. Specifically, the early identification of abnormal fetal growth is critical in mitigating the risk of stillbirth[3,4] and ensuring favorable perinatal outcomes[5]. Throughout the past 40 years, the standard practice for identifying abnormal fetal growth has been ultrasound-based estimates of fetal biometry using the Hadlock formula. Yet only about half of all SGA and LGA fetuses are correctly identified using this method, resulting in missed opportunities to mitigate the risk of adverse outcomes through appropriate obstetric management[6–9]. Moreover, the accuracy of the scans remains subject to considerable variability due to differences in patient characteristics and the quality of ultrasound equipment, thereby exacerbating existing health inequalities[10].

New deep learning technologies offer the potential to harness features from ultrasound images, providing more accurate fetal weight estimates than existing methods that are based on simple fetal biometry measurements[11]. Deep learning is a subset of artificial intelligence (AI) that is particularly well-suited for image analysis, especially when large-scale image and outcome data are available for training and testing these models. For example, deep learning has been used for standard plane classification[12], placenta segmentation[10], gestational age estimation[13], and neurosonographic diagnostic support[14] to mention a few. Existing approaches to fetal growth assessment have focused on automated biometry measurements[15,16], utilizing maternal characteristics[17,18], or a combination of maternal factors and fetal biometry measured by clinicians[19,20]. However, these approaches often fail to capture the extensive additional details present in ultrasound images, such as subcutaneous fat, organ texture, and fetal brain convolutions that may further inform growth estimates[21–23].

In this study, we aimed to investigate the potential of deep learning for improving the accuracy and precision of fetal weight estimation in clinical practice based on images alone. Our model was trained on population-wide data to estimate the fetal weight. We compared these estimates to the current clinical standard practice using the Hadlock formula[24] and various commonly used growth curves[25–27]. By also determining anatomical features of importance to the model estimate, we were able to point out directions for defining different sub-types and etiologies of abnormal fetal growth.

Results

We trained a deep learning model to estimate fetal weight based on 433,096 images from 94,538 examinations performed on 65,752 patients. The mean

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark. [2]Copenhagen Academy for Medical Education and Simulation (CAMES), Copenhagen, Denmark. [3]Center for Fetal Medicine, Copenhagen University Hospital Rigshospitalet, Copenhagen, Denmark. [4]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark. [5]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark. [6]Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark.
✉e-mail: martin.groennebaek.tolsgaard@regionh.dk

gestational age, in weeks, in the train and test sets was 30 ± 6.1 and 34 ± 3.2, respectively, and the mean gestational age at birth was 40 ± 1.6. The mean maternal age was 31.5 ± 5.26 years, BMI was 24.4 ± 5.6, 40.7% were nulliparous, and 59.3% were multiparous. Further baseline demographics for the training set can be found in Supplementary Table 1.

Detection rate and classification

The model significantly improved the sensitivity of SGA and LGA compared with Hadlock estimates based on clinician-performed fetal biometries. At the fixed specificity inherent to the Hadlock formula, the model's sensitivity for SGA was 70% (0.69, 0.71) vs. 58% (0.56, 0.59) for the Hadlock formula at a specificity of 91%. Similarly, for LGA detection, the model sensitivity was 55% (0.53, 0.57) vs. 41% (0.40, 0.43) for the Hadlock formula at a specificity of 96%. In Fig. 1 the ROC curve is shown for the model and Hadlock for the SGA and LGA groups.

Relative error

To validate weight predictions, birth weight is translated to fetal weight at scan time using one of three commonly used growth curves (Marsál, INTERGROWTH, and WHO). The relative errors between fetal weight and predictions obtained from the Hadlock formula and the model are shown in Table 1. The relative errors for the model and Hadlock are reported across demographic confounders such as BMI, parity, ethnicity, and ultrasound machine model in Table 2. These results reveal three significant observations. First, the model surpasses the Hadlock formula in all categories. Secondly, the model displays higher consistency, as indicated by smaller standard deviations. Finally, our model is more robust against demographic biases and other confounders than the Hadlock formula regardless of the growth curve used, with lower errors across all subgroups. These differences in performance between the two methods were found to be statistically significant, with a two-sample Welch's t-test yielding p-values < 0.0001 for SGA, AGA, and LGA, with effect sizes ranging from trivial for Appropriate for Gestational Age (AGA) to moderate for Large for Gestational Age (LGA).

Uncertainty estimation

We quantified the uncertainty associated with the growth estimates and observed a good fit for 98% of the data with predicted uncertainty in the range between 20 and 130 SD of 10 different estimates. We divided the estimated uncertainty into bins and found a linear relation between the standard deviation of the errors and the predicted uncertainty. An example of the model output is provided in Fig. 2. Further examples can be found in Supplementary Material E.

Analysis of pixel-level information

To assess whether the model used information beyond the fetal biometry used by the Hadlock formula, a total of 1800 heatmaps (see Supplementary Material E) (600 transthalamic, 600 transabdominal, and 600 femur images) were analyzed by two fetal medicine experts (MGT, CTV), Fig. 3. The analysis showed that certain anatomical features were of particular importance to the model predictions. For the transthalamic view of the fetal brain, we found that the brain area and cortical structures were of importance. For the transabdominal plane, we found that the texture of the liver and the skin was of importance. Finally, for the images of the fetal femur, the proximal and distal epiphyses as well as the subcutaneous thickness of the medial thigh were of importance. A complete overview can be found in Supplementary Table 3.

Discussion

We present a deep learning model for fetal growth estimation that, utilizing a comprehensive dataset, demonstrates superior sensitivity compared to existing best practices in identifying LGA and Small for Gestational Age (SGA) fetuses.

The clinical significance of improved detection of SGA and LGA fetuses is substantial, as abnormal growth remains a core challenge in obstetric management[4]. Notably, inaccuracies in fetal biometry measurements can significantly impact EFW, with even minor measurement errors potentially altering crucial clinical management decisions[28]. Our deep learning approach represents a valuable advancement in addressing these challenges and optimizing perinatal care.

To reduce variability in estimating fetal weight, strategies have included pinpointing pregnancies at risk, improving ultrasound proficiency among clinicians, and automating fetal biometry measurements[6,10,16,29]. Yet, these methods depend on a limited set of variables (risk factors, biometric measurements), neglecting other potentially significant features that could improve predictive performance. Our deep learning model surpasses current clinical practice, likely because it integrates a broader array of pixel-level features from ultrasound images than conventional fetal biometrics. Saliency maps revealed that our model identifies key anatomical features for accurately predicting fetal weight. These features, including subcutaneous fat[30,31], liver texture[23], and cortical structures[22] have been explored in previous smaller studies primarily using MRI or 3D ultrasound[32]. However, their integration into routine clinical practice has been hindered until now by the need for simplistic, time-efficient measures and by the extensive training required for clinicians when introducing new measurement practices. In this study, we apply deep learning algorithms to automate the analysis of pixel-level information from key anatomical features, enhancing

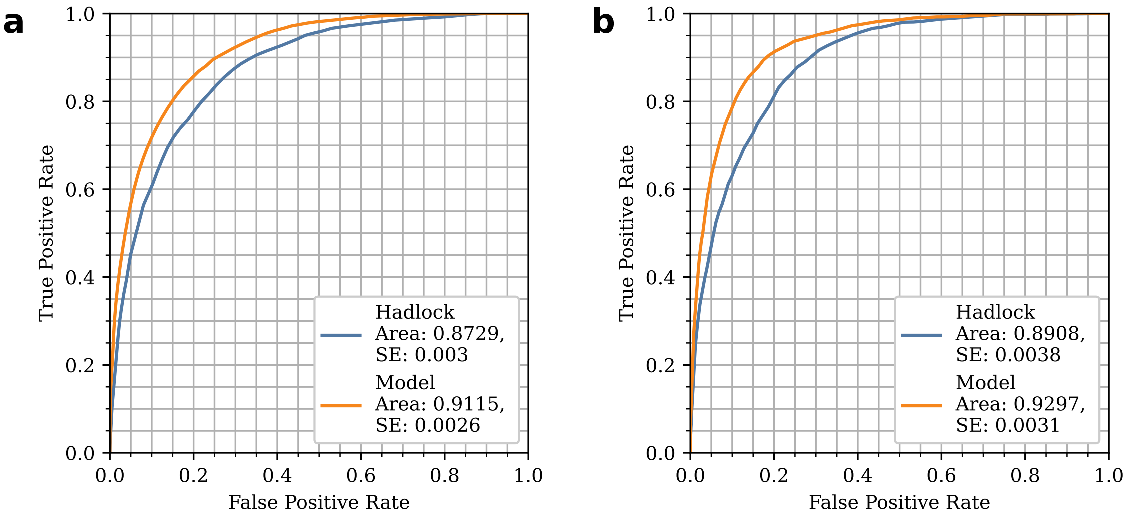


Fig. 1 | ROC curve analysis. a, b ROC curve comparison for classification of (**a**) Small for Gestational Age (SGA) and (**b**) Large for Gestational Age (LGA). The total number of samples was 31,386, of which 6152 were SGA and 3270 were LGA.

their predictive value without additional time, imaging, or training requirements for clinicians. This advancement not only streamlines the diagnostic process but also enables the exploration of specific phenotypes associated with abnormal fetal growth, linking these directly to their underlying pathophysiology. For instance, better identification of fetuses with excessive subcutaneous fat may improve detection rates of maternal gestational diabetes mellitus (GDM) even in the absence of LGA[33]. Similarly, identifying subtypes of SGA growth patterns could help differentiate between fetuses at risk of adverse outcomes and those who are constitutionally small but healthy.

There are other anatomies that may hold even greater potential for predicting accurate fetal weight, such as subcutaneous thickness of the fetal humerus or the fetal pancreas volume[34,35]. However, in this study, we focused on comparing a deep learning based approach to estimating fetal weight without introducing additional requirements to the planes used for the growth scan.

Additionally, our method addresses the frequently neglected aspect of uncertainty in fetal weight estimation, a factor that can lead to unwarranted diagnostic tests, obstetric procedures, and subsequent ultrasound assessments, ultimately inflating healthcare expenses[36]. By offering uncertainty metrics alongside weight predictions, our model supports more nuanced clinical decision-making. This enhancement allows clinicians to assess management strategies with a clear understanding of the certainty level regarding the fetal size relative to gestational age, prompting a specialist follow-up if the uncertainty is high. Moreover, having the opportunity to share uncertainty in these estimates with patients may contribute to shared decision-making and improved patient involvement in management strategies. However, it is important to note that our approach does not accurately measure error for different images within the same examination. Addressing this limitation requires further research aimed at identifying

Table 1 | Mean Relative Error of Estimated Fetal Weight Using the Hadlock Formula and the Model

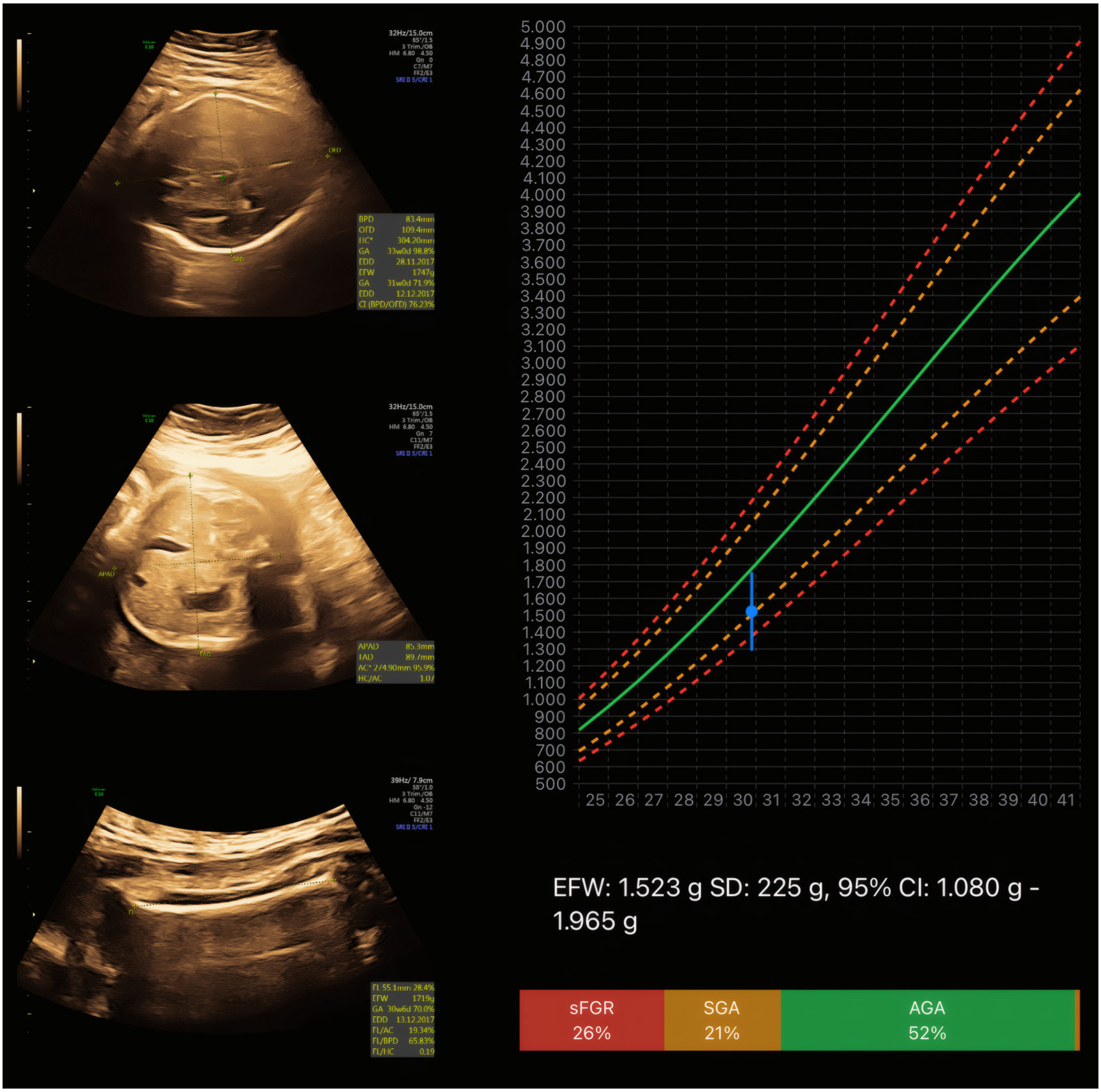
	SGA (*N* = 6152)	AGA (*N* = 21,964)	LGA (*N* = 3270)
Hadlock MRE [%]	9.12 ± 7.68	7.14 ± 5.38	9.57 ± 6.71
Model MRE [%]	7.31 ± 6.47	6.49 ± 5.02	7.31 ± 5.71
p-value	< 0.0001	< 0.0001	< 0.0001
Effect size—Cohen's d	0.25	0.12	0.36

Mean Relative Error (MRE) of Estimated Fetal Weight (EFW) ± standard deviation using Hadlock formula and Model on the test set. Columns denote 3 groups - Small for Gestational Age (SGA), Appropriate for Gestational Age (AGA), and Large for Gestational Age (LGA) for those with birthweights below, between, and above the 10th and 90th percentile, respectively.

Table 2 | Comparison of model and Hadlock, using three different growth curves: Maršál, Intergrowth, and WHO

			Maršál			Intergrowth			WHO		
Strat	**val/range**	***N***	**Model**	**Hadlock**	***p*-val**	**Model**	**Hadlock**	***p*-val**	**Model**	**Hadlock**	***p*-val**
All	–	31386	**6.74** ± **5.42**	7.78 ± 6.12	<0.0001	**7.71** ± **6.28**	8.73 ± 6.98	<0.0001	**6.86** ± **5.74**	7.90 ± 6.41	<0.0001
GA	196–224	7301	**8.29** ± **6.32**	9.12 ± 6.70	<0.0001	**9.00** ± **7.26**	9.59 ± 7.52	<0.0001	**8.22** ± **6.71**	8.70 ± 6.86	<0.0001
	224–259	18296	**6.47** ± **5.18**	7.53 ± 6.07	<0.0001	**7.57** ± **6.12**	8.79 ± 7.07	<0.0001	**6.68** ± **5.53**	7.86 ± 6.47	<0.0001
	259–294	5789	**5.61** ± **4.44**	6.89 ± 5.20	<0.0001	**6.49** ± **5.06**	7.44 ± 5.71	<0.0001	**5.70** ± **4.62**	7.02 ± 5.42	<0.0001
BMI	<18.5	3000	**6.84** ± **5.57**	7.46 ± 5.81	<0.0001	**7.57** ± **6.26**	8.25 ± 6.79	<0.0001	**7.13** ± **6.09**	7.74 ± 6.37	0.0001
	18.5–25	15063	**6.60** ± **5.37**	7.60 ± 6.03	<0.0001	**7.41** ± **6.07**	8.49 ± 6.81	<0.0001	**6.80** ± **5.66**	7.78 ± 6.31	<0.0001
	25–30	7192	**6.72** ± **5.37**	8.09 ± 6.05	<0.0001	**7.81** ± **6.38**	9.00 ± 6.89	<0.0001	**6.66** ± **5.62**	7.88 ± 6.11	<0.0001
	30–35	4166	**6.78** ± **5.29**	8.03 ± 6.47	<0.0001	**7.90** ± **6.43**	9.02 ± 7.34	<0.0001	**6.81** ± **5.57**	8.19 ± 6.80	<0.0001
	35 <	1965	**7.55** ± **5.96**	8.03 ± 6.70	0.0184	9.36 ± 6.90	9.61 ± 7.97	0.2947	**7.74** ± **6.48**	8.49 ± 7.38	0.0007
Parity	1	12995	**6.67** ± **5.40**	7.74 ± 6.26	<0.0001	**7.66** ± **6.25**	8.71 ± 7.12	<0.0001	**7.14** ± **6.00**	8.29 ± 6.87	<0.0001
	1 <	18391	**6.79** ± **5.43**	7.82 ± 6.02	<0.0001	**7.74** ± **6.31**	8.73 ± 6.89	<0.0001	**6.66** ± **5.54**	7.63 ± 6.05	<0.0001
Conception	FER	49	**4.28** ± **3.18**	8.41 ± 7.60	0.0008	5.68 ± 4.38	6.24 ± 8.43	0.6825	5.52 ± 4.39	6.17 ± 6.90	0.5790
	ICSI	512	**5.99** ± **4.96**	7.49 ± 5.36	<0.0001	**6.68** ± **5.60**	8.49 ± 5.85	<0.0001	**6.37** ± **5.56**	8.47 ± 5.90	<0.0001
	IVF	1133	**7.01** ± **5.87**	8.41 ± 6.27	<0.0001	**8.49** ± **7.27**	9.88 ± 7.93	<0.0001	**6.89** ± **5.87**	8.51 ± 6.57	<0.0001
	Insemination	669	6.24 ± 5.01	6.65 ± 5.30	0.1450	7.63 ± 6.37	8.12 ± 6.81	0.1761	6.59 ± 5.21	7.01 ± 5.86	0.1675
	Ov. Stim.	180	**6.01** ± **4.89**	8.68 ± 5.70	<0.0001	7.80 ± 6.55	8.99 ± 6.88	0.0954	**5.89** ± **4.62**	7.47 ± 5.33	0.0030
	Spontaneous	28385	**6.75** ± **5.41**	7.79 ± 6.14	<0.0001	**7.69** ± **6.23**	8.70 ± 6.96	<0.0001	**6.87** ± **5.74**	7.89 ± 6.42	<0.0001
	Unknown	458	7.24 ± 6.17	7.48 ± 6.25	0.5529	8.06 ± 7.01	8.59 ± 6.77	0.2435	7.65 ± 6.55	8.05 ± 6.73	0.3676
Ethnicity	Afro-Caribbean	424	6.69 ± 5.42	7.45 ± 6.16	0.0573	**7.94** ± **6.56**	9.20 ± 7.10	0.0075	**6.60** ± **5.74**	7.86 ± 6.80	0.0038
	Asian	1072	**6.72** ± **5.29**	7.81 ± 5.66	<0.0001	**7.77** ± **6.00**	9.10 ± 6.77	<0.0001	**7.06** ± **5.63**	8.01 ± 6.35	0.0002
	Caucasian	27053	**6.73** ± **5.40**	7.83 ± 6.18	<0.0001	**7.77** ± **6.30**	8.79 ± 7.07	<0.0001	**6.83** ± **5.71**	7.91 ± 6.45	<0.0001
	Oriental	582	**6.45** ± **4.53**	8.28 ± 5.65	<0.0001	**6.93** ± **5.67**	8.93 ± 6.09	<0.0001	**5.98** ± **4.54**	7.67 ± 5.42	<0.0001
	Unknown	2255	6.85 ± 5.95	7.16 ± 5.67	0.0804	7.11 ± 6.26	7.57 ± 6.07	0.0954	7.41 ± 6.38	7.79 ± 6.09	0.0420
Machine Model	V730	131	7.53 ± 8.39	7.71 ± 5.81	0.8412	8.91 ± 9.39	9.72 ± 7.20	0.4329	8.86 ± 9.54	9.86 ± 6.54	0.3245
	V830	29572	**6.75** ± **5.40**	7.82 ± 6.14	<0.0001	**7.71** ± **6.28**	8.75 ± 7.00	<0.0001	**6.87** ± **5.73**	7.94 ± 6.43	<0.0001
	Voluson E10	1097	**6.16** ± **5.49**	6.80 ± 5.61	0.0078	7.49 ± 6.37	8.01 ± 6.86	0.0659	6.73 ± 5.89	6.95 ± 5.96	0.3908
	Voluson S	458	**6.68** ± **5.05**	7.66 ± 6.13	0.0083	**7.05** ± **4.99**	8.25 ± 5.89	0.0009	**6.29** ± **4.90**	7.31 ± 5.56	0.0032
	Voluson S10	128	7.15 ± 5.55	8.05 ± 6.33	0.2299	8.59 ± 6.51	9.41 ± 7.53	0.3501	6.81 ± 4.97	7.63 ± 6.33	0.2473

All numbers are in % ± SD. The significance level was adjusted using the Benjamini-Hochberg procedure[46] and significant values are highlighted in bold.



EFW: 1.523 g SD: 225 g, 95% CI: 1.080 g - 1.965 g

Fig. 2 | Example model output for clinical decision-making. The blue dot is the estimated weight, and the blue line is the standard deviation. The green line is the growth curve, and the orange and red dotted lines are the 10/90% percentile and the 3/97% percentile, respectively. EFW Estimated Fetal Weight, with standard deviation and 95% confidence interval. In the bottom is shown the estimated probabilities based on the uncertainty, sFGR Severe growth restriction (below 3% percentile), SGA Small for Gestational Age (below 10% percentile), AGA Average for Gestational Age.

combinations of images that minimize diagnostic error during the dynamic nature of an ultrasound examination.

Deep learning models have been recognized for their superiority over clinician-based diagnostics in various domains[37], yet concerns about these models amplifying existing biases in care delivery persist[38]. Another issue is the potential for poor generalization of deep learning models across diverse demographic groups and different ultrasound equipment types. To address these concerns, we evaluated our AI model's performance against a range of demographic factors, including parity, BMI, age, gestational age, ethnicity, and mode of conception. In every scenario, our model exhibited lower errors compared to the Hadlock formula, irrespective of the growth chart applied (Marsal, Intergrowth, WHO). Additionally, the model consistently outperformed the Hadlock formula, regardless of the ultrasound equipment used. While the model's generalizability to other populations requires further exploration, we have previously shown that retraining fetal ultrasound models on smaller, diverse datasets may overcome such performance reductions[12,39,40].

We evaluated our model and the Hadlock formula based on how well they predicted birth weight back-calculated to the time of the scan. The precision is therefore lower the longer the interval from date of scan to date of birth, highlighting that abnormal growth can occur in this interval for multiple different reasons. Consequently, the true weight at the time of the scan is difficult to obtain in a reliable fashion. Still, the main priority when estimating fetal weight is to identify those that over the course of pregnancy, are at risk of adverse events due to abnormal fetal growth.

We demonstrate that deep learning improves the sensitivity of ultrasound-based screening for abnormal fetal growth over current clinical standards by 20% for SGA and 34% for LGA. Our model is particularly

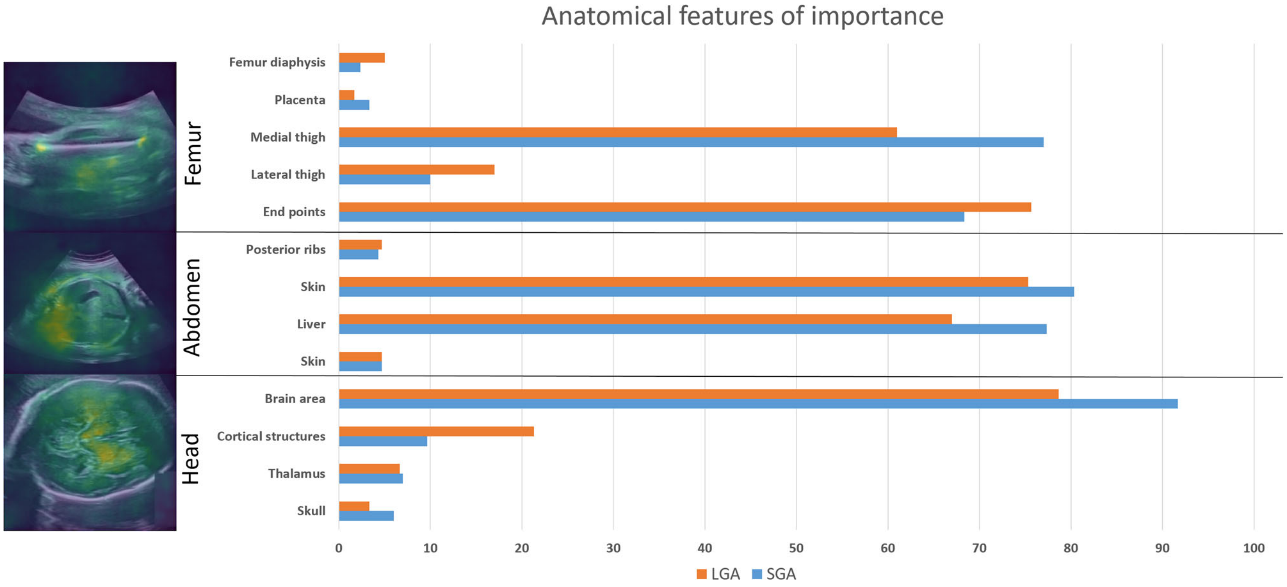


Fig. 3 | Anatomical features of importance. Frequency in % of structures found in the GradCAM heatmaps (see Supplementary Material E), i.e., used by the model. Examples of heatmaps shown on the left. There were no differences in anatomical regions of importance between LGA and SGA fetuses. Image optimization such as gain settings, did not seem to influence the annotations of anatomies.

effective in mitigating demographic biases, offering a more universally applicable tool. Incorporating pixel-level image data, including key anatomical features such as cortical structures, liver texture, and subcutaneous fat, significantly enhanced the model's sensitivity beyond traditional regression models. Looking forward, leveraging these anatomical features could enhance our ability to pinpoint different growth abnormalities and their causes.

Methods

Study design

In this retrospective, multi-center cohort study, we used a deep learning model for estimating fetal weight based on ultrasound images obtained across 17 hospitals in Denmark between 2008 and 2018. Birth weight data were obtained through the Danish Fetal Medicine Database, and imaging data were collected from four central servers. The Danish Patient Safety Authority, Islands Brygge 67, 2300-Copenhagen, Denmark, waived patient consent for this study (Record No. 3-3013-2915/1), and the Danish Data Protection Agency, Carl Jacobsens Vej 35 2500-Valby, Denmark, approved the study (Protocol No. P-2019-310).

The Hadlock formula is based on measurements of Abdominal Circumference (AC), Head Circumference (HC), and Femur Length (FL) performed by the clinician during the scan. The measurements were obtained automatically using Optical Character Recognition (OCR). For more detail refer to Supplementary Material G.

The confidence intervals and standard errors for the Receiver Operating Characteristic (ROC) curves were calculated using the Hanley method[41,42].

Dataset

Images used for fetal biometry measurements often come with embedded markings placed by clinicians during the scan. One example of such marking can be seen in Supplementary Fig. 6a, where calipers (yellow crosses) are placed on the picture to outline the anatomy to be measured, and the result is placed in a table in the lower-right corner. The table contains the value and the code of what is being measured: FL, AC, HC, and Biparietal Diameter (BPD).

These measurements were performed by the sonographers and clinicians on the three standard ultrasound planes required for estimating fetal weight and served as an input to the Hadlock formula. All 17 hospitals follow the international criteria for obtaining standard planes (ISUOG criteria for 3rd trimester ultrasound), and measurements are performed according to national guidelines (dfms.dk).

Optical Character Recognition (OCR) based on Tesseract was used to automatically classify the images as head, abdomen, femur, and other and extract the relevant measurements. The "other" class was discarded. Next, the images were aggregated based on the patient's identification number and study date to obtain sets of images from the same examination. Furthermore, sets that did not have at least one image from each class were excluded. Lastly, fetal weight at scan time was extrapolated from the birth weight using the Marsal growth curve[25].

The data was limited to singleton pregnancies only and was divided on a patient basis between training, validation, and test sets (85%,5%,10%), ensuring no patient overlap. In the event that multiple images of each anatomical region (femur, abdomen, head) were obtained during an examination, multiple observations were created by generating all feasible permutations of the available images. The training included 27% of 2nd-trimester images to increase the amount of training data. However, the test set contains only 3rd-trimester images with gestational age above 28 weeks. Before training, the calipers were removed from the images to avoid shortcut learning, see Supplementary Material G.

The standard deviation of the fetal weight is not fixed and varies as a function of the weight itself. Similarly, as in Maršál et al., it was set to 12%. Therefore, the standard score is calculated $z = \frac{x-\mu}{0.12\mu}$. Moreover, fetuses with a fetal weight below the 10th percentile ($z < -1.282$) are referred to as SGA, while fetal weights above 90th ($z > 1.282$) are referred to as LGA. Normal weight fetuses are referred to as AGA.

Model

The model used in this study was based on RegNetX 400 Mf[43] and was comprised of two distinct parts. The first part of the model processed the images to generate a measurement of the input anatomical structure and an embedding vector that corresponded to this structure. This vector enabled the model to encode additional information about the input images beyond the measurements. The entire model was composed of three subnetworks, each responsible for processing a different standard plane (head, abdomen, femur).

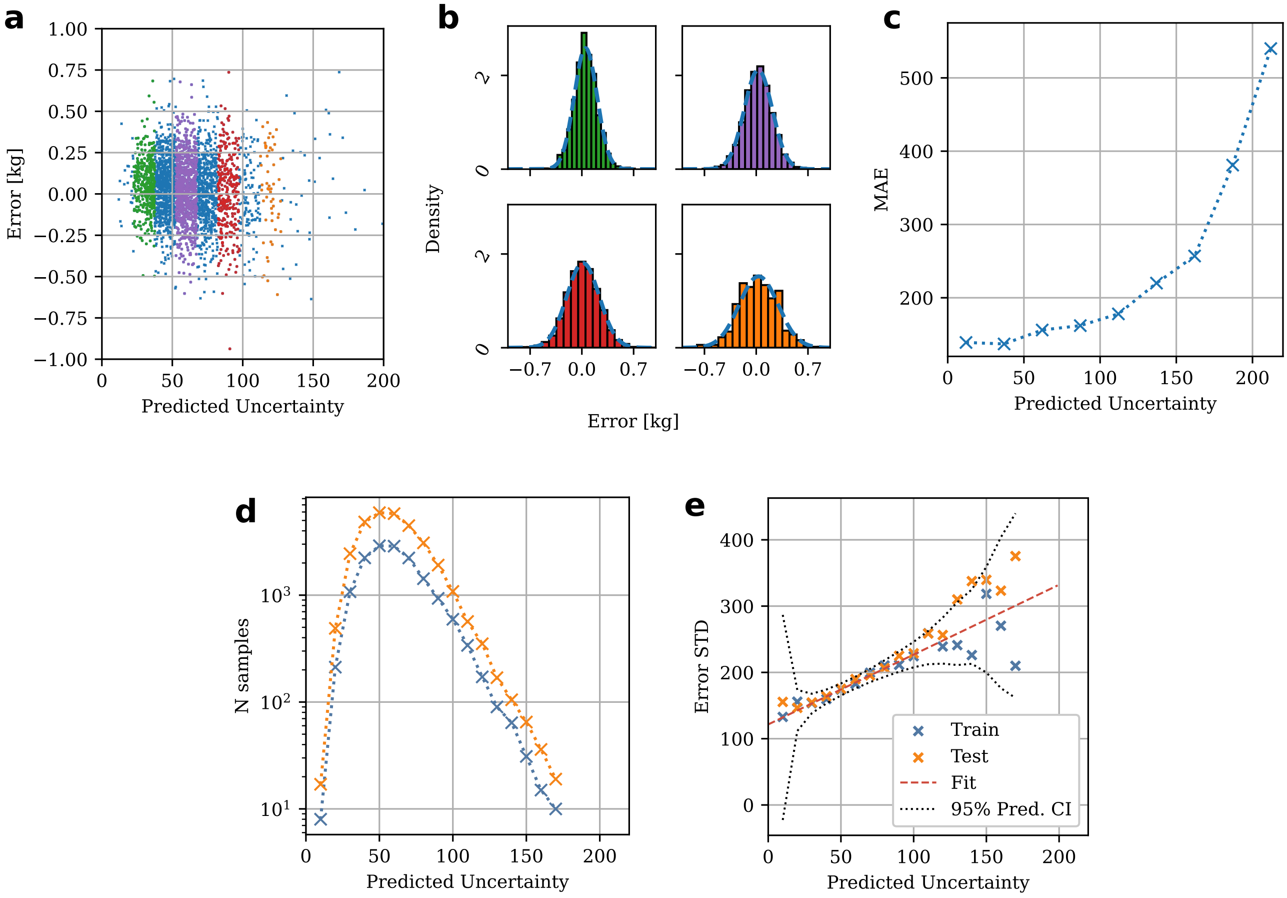


Fig. 4 | High predicted uncertainty correlates well with predictive error. **a** Predicted uncertainty is plotted vs. prediction error and binned into levels of predicted uncertainty, indicated by color. **b** For every second bin from (**a**), indicated by color, we show how higher predictive uncertainty comes with a broader distribution of predictive errors. **c** The Mean Absolute Error (MAE) grows with increased predicted uncertainty. **d**, **e** The STD of the prediction error (**e**) matches the smaller samples (**d**) found in the same predicted uncertainty bins.

The second part of the model was composed of two fully connected layers that accepted the predicted measurements and embedding vectors. The output was the Estimated Fetal Weight (EFW). The block diagram of the model can be seen in Supplementary Material C.

Lastly, the anatomy presented on an ultrasound image can vary in scale depending on the zoom level chosen by the operator. To alleviate this problem, pixel spacing (spatial resolution) was input into the first part of the model to provide information about the relative scale of the image. This parameter is saved in the DICOM files exported from the ultrasound machine.

Training

The models were trained using the AdamW optimizer with a learning rate of 1e-4, weight decay of 1e-6, and batch size of 8. To reduce training time, the RegNetX parameters obtained from training on ImageNet data[44] were used as a starting point. The training images were center cropped and resized to 224 × 224px, converted to grayscale, and further augmented with random rotation (±25°), shear (±10°); translation (0.05 of image size); brightness (0.2), contrast (0.2), and random horizontal flip ($P = 0.5$). The model was trained using a multi-task learning scheme to output the measurements: HC, BPD, AC, FL, and EFW. Images as well as all measurements were normalized to fit 0 to 1 interval.

In our study, we incorporated an additional weighting parameter into the loss function used for estimating fetal weight predictions. Specifically, we utilized relative error as the base loss function and added a weighting parameter based on the z-score to further emphasize the loss on abnormal fetuses. This is illustrated in Equation (1). The same loss function was also utilized for the measurement predictions.

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} \cdot (0.5 + |z_i|) \qquad (1)$$

Where:

y_i = fetal weight based on Maršál growth curve[25]
$\hat{y}_i$ = Estimated Fetal Weight (EFW)
z_i = fetal weight z-score

The training dataset is organized such that each unique scan corresponds to one entry, but the scans can contain more than one image of each standard plane. Therefore, during training, a set of three images (head, abdomen, and femur) is randomly sampled from each scan.

Uncertainty estimation

Test time augmentation[45] was used to estimate prediction uncertainty. Each set of images was augmented 10 times and passed through the model to obtain multiple predictions for the fetal weight; the standard deviation of the predictions is used as the initial uncertainty estimate. The augmentation parameters used in this step were the same as the ones used in training.

Values obtained in this way correlate with the prediction errors, as detailed in Fig. 4: Figure 4a shows the distribution of errors. Moreover, to evaluate how the prediction error changes as a function of predicted uncertainty, the data was divided into bins with a width of 10. In Fig. 4a, 4 out of the 22 bins are highlighted in color, and Fig. 4b shows the distribution of errors in those same bins. Notice that the errors in bins are normally distributed and that the standard deviation increases as the uncertainty increases. Additionally, Fig. 4c shows the mean absolute error.

Next, the error standard deviation in each bin was paired with the mean predicted uncertainty in that bin. Using the number of samples, shown in Fig. 4d, as a weighting factor, a weighted linear regression model was fitted to this data as shown in Fig. 4e. This linear relationship can be utilized to convert the uncertainty to the scale of the errors.

Analysis of pixel-level information

Saliency heatmaps were developed as described in Supplementary Material E. A subset of the test data (1800 images of the transthalamic, transabdominal, and the fetal femur) was analyzed by two feteal medicine clinicians. The two most intensely highlighted regions of each image were annotated, and the frequency of different anatomical features was calculated across the dataset. For a detailed description of the annotation protocol, please see Supplementary Material B.

Data availability

Due to the sensitive nature of the data, it is not possible to share model weights and source data within the permissions with which we can access the data. To access these data, the Danish Regions and the Danish Data Protection Agency need to be applied (https://www.datatilsynet.dk/english and https://www.regioner.dk/).

Code availability

The code is available on request.

Received: 8 October 2024; Accepted: 3 May 2025;
Published online: 29 May 2025

References

1. Oral, E. et al. Perinatal and maternal outcomes of fetal macrosomia. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **99**, 167–171 (2001).
2. Andreasen, L. A. et al. Detection of growth-restricted fetuses during pregnancy is associated with fewer intrauterine deaths but increased adverse childhood outcomes: an observational study. *BJOG Int. J. Obstet. Gynaecol.* **128**, 77–85 (2021).
3. Moraitis, A. A., Wood, A. M., Fleming, M. & Smith, G. C. S. Birth weight percentile and the risk of term perinatal death. *Obstet. Gynecol.* **124**, 274–283 (2014).
4. Gardosi, J., Madurasinghe, V., Williams, M., Malik, A. & Francis, A. Maternal and fetal risk factors for stillbirth: population based study. *BMJ* **346**, f108 (2013).
5. Boulvain, M., Irion, O. & Thornton, J. G. Induction of labour at or near term for suspected fetal macrosomia. *Cochrane Database Syst. Rev.* **3**, CD000938 (2016).
6. 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).
7. Henrichs, J. et al. Effectiveness of routine third trimester ultrasonography to reduce adverse perinatal outcomes in low risk pregnancy (the IRIS study): nationwide, pragmatic, multicentre, stepped wedge cluster randomised trial. *BMJ* **367**, l5517 (2019).
8. Hugh, O., Williams, M., Turner, S. & Gardosi, J. Reduction of stillbirths in England from 2008 to 2017 according to uptake of the growth assessment protocol: 10-year population-based cohort study. *Ultrasound Obstet. Gynecol. Off. J. Int. Soc. Ultrasound Obstet. Gynecol.* **57**, 401–408 (2021).
9. Relph, S. et al. Antenatal detection of large-for-gestational-age fetuses following implementation of the growth assessment protocol: secondary analysis of a randomised control trial. *BJOG Int. J. Obstet. Gynaecol.* **130**, 1167–1176 (2023).
10. Andreasen, L. A. et al. Multicenter randomized trial exploring effects of simulation-based ultrasound training on obstetricians' diagnostic accuracy: value for experienced operators. *Ultrasound Obstet. Gynecol. Off. J. Int. Soc. Ultrasound Obstet. Gynecol.* **55**, 523–529 (2020).
11. Stirnemann, J., Salomon, L. J. & Papageorghiou, A. T. Intergrowth-21st standards for Hadlock's estimation of fetal weight. *Ultrasound Obstet. Gynecol.* **56**, 946–948 (2020).
12. Tolsgaard, M. G. et al. Does artificial intelligence for classifying ultrasound imaging generalize between different populations and contexts? *Ultrasound Obstet. Gynecol. Off. J. Int. Soc. Ultrasound Obstet. Gynecol.* **57**, 342–343 (2021).
13. Lee, L. H. et al. Machine learning for accurate estimation of fetal gestational age based on ultrasound images. *npj Digit. Med.* **6**, 1–11 (2023).
14. Xie, H. N. et al. Using deep-learning algorithms to classify fetal brain ultrasound images as normal or abnormal. *Ultrasound Obstet. Gynecol. Off. J. Int. Soc. Ultrasound Obstet. Gynecol.* **56**, 579–587 (2020).
15. Salim, I. et al. Evaluation of automated tool for two-dimensional fetal biometry. *Ultrasound Obstet. Gynecol.* **54**, 650–654 (2019).
16. Yazdi, B. et al. Optimal caliper placement: manual vs automated methods. *Ultrasound Obstet. Gynecol.* **43**, 170–175 (2014).
17. Hussain, Z. & Borah, M. D. Birth weight prediction of new born baby with application of machine learning techniques on features of mother. *J. Stat. Manag. Syst.* **23**, 1079–1091 (2020).
18. Khan, W. et al. Infant birth weight estimation and low birth weight classification in United Arab Emirates using machine learning algorithms. *Sci. Rep.* **12**, 12110 (2022).
19. Feng, M., Wan, L., Li, Z., Qing, L. & Qi, X. Fetal weight estimation via ultrasound using machine learning. *IEEE Access* **7**, 87783–87791 (2019).
20. Lu, Y., Zhang, X., Fu, X., Chen, F. & Wong, K. K. L. Ensemble machine learning for estimating fetal weight at varying gestational age. *Proc. AAAI Conf. Artif. Intell.* **33**, 9522–9527 (2019).
21. Roelants, J. et al. Foetal fractional thigh volume: an early 3d ultrasound marker of neonatal adiposity. *Pediatr. Obes.* **12**, 65–71 (2017).
22. Husen, S. C. et al. Three-dimensional ultrasound imaging of fetal brain fissures in the growth restricted fetus. *PLoS One* **14**, e0217538 (2019).
23. Zeidan, A. M. et al. An approach to automated diagnosis and texture analysis of the fetal liver & placenta in fetal growth restriction. *Mach. Learn. Biomed. Imaging* **1**, 1–37 (2022).
24. Hadlock, F. P., Harrist, R. B., Sharman, R. S., Deter, R. L. & Park, S. K. Estimation of fetal weight with the use of head, body, and femur measurements—a prospective study. *Am. J. Obstet. Gynecol.* **151**, 333–337 (1985).
25. Maršál, K. et al. Intrauterine growth curves based on ultrasonically estimated foetal weights. *Acta Paediatr.* **85**, 843–848 (1996).
26. Papageorghiou, A. T. et al. The intergrowth-21st fetal growth standards: toward the global integration of pregnancy and pediatric care. *Am. J. Obstet. Gynecol.* **218**, S630–S640 (2018).
27. Kiserud, T. et al. The World Health Organization fetal growth charts: a multinational longitudinal study of ultrasound biometric measurements and estimated fetal weight. *PLOS Med.* **14**, e1002220 (2017).
28. Wright, D., Wright, A., Smith, E. & Nicolaides, K. H. Impact of biometric measurement error on identification of small- and large-for-gestational-age fetuses. *Ultrasound Obstet. Gynecol. Off. J. Int. Soc. Ultrasound Obstet. Gynecol.* **55**, 170–176 (2020).
29. 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.* **83**, 102629 (2023).
30. Simcox, L. E., Myers, J. E., Cole, T. J. & Johnstone, E. D. Fractional fetal thigh volume in the prediction of normal and abnormal fetal growth during the third trimester of pregnancy. *Am. J. Obstet. Gynecol.* **217**, 453–e1 (2017).
31. Gardeil, F., Greene, R., Stuart, B. & Turner, M. J. Subcutaneous fat in the fetal abdomen as a predictor of growth restriction. *Obstet. Gynecol.* **94**, 209–212 (1999).

32. Schwartz, J. & Galan, H. Ultrasound in assessment of fetal growth disorders: is there a role for subcutaneous measurements? *Ultrasound Obstet. Gynecol.* **22**, 329–335 (2003).
33. Borboa-Olivares, H. et al. AI-enhanced analysis reveals impact of maternal diabetes on subcutaneous fat mass in fetuses without growth alterations. *J. Clin. Med.* **12**, 6485 (2023).
34. Sood, A. K., Yancey, M. & Richards, D. Prediction of fetal macrosomia using humeral soft tissue thickness. *Obstet. Gynecol.* **85**, 937–940 (1995).
35. Gilboa, Y. et al. Predictive capacity of fetal pancreatic circumference for gestational diabetes mellitus. *Ultrasound Obstet. Gynecol.* **64**, 348–353 (2024).
36. Bhise, V. et al. Defining and measuring diagnostic uncertainty in medicine: a systematic review. *J. Gen. Intern. Med.* **33**, 103–115 (2018).
37. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
38. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y. & Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* **27**, 2176–2182 (2021).
39. Sendra-Balcells, C. et al. Generalisability of fetal ultrasound deep learning models to low-resource imaging settings in five African countries. *Sci. Rep.* **13**, 2728 (2023).
40. Andreasen, L. A. et al. Multi-centre deep learning for placenta segmentation in obstetric ultrasound with multi-observer and cross-country generalization. *Sci. Rep.* **13** https://doi.org/10.1038/s41598-023-29105-x (2023).
41. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
42. Hanley, J. A. & McNeil, B. J. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **148**, 839–843 (1983).
43. Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K. & Dollár, P. Designing network design spaces. *arXiv:2003.13678 [cs]* http://arxiv.org/abs/2003.13678. 2003.13678.
44. Russakovsky, O. et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
45. Wang, G. et al. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* **338**, 34–45 (2019).
46. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).

Acknowledgements

The study was conducted with the Danish Fetal Medicine Study Group.

Author contributions

A.N.C., A.F., M.N., M.B.S., K.M., and M.G.T. conceived and planned the study. M.G.T. and O.B.P. secured data for this study. C.T.V. and M.G.T. provided clinical annotations. M.G.T. and K.M. drafted the first draft of the manuscript. M.L. and K.M. performed deep learning analyses. A.N.C., A.F., and M.N. provided technical revision and feedback. All authors helped revise the manuscript into its current form.

Competing interests

Two patents have been submitted by the institution: Danmarks TekniskeUniversitet, CVR 30060946, Anker Engelunds Vej 101, Kongens Lyngby 2800, Denmark. UK Patent Application No. 2318746.1 A Method of, and Apparatus for, Improved Estimation of Fetal Characteristics. Pending. Covers the computation of weight and its uncertainty. Inventors: K.W.M., A.N.C., A.F., M.G.T., M.N. UK Patent Application No. 2318747.9 An Improved Method of, and Apparatus for, Ultrasound Examination to Extract Fetal Characteristics. Pending. Covers the improved workflow minimizing uncertainty. Inventors: K.W.M., A.N.C., A.F., M.G.T., M.N. A.N.C., A.F., M.N., and M.G.T. own stocks in Prenaital ApS. Other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01704-0.

Correspondence and requests for materials should be addressed to Martin Grønnebæk Tolsgaard.