Fetal biometry and amniotic fluid volume assessment end-to-end automation using Deep Learning

Saad Slimani (saadslimani@outlook.com)
Ibn Rochd University Hospital, Hassan II University https://orcid.org/0000-0001-5258-0252

Saad Slimani
Ibn Rochd University Hospital, Hassan II University

Salaheddine Hounka
Telecommunications Systems Services and Networks lab (STRS Lab), INPT

Abdelhak Mahmoudi
Mohammed V University https://orcid.org/0000-0003-4141-0623

Dalal Laoudiyi
Ibn Rochd University Hospital, Hassan II University

Hanane Saadi
Mohammed VI University Hospital

Amal Bouzyiane
Université Mohammed VI des Sciences de la Santé, Hôpital Universitaire Cheikh Khalifa

Amine Lamrissi
Ibn Rochd University Hospital, Hassan II University

Mohammed Jalal
Ibn Rochd University Hospital, Hassan II University

Said Bouhya
Ibn Rochd University Hospital, Hassan II University

Taha Rehah
Deepecho

Youssef Bouyakhf
Deepecho

Bouabid Badaoui
Mohammed V University in Rabat

Musa Mhlanga
Radboud Institute for Molecular Life Sciences, Epigenomics & Single Cell Biophysics
https://orcid.org/0000-0003-1381-3409

Amina Radgui
Telecommunications Systems Services and Networks lab (STRS Lab), INPT

El Houssine Bouyakhf
Deepecho

Article

Keywords: Fetal ultrasound, Fetal biometry, Deep learning, Amniotic fluid volume, Ultrasound, Artificial Intelligence

Posted Date: November 10th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2239392/v1

Abstract

Fetal biometry (FB) and amniotic fluid volume (AFV) assessments are two crucial yet repetitive tasks of fetal ultrasound screening scans that help detect potential life-threatening conditions, however, they suffer from reproducibility and reliability issues. Advances in deep learning have led to new applications in measurement automation in fetal ultrasound, showcasing human-level performances in several fetal ultrasound tasks. However, most of the studies performed are retrospective “in silico” studies and few include African patients in their dataset.

Here we develop and prospectively assess the performance of deep learning models for an end-to-end FB and AFV automation from a newly constructed database of 172 293 de-identified Moroccan fetal ultrasound images in addition to publicly available datasets. They were tested on prospectively acquired video clips from 172 patients forming a consecutive series gathered at four healthcare centers in Morocco.

Our results show the 95% limits of agreement between the models and practitioners for the studied measurements were narrower than reported intra and inter-observer variability for human expert sonographers for all the studied parameters.

This means that these models could be deployed in clinical conditions, to alleviate time-consuming, repetitive tasks, and to make fetal US more accessible in limited resources environments.

Introduction

Ultrasound (US) is a low-cost, non-invasive imaging modality that has been shown to independently reduce fetal mortality by up to 20%[1]. Yet, 99% of preventable fetal and maternal deaths occur in developing countries where access to fetal ultrasound is scarce, and more than a third of operators have no training at all[2,3]. The WHO recommends at least one US examination for each pregnancy[4], however, there is a shortage of physicians and sonographers able to perform this examination primarily in countries of the Global South[5]. These countries are not the only ones suffering from excessive and increasing fetal and maternal mortality. The USA ranks last amongst industrialized countries in terms of maternal mortality with notable ethnic differences: African-American women are three times more likely to die during pregnancy compared to non-Hispanic White women[6]. Thus, democratizing access to healthcare resources dedicated to fetal and maternal health, regardless of ethnicity, socioeconomic status, and geographic location is a global healthcare priority.

The current US machines market is ongoing a dynamic positive change with the advent of low-cost point-of-care portable US devices that offer quality images at a fraction of the cost of those sold by traditional manufacturers. As such, US devices are becoming more affordable in low-income countries. But this “hardware” technological transformation would only solve part of the problem. More ultrasound devices in inexperienced hands will not benefit patients. Efficient tools that would enable minimally trained operators to perform parts or the entirety of fetal US scans would radically change how we engage with fetal and maternal health.

Two vital and systematic assessments of all routine screening scans are fetal biometry (FB) and amniotic fluid volume (AFV). FB and AFV help detect and manage potential life-threatening conditions. On the one hand, FB is used to determine gestational age (GA), which is essential to guide therapeutic interventions in the case of pre-term labor or pre-eclampsia and detect pregnancy-related complications, such as fetal growth restriction (FGR). FGR, sometimes defined as the "failure of the fetus to meet its growth potential due to a pathological factor"[7], is responsible for 30% of all stillbirths and poor neonatal outcomes. Its diagnosis can rely solely on US FB assessment when abdominal circumference (AC) or estimated fetal weight (EFW) falls below the 3rd percentile[8,9]. On the other hand, AFV abnormalities are strongly associated with increased mortality in the case of low AFV (oligohydramnios)[10]. The single Deepest pocket (SDP) method has proved to be as reliable as the amniotic fluid index method (AFI) for AFV assessment but to cause fewer false positive diagnoses for oligohydramnios and therefore, fewer unnecessary labor inductions.

FB coupled with AFV assessments are time-consuming, repetitive, and error-prone tasks, and several studies have stressed the need for quality audits to ensure measurements reproducibility and lower inter and intra-observer variability[11–13].

Advances in deep learning (DL) applied to medical imaging have sparked interest in its application to measurement automation in fetal ultrasound, with studies showcasing human-level performances of DL models in standard plane classification and segmentation[14–17]. Most of them are retrospective "in silico" studies conducted on Caucasian populations on fixed images except for a few exceptions[18].

An end-to-end FB and AFV assessment workflow automation could potentially alleviate practitioners' burden, increase ultrasound's sensitivity and specificity, and even enable non-trained healthcare workers to perform these measurements in resource-stranded environments.

Results

Data

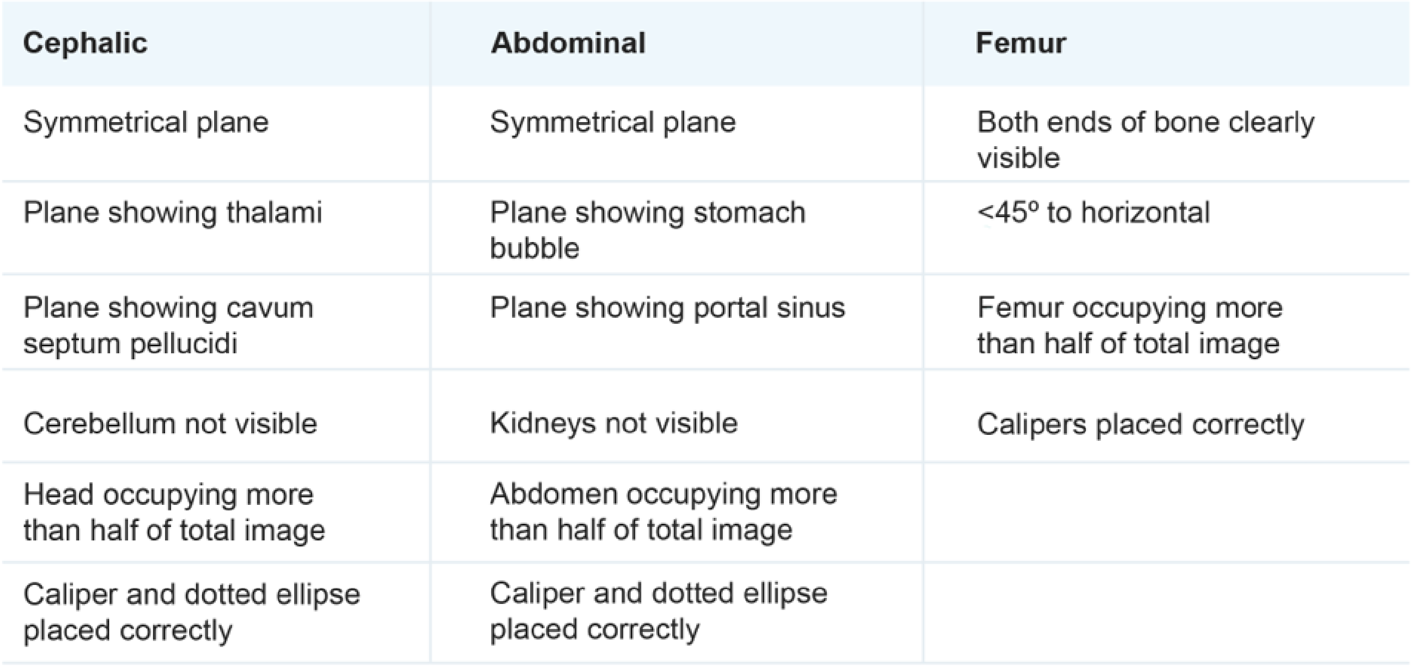
In order to develop and prospectively test DL models designed to fully automate FB and AFV assessment, the models were trained on a newly constructed database of 172,293 de-identified fetal ultrasound images that were collected from 12,356 US examinations performed in six health centers in two different cities of Morocco between 2015 and 2021. In addition publicly available datasets, using the following ultrasound machines: General Electric's Voluson E6, E8, E10, S8 and S10, and Aloka[17].

Within the collected data, 30,249 2D standard biometry planes of the abdomen, brain, and femur were preprocessed and annotated based on the pixelated annotations (images containing calipers - acronyms referring to biometry measurements) and annotated. In addition, ground truth masks were automatically extracted at the preprocessing step to alleviate the work of annotators in the segmentation tasks. In total, fifteen human annotators (ranging from medical students to Radiology and Obstetrics professors) participated in the annotation process using our bespoke annotation platform based on the open-source tool Label Studio[19] that we adapted to our needs. Each annotation included the type of standard plane (abdomen, brain, femur) and some of the quality criteria associated with it as described by the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG) guidelines[20] (Table 1) and polygonal segmentation in the case of the femur. A further distinction between transthalamic, transcerebellar and transventricular planes was made by the annotators. Quality criteria such as the zoom (head, abdomen, femur occupying more than half of the image – caliper placement – angle of the femur to the horizontal < 45 °) were omitted in the annotation process. Instead, their detection was automated through fetal structure segmentation: calculating the surface ratio of the structure to the whole image or the angle of the femur to the horizontal to determine conformity to the criteria described by Salomon et al[21] (Table 1). That step was designed to ensure that the models select the best suitable plane on a given video loop, detecting the presence or absence of the quality criteria, and displaying them with the measurement, allowing an insight into the model's choice as well as a correction if necessary.

Images were also annotated according to the presence or absence of an AF-pocket defined as an in-utero fluid pocket free of fetal parts or umbilical cord. In the case of the presence of the AF pocket, annotators were asked to segment it manually.

Figure 1 summarizes the amount of annotated data for the segmentation of the three biometric structures and their classification based on their respective quality criteria, along with the number of individual measurements in the annotated data for the classification and segmentation of AF pockets.

Table 1: Criteria for score-based biometry plane assessment developed by Salomon et al[21] Models' performance on the retrospective data



Cephalic	Abdominal	Femur
Symmetrical plane	Symmetrical plane	Both ends of bone clearly visible
Plane showing thalami	Plane showing stomach bubble	<45º to horizontal
Plane showing cavum septum pellucidi	Plane showing portal sinus	Femur occupying more than half of total image
Cerebellum not visible	Kidneys not visible	Calipers placed correctly
Head occupying more than half of total image	Abdomen occupying more than half of total image	
Caliper and dotted ellipse placed correctly	Caliper and dotted ellipse placed correctly	

Segmentation and Classifications Models

We assessed the performances of our models on the retrospective test sets comparing them to the experts' annotations for standard plane detection, quality criteria detection, fetal structure segmentation and AF-pocket detection and segmentation.

For the standard plane detection and anatomical regions (brain, abdomen, and femur) segmentation, four MASK-RCNN models were finetuned (R_101_C4_3x, R_101_DC5_3x, R_50_C4_3x, R_50_DC5_3x). The R_50_DC5_3x model achieves the best performance with an average DICE score of 0.89 and an Intersection over Union (IoU) score of 0.82 versus 0.96 and 0.90 respectively reported with the FUVAI model[14] (figure 2). The Segmentation of the brain region achieved the best performance with a DICE score of 0.95 and an IoU of 0.91.

For each biometry plane, classification models for quality criteria detection were assessed on the test set of the retrospective data (figure 3). Assessment of the quality of the standard biometry plane allows for better reproducibility of the AC measurement, we assessed 4 quality criteria (kidneys not visible (A_KN), plane showing portal sinus (A_PS), plane showing stomach bubble (A_SB), symmetrical plane (A_SYM)) leaving out the image zoom quality criteria that is the only one that is not qualitative and can be inferred directly from the abdomen segmentation. Based on three fine-tuned models (INCEPTIONV3, RESNET50V2 and VGG16), INCEPTIONV3 shows the best results for all the criteria with an average area under the curve (AUC) of 0.86. The results also show that A_SB criterion is detected better compared to other criteria with an AUC of INCEPTIONV3 of 0.93.

For the classification of the brain plane, five quality criteria were assessed: cerebellum not visible (B_CB), plane showing cavum septum pellucidity (B_CS), plane showing posterior horn of lateral ventricles

(B_PVV), symmetrical plane (B_SYM) and plane showing thalami (B_TH). Similarly, the 3 classification models were finetuned for this task. They show very similar results with an average AUC of 0.83. The results also show that the B_CB criterion is well detected compared to other criteria with an AUC of INCEPTIONV3 of 0.95 (figure 3).

For the femoral plane, the performances of the model designed to detect if both ends of the femur are clearly visible were poor as inter-observer variability was high in the training set, thus, it was not used for image quality scoring. For the femoral plane on the prospective part of the study, the size, subsequent femur to image sizes ratio, and angle of the femur were directly obtained from the femur segmentation stage and kept as the only quality criteria.

For the AF Pocket classification, we compared the finetuned models (RESNET50V2, INCEPTIONV3 and VGG16) on the retrospective test set (figure 4). The results show almost equivalent AUC scores of 0.89. Similarly, we compared 7 finetuned MASK-RCNN models ('R_101_C4_3x', 'R_101_DC5_3x', 'R_101_FPN_3x', 'R_50_C4_3x', 'R_50_DC5_3x', 'R_50_FPN_3x', 'X_101_32x8d_FPN_3x') for the segmentation of the AF pocket region (figure 4). The results show that 'X_101_32x8d_FPN_3x' achieved the best performance with a DICE score of 0.78 and an IoU of 0.71 versus a DICE of 0.877 for the state of the art model by Cho et al.[33] who tested the model on only 125 images.

From this retrospective study, we adopted the finetuned R_50_DC5_3x model for the segmentation of the fetal structures, the finetuned INCEPTIONV3 models for the quality criteria and the AF pocket detection, and the finetuned X_101_32x8d_FPN_3x model for the AF pocket segmentation. These models will then be evaluated on the prospectively acquired data.

Models performance on the prospective evaluation

Study population

From October 2021 to April 2022, 172 patients with singleton pregnancies were included in our prospective study. Multiple pregnancies were not an exclusion criterion, and patients were included even in the case of partially complete examinations. However, duplicates and patients without an image nor cine-loop available or no corresponding ground truth measurement obtained were excluded (figure 5). In total, the study gathered: 142 different cine-loops containing a femoral plane; 144 containing an abdominal plane; 123 containing a cephalic plane; and 90 containing AF-pockets.

The US machines and healthcare centers from which the prospective data differed from those of the retrospective data were retained. Three of the four centers where the prospective part of the study was conducted did not participate in the retrospective data collection. Several US machines used in the prospective testing were not present in the retrospective data as well: Mindray DC 40 and Resona 6, Philips Medical Systems Affinity 50W and 70G, GE Voluson P8.

When possible, EFW and GA were computed from all measurements using the recommended Hadlock and Intergrowth formulae[29,30] and all necessary measurements performed by the doctors with the corresponding available cine-loops.

Hadlock formula for EFW estimation[29]:
Intergrowth recommended formula for GA estimation > 14 weeks[30]:

Overall, the mean GA estimated by the operators was of 30 weeks and 3.13 days ± 6 weeks and 3.1 days (range: 15 weeks and 2 days – 41 weeks and 2 days), the mean measured HC, BPD, AC, FL, EFW and SDP were respectively of 26.37 ± 5.88 cm (range: 11.29 – 34.71 cm), 7.41 ± 1.72 cm (range: 3.09 – 10.07 cm), 23.98 ± 6.58 cm (range: 8.95 – 38.18 cm), 5.28 ± 1.44 cm (range: 1.52 – 7.86 cm), 1606.78 ± 957.56 g (range: 108.81 – 3783.86 g and 5.25 ± 2.22 cm (range: 2.15 – 17.37 cm).

The models segmented each relevant anatomical region and then extracted the planes with the highest composite score, including quality score according to the ISUOG subjective quality criteria, the zooming of the image inferred from the anatomical segmentation to total image ratio, and the confidence of the model's prediction (figure 6).

The models were able to extract measurements from all the videos containing standard biometry planes. The 95% limits of agreement expressed in percentage using the Bland-Altman method were of 2% for HC, 4.2% for BPD, 3 % for AC, 5.1% for FL, 2.7% for GA, 8% for EFW and 26 % for SDP. All percentages found are narrower than reported inter and intra observer limits of agreements among sonographers (HC: 3.0%, AC: 5.3%, FL: 6.6% for intraobserver difference and HC: 4.9%, AC: 8.8%, FL: 11.1 for interobserver difference)[31](figure 6). Visual assessment of the Bland-Altman plots shows random artifactual bias for every parameter, the variability increasing with the size of the parameter. However, our results also show constant bias for SDP and FL, the predicted measurements for both parameters being consistently greater than those of the physicians.

This over-expectation of the femur segmentation by the model can be mitigated by reviewing the images manually. By selecting images with abnormal results, we found (figure 7) that the model often selected planes showcasing strictly horizontal femurs, and that the predicted calipers were placed avoiding the grand trochanter in accordance with measurement guidelines in contrast to some of the participating physicians[32].

As for the SDP discrepancy, it appears as though the model actually detected deeper pockets not selected or measured by the clinician. However, the model's failure can also be explained by a slight angulation of the probe from 90° results in a larger antero-posterior pocket diameter at the time of examination which will be construed as the SDP by our approach (figure 7).

The ICC for each measurement was high (>0.9 for all parameters apart from SDP) showing excellent reliability of the performed measurements: AC = 0.982, HC = 0.987, BPD = 0.975, FL = 0.945, GA = 0.978, EFW = 0.9713, SDP = 0.692.

The MAE for each biometric parameter was of 0.67 ± 0.69 cm for HC, 0.33 ± 0.22 cm for BPD, 0.27 ±0.40 cm for FL, 0.91 ± 0.81 cm for AC, 9.85 weeks ± 14.36 days for GA, 147.18 ± 177.97 g for the EFW and 1.46 ± 1.10cm for SDP (table 2).

The FUVAI model is the closest one to our approach for end-to-end automated biometric assessment from cine-loops and showed similar performances to those of trained sonographers[14].

We computed the MAE of each parameter using the open source FUVAI model developed by Plotka et al.[14] and compared them with our approach (table 2).

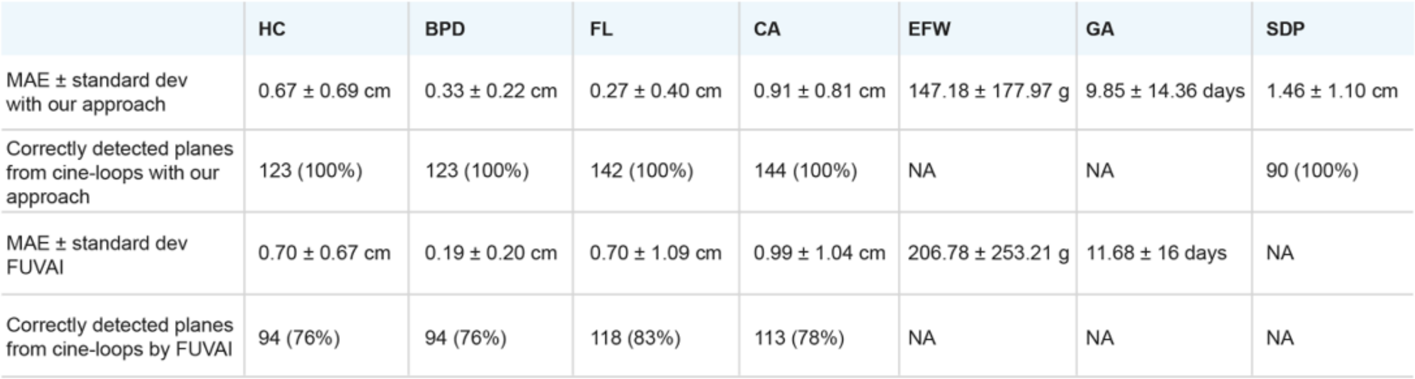
It showed inferior MAE compared with our approach for every biometric parameter except for BPD. We also note that our approach was able to correctly detect the entirety of the corresponding biometry plane while FUVAI failed to do so.

The MAE between the predicted SDP and the measured SDP was also lower than the one reported by Cho et al[33] with their state of the art model for AF pocket segmentation: AF-net (1.46 cm with our approach vs 2.666 cm for Cho et al[33] on a retrospectively annotated data-set).

There were no cases of oligohydramnios in the prospective set and 7 cases (7.07%) of polyhydramnios. The sensitivity and specificity of the models at detecting polyhydramnios was 86.6%, and 85.7% respectively when comparing them to the experts' estimation.

The models' estimated biometric parameters were computed during the prospective phase of the study at the earliest time after each examination was complete. No adverse effect was reported during the entirety of this study.

Table 2: Mean Absolute Error (MAE) for each predicted biometric value, EFW, and GA compared to clinicians and state of the art model FUVAI showing superior correct detection rates and lower MAE with our approach, except for the BPD measurement.



	HC	BPD	FL	CA	EFW	GA	SDP
MAE ± standard dev with our approach	0.67 ± 0.69 cm	0.33 ± 0.22 cm	0.27 ± 0.40 cm	0.91 ± 0.81 cm	147.18 ± 177.97 g	9.85 ± 14.36 days	1.46 ± 1.10 cm
Correctly detected planes from cine-loops with our approach	123 (100%)	123 (100%)	142 (100%)	144 (100%)	NA	NA	90 (100%)
MAE ± standard dev FUVAI	0.70 ± 0.67 cm	0.19 ± 0.20 cm	0.70 ± 1.09 cm	0.99 ± 1.04 cm	206.78 ± 253.21 g	11.68 ± 16 days	NA
Correctly detected planes from cine-loops by FUVAI	94 (76%)	94 (76%)	118 (83%)	113 (78%)	NA	NA	NA

Discussion

In this study, we successfully developed an end-to-end approach to automate FB and AFV estimations from ultrasound cine-loops using the ISUOG quality criteria for standard biometry planes with performances similar to expert operators. These two tasks are part of the six fundamental items listed by the ISUOG in the recently updated practice guideline for the routine mid-trimester scan[35]. They allow early detection of life threatening conditions such as FGR, oligohydramnios and polyhydramnios that are associated with increases of the risk of fetal mortality by respectively 19, 5 and 3 fold[36–38]. The 95% limits of agreement expressed in percentage between the models measurements and the doctors for AC, HC, FL and SDP were narrower than both reported intra- and inter-observer variability for human expert sonographers[13,31]. The difference between the US machines, the operators and the healthcare facilities in the retrospective and the prospective data indicate that the developed models are generalizable. Furthermore, our deterministic method has the advantage of always giving the same output given the same cine-loop which is not the case for human operators. This means that AI can reliably assess fetal growth status and potentially detect AFV abnormalities on fetal US cine-loops automating the third of the six items showcased in the ISUOG guidelines; and has the potential to address the shortage of sonographers in countries of the Global South.

HC, BPD, AC and FL have been shown to be more reliable and reproducible amongst expert operators than SDP measurement with intra and inter CC > 0.990 amongst expert sonographers and clinically acceptable 95% limits of agreement[31,39]. Our models showed intra CC superior to 0.94 for all the biometry metrics (AC = 0.982, HC = 0.987, BPD = 0.975, FL = 0.945) and reached narrower 95% limits of agreement than those reported in studies assessing their reliability and reproducibility between human sonographers.

The models we developed were specifically designed to extract the best biometric planes according to the ISUOG criteria. Although other models have been developed to automate quality control of 2D fetal ultrasound images through anatomical structures recognition, our study is the first study to explicitly use the ISUOG quality criteria specifically for biometry planes classification[40,41]. Such an approach, if integrated in the clinical workflow, could be used to automate the biometry plane's quality control. It could allow fast inexpensive quality audits, accelerate the workflow of trained sonographers, and be a pedagogical tool to the sonographer in-training. This could prove particularly useful in resource stranded regions such as Africa, where only 38.3% of fetal US operators have received formal training, and only 40.4% of them have received a short theoretical course[3].

A similar study to ours compared the performances of a multi-task deep neural network (DNN) on FB assessment, testing it on 50 free-hand ultrasound videos with results comparable to those of trained sonographers. Our models outperformed the one described in the study (FUVAI)[14] when comparing proximity of the results showcased by the model vs sonographers expressed in MAE (table 2)-even if the DICE score coefficients and IoU were lower for the same tasks potentially indicating a greater generalizability of our models. FUVAI's choice of standard biometry planes didn't rely on the quality of the plane but rather on the confidence of the model when selecting it; in other words, on how closely it resembled images from the training set which are not necessarily the best standard planes according to the ISUOG guidelines.

Another vast prospective study by Pokaprakarn et al.[42] took an original approach and assessed the performance of a DNN to estimate GA from blind loops taken by non-trained operators. The DNN proved to be more accurate than expert sonographers at estimating GA with an MAE of 3.9±0.12 days vs 9.85 weeks ± 14.36 days with our approach which could be a game changer in resource stranded environments. Due to the nature of DNNs and the choice of blind sweeps, it is challenging to get a sense of how the model came up with its output and impossible to extract AC or EFW for FGR risk assessment. Instead, our models mimic trained sonographers thanks to the separation of the FB workflow in classification, quality scoring and segmentation tasks. They are thus understandable, errors in the models' outputs being easily detectable by sonographers.

Our approach for AFV assessment is vastly more reliable with limits of agreement of only ± 26% and an ICC of 0.692 for SDP measurement. SDP estimation has the widest variability with reported inter-observer limits of agreements of -51% to + 52% and an ICC of 0.42[13,43]. This high variability amongst human operators might be explained by the "subjective" choice of the SDP. We brought the subjective choice closer to an objective one by segmenting and measuring every single AF-pocket in a given cine-loop. In contrast, several studies present automated techniques to segment AF-pockets and measure the pocket's depth. Cho et al.[33], for example, developed a CNN showcasing results similar to those of sonographers in segmenting AF-pockets (DICE similarity coefficient : 0.877 ± 0.086) and with a MAE of 2.666 ± 2.986 cm in the measurement of the pocket's depth versus a DICE score of 0.783 in our study but a MAE of 1.46 ± 1.10cm on prospectively acquired video loops. However, these come from retrospective studies using 2D fixed images, only automating the segmentation part of the clinical workflow of AFV assessment. Ours proved to be clinically more precise and useful as they detected polyhydramnios with a sensitivity and specificity of 86.6%, and 85.7% respectively .

To the best of our knowledge, our study is the first one to prospectively assess the performance of a model aimed at AFV estimation on US videos. In the context of deployment, clinicians could validate the image selected by the model, potentially correcting an error they or the model committed, or scroll through the selected pockets until a satisfying one is found. Conversely this approach could be compatible with AFV assessment from blind repeated cranio-caudal perpendicular sweeps allowing even minimally trained healthcare workers to perform it.

The limits of our study include the use of cine-loops acquired by one expert operator per examination to compare the performances of the models with those of the physicians. Comparing their performances from US cine-loops acquired from sonographers in training, with those of expert operators for each US scan would give a better assessment of their intended use.

Methods

Models and training

In this study, we finetuned different MASK-RCNN segmentation models on the retrospective data. The MASK-RCNN architectures performance supremacy as well as their easy generalization to other tasks have been proven[24,25]. For the segmentation of the relevant biometric plane, 30249 annotated images were used, (10527 brains, 10227 abdomens and 9495 femurs). For the segmentation of the AF-pockets, only 3773 images were manually annotated with polygons by the experts out of 6199 that were annotated as containing AF-Pocket from the total number of 11926 images. The segmentation models were trained with 80% of the data, validated with 10% and tested with 10%. We also finetuned three classification models (INCEPTIONV3, RESNET50V2 and VGG16) to detect the quality criteria of the abdomen and brain plans with annotated images and to classify 11926 annotated images as containing AF pockets or not. The classification models were trained with 60% of the data, validated with 20% and tested with 20% (Figure 1).

Study design

We validated the DL models on prospectively consecutively acquired transabdominal US videos from pregnant patients (>18 years, evolutive pregnancy > 14 weeks, non-emergency related scan indication, written consent obtained) gathered at four health care centers from October 2021 to April 2022 by 7 different radiologists and obstetricians (experience in fetal US > 10 years) and annotated during the examination using the machine's ellipse and caliper facilities. The participating physicians were asked to measure HC, BPD, AC, and FL following the ISUOG criteria as well as the single deepest AF-pocket (SDP) to assess AFV.

On top of their routine examination, the physicians had to take three additional cine-loops containing all the standard biometry planes, and a cine-loop containing all AF pockets: an axial cephalic loop going from the base of the skull to the vertex, an axial abdominal loop going from the four-chamber view of the heart to a cross section of the kidneys, a sagittal femur loop, and an amniotic loop sweeping perpendicularly through all the right, then the left AF pockets (figure 1).

The physicians had no knowledge of the predicted values for all biometric parameters until the end of the study, the team evaluating the models' performances was also tasked to gather the prospective data, and hence had access to the predicted and measured values for each. On the modeling side, the best segmentation and classification models that were trained on retrospective data were run on each video to extract HC, BPD, AC or FL measurements depending on the plane. All the detected AF-pockets on the "amniotic" cine-loops were segmented and their depth computed, retaining the deepest one as the predicted SDP. This approach is directly inspired by the standard steps taken by expert-trained sonographers to select the single deepest pocket. They consist of the following tasks: 1) Sweep through all AF-pockets, 2) Subjectively select the SDP, 3) Measure the SDP's depth. Oligohydramnios was defined as a SDP < 2 cm and polyhydramnios as a SDP > 8 cm[28].

Evaluation and statistical analysis

DICE score coefficients and Intersection of Union (IoU) were computed for the MASK-RCNNs on the retrospective dataset. For the classification tasks, the receiving operating characteristics (ROC) curves were computed.

The intended sample size was estimated at 122 patients with all corresponding measurements and cine-loops correctly performed. We computed the mean absolute errors (MAE) between the models' measurements and the operators on the prospective cine-loops using the R package 'Metrics' (version 0.1.4) of R software (R version 4.2.1). Intraclass correlation coefficients (ICC) were calculated using the Package 'merTools' (version 0.5.2). ICC is a desirable measure of reliability that reflects both the degree of correlation and agreement between measurements. Wilcoxon rank sum test was calculated for each measurement using the 'PairedData' (version 1.1.1) R package. We also compared the performance of our approach to FUVAI[14] model using the percentage of correctly classified planes and MAE using the R package 'Metrics' (version 0.1.4). Bland-Altman plots were used for the visual assessment of the models' reliability and the 95% limits of agreement were calculated and expressed in percentage using the 'blandr' package (version 0.5.1) of the R software. Firstly, The measurements from the operators and the model were passed to the blandr.statistics function to generate Bland-Altman statistics. Afterwhich, plots were generated using the package ggplot2 (version 3.3.6). Assessment of the models' performances was carried alongside prospective data collection. Approval for this study was granted by the Institutional Review Board of Oujda's Faculty of Medicine (Comité d'Ethique pour la Recherche Biomédicale d'Oujda).

The full protocol of this study can be found on clinicaltrials.gov under the ID: NCT05059093. This study was funded by Deepecho.inc

Declarations

DATA AVAILABILITY

Part of the de-identified fetal ultrasound data used in this study comes from a publicly available dataset available at https://zenodo.org/record/3904280

The rest of the data is not made publicly available.

CODE AVAILABILITY

The pretrained models used in this study are available publicly. Their fine-tuned weights are not made publicly available. API calls for the developed models can be made available for research purposes on demand.

ACKNOWLEDGEMENTS

Authors would like to thank Leila Noureddine, Yasmine Guenni, Abed Tlemcani, Amine Fourari and Nacer Abid for their hard and continuous work on image annotation.

References

1. Grytten, J., Skau, I., Sørensen, R. & Eskild, A. Does the Use of Diagnostic Technology Reduce Fetal Mortality? *Health Serv Res* **53**, 4437–4459 (2018).
2. Wiafe, Y., Odoi, A. & Dassah, E. The Role of Obstetric Ultrasound in Reducing Maternal and Perinatal Mortality. in (2011). doi:10.5772/22847.
3. Carrera, J. M. Obstetric Ultrasounds in Africa: Is it Necessary to Promote their Appropriate Use? *Donald School Journal of Ultrasound in Obstetrics and Gynecology* **5**, 289–296 (2011).
4. WHO recommendations on antenatal care for a positive pregnancy experience. https://www.who.int/publications-detail-redirect/9789241549912.
5. Kim, E. T., Singh, K., Moran, A., Armbruster, D. & Kozuki, N. Obstetric ultrasound use in low and middle income countries: a narrative review. *Reprod Health* **15**, (2018).
6. Maternal Mortality Rates in the United States, 2020. https://www.cdc.gov/nchs/data/hestat/maternal-mortality/2020/maternal-mortality-rates-2020.htm (2022).
7. Melamed, N. *et al.* FIGO (International Federation of Gynecology and Obstetrics) initiative on fetal growth: Best practice advice for screening, diagnosis, and management of fetal growth restriction. *International Journal of Gynecology & Obstetrics* **152**, 3–57 (2021).
8. Nardozza, L. M. M. *et al.* Fetal growth restriction: current knowledge. *Arch Gynecol Obstet* **295**, 1061–1077 (2017).
9. ISUOG. Diagnosis and management of small-for-gestational-age fetus and fetal growth restriction. https://www.isuog.org/resource/isuog-practice-guidelines-diagnosis-and-management-of-sga-and-fgr.html.
10. Morris, R. K. *et al.* Association and prediction of amniotic fluid measurements for adverse pregnancy outcome: systematic review and meta-analysis. *BJOG: An International Journal of Obstetrics & Gynaecology* **121**, 686–699 (2014).
11. Yaqub, M. *et al.* Quality-improvement program for ultrasound-based fetal anatomy screening using large-scale clinical audit. *Ultrasound Obstet Gynecol* **54**, 239–245 (2019).
12. Kilani, R. *et al.* Inter-observer variability in fetal biometric measurements. *Taiwanese Journal of Obstetrics and Gynecology* **57**, 32–39 (2018).
13. Sande, J. A., Ioannou, C., Sarris, I., Ohuma, E. O. & Papageorghiou, A. T. Reproducibility of measuring amniotic fluid index and single deepest vertical pool throughout gestation. *Prenat Diagn* **35**, 434–439 (2015).
14. Płotka, S. *et al.* Deep learning fetal ultrasound video model match human observers in biometric measurements. *Phys. Med. Biol.* (2022) doi:10.1088/1361-6560/ac4d85.
15. Zeng, Y., Tsui, P.-H., Wu, W., Zhou, Z. & Wu, S. Fetal Ultrasound Image Segmentation for Automatic Head Circumference Biometry Using Deeply Supervised Attention-Gated V-Net. *J Digit Imaging* **34**, 134–148 (2021).

16. Kim, H. P. *et al.* Automatic evaluation of fetal head biometry from ultrasound images using machine learning. *Physiol Meas* **40**, 065009 (2019).

17. Burgos-Artizzu, X. P. *et al.* Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. *Scientific Reports* **10**, 10200 (2020).

18. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat Med* **28**, 31–38 (2022).

19. heartexlabs/label-studio. (2022).

20. ISUOG. Ultrasound assessment of fetal biometry and growth. https://www.isuog.org/resource/isuog-practice-guidelines_ultrasound-assessment-of-fetal-biometry-and-growth-pdf.html.

21. Salomon, L. J. *et al.* Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. *Ultrasound in Obstetrics & Gynecology* **37**, 116–126 (2011).

22. Salomon, L. J. *et al.* Feasibility and reproducibility of an image-scoring method for quality control of fetal biometry in the second trimester. *Ultrasound in Obstetrics & Gynecology* **27**, 34–40 (2006).

23. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask R-CNN. in *2017 IEEE International Conference on Computer Vision (ICCV)* 2980–2988 (2017). doi:10.1109/ICCV.2017.322.

24. Evain, E. *et al.* Breast nodule classification with two-dimensional ultrasound using Mask-RCNN ensemble aggregation. *Diagnostic and Interventional Imaging* **102**, 653–658 (2021).

25. Liu, Z. *et al.* Deep learning framework based on integration of S-Mask R-CNN and INCEPTION-v3 for ultrasound image-aided diagnosis of prostate cancer. *Future Generation Computer Systems* **114**, 358–367 (2021).

26. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the INCEPTION Architecture for Computer Vision. Preprint at https://doi.org/10.48550/arXiv.1512.00567 (2015).

27. Morid, M. A., Borjali, A. & Del Fiol, G. A scoping review of transfer learning research on medical image analysis using ImageNet. *Computers in Biology and Medicine* **128**, 104115 (2021).

28. Kehl, S. *et al.* Single deepest vertical pocket or amniotic fluid index as evaluation test for predicting adverse pregnancy outcome (SAFE trial): a multicenter, open-label, randomized controlled trial. *Ultrasound in Obstetrics & Gynecology* **47**, 674–679 (2016).

29. 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. *American Journal of Obstetrics and Gynecology* **151**, 333–337 (1985).

30. Papageorghiou, A. T. *et al.* Ultrasound-based gestational-age estimation in late pregnancy. *Ultrasound in Obstetrics & Gynecology* **48**, 719–726 (2016).

31. Sarris, I. *et al.* Intra- and interobserver variability in fetal ultrasound measurements. *Ultrasound in Obstetrics & Gynecology* **39**, 266–273 (2012).

32. Article: Ultrasound Operations Manual • INTERGROWTH-21st. https://intergrowth21.tghn.org/articles/ultrasound-operations-manual/.

33. Cho, H. C. *et al.* Automated ultrasound assessment of amniotic fluid index using deep learning. *Medical Image Analysis* **69**, 101951 (2021).

34. Cho, A. Y., Yoon, H. J., Lee, K. Y. & Sun, I. O. Clinical characteristics of sepsis-induced acute kidney injury in patients undergoing continuous renal replacement therapy. *Ren Fail* **40**, 403–409 (2018).

35. ISUOG. ISUOG Practice Guidelines (Updated): Performance of the routine mid-trimester fetal ultrasound scan. https://www.isuog.org/resource/isuog-practice-guidelines-updated-performance-of-the-routine-mid-trimester-fetal-ultrasound-scan.html.

36. Pels, A., Beune, I. M., van Wassenaer-Leemhuis, A. G., Limpens, J. & Ganzevoort, W. Early-onset fetal growth restriction: A systematic review on mortality and morbidity. *Acta Obstet Gynecol Scand* **99**, 153–166 (2020).

37. Figueroa, L. *et al.* Oligohydramnios: a prospective study of fetal, neonatal and maternal outcomes in low-middle income countries. *Reproductive Health* **17**, 19 (2020).

38. Tashfeen, K. & Hamdi, I. M. Polyhydramnios as a Predictor of Adverse Pregnancy Outcomes. *Sultan Qaboos Univ Med J* **13**, 57–62 (2013).

39. Perni, S. C. *et al.* Intraobserver and interobserver reproducibility of fetal biometry. *Ultrasound in Obstetrics & Gynecology* **24**, 654–658 (2004).

40. Zhang, B., Liu, H., Luo, H. & Li, K. Automatic quality assessment for 2D fetal sonographic standard plane based on multitask learning. *Medicine (Baltimore)* **100**, e24427 (2021).

41. Wu, L. *et al.* FUIQA: Fetal Ultrasound Image Quality Assessment With Deep Convolutional Networks. *IEEE Transactions on Cybernetics* **47**, 1336–1349 (2017).

42. Pokaprakarn, T. *et al.* AI Estimation of Gestational Age from Blind Ultrasound Sweeps in Low-Resource Settings. *NEJM Evidence* **0**, EVIDoa2100058.

43. Hughes, D. *et al.* <p>Amniotic Fluid Volume Estimation from 20 Weeks to 28 Weeks. Do You Measure Perpendicular to the Floor or Perpendicular to the Uterine Contour?</p>. *IJWH* **13**, 1139–1144 (2021).

Figures

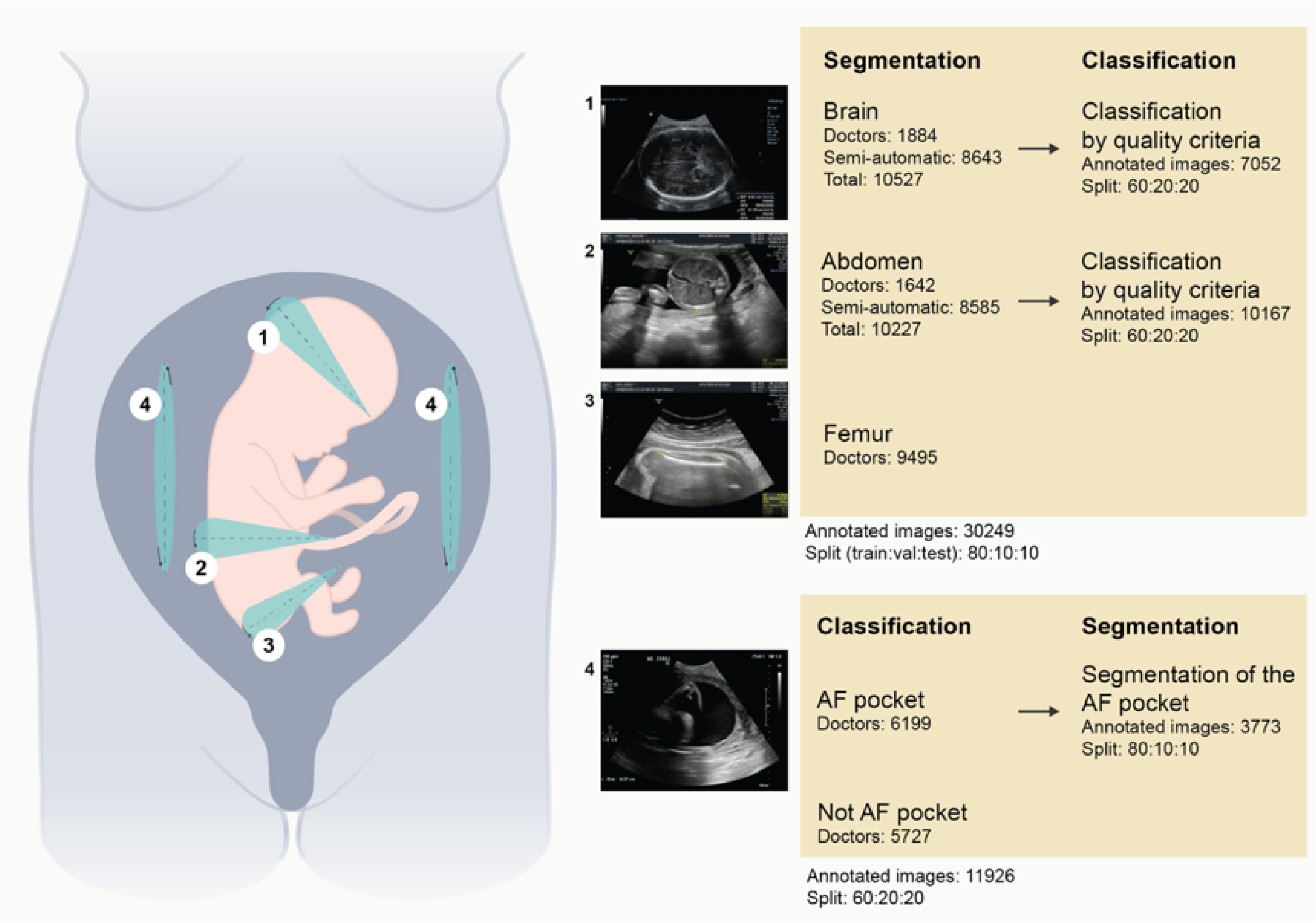


Figure 1

Summary of the retrospective data used during the segmentation and classification tasks along with the data amount used for training, validation, and testing .

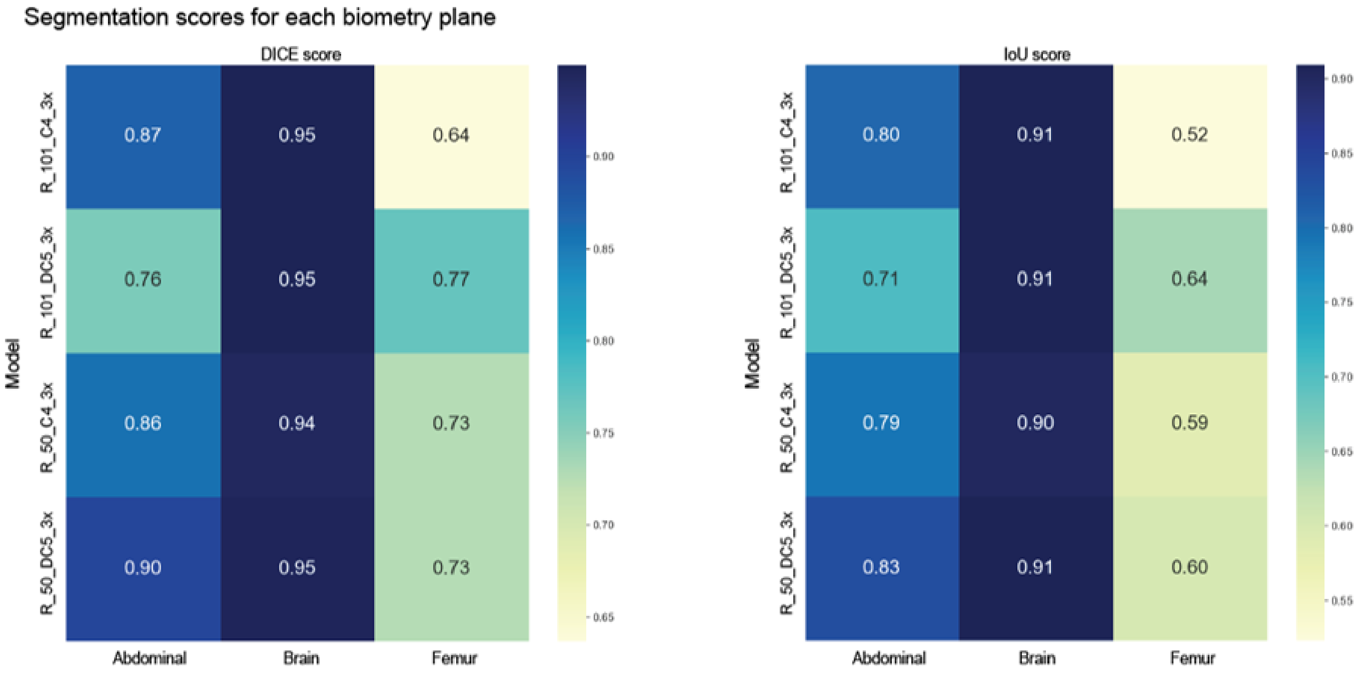


Figure 2

Bar and heatmap plots showing the overall DICE and IoU scores of four finetuned MASK-RCNN models (R_101_C4_3x, R_101_DC5_3x, R_50_C4_3x, R_50_DC5_3x) for the segmentation of the abdominal, femoral and brain planes on the retrospective test set. The bar plot shows the segmentation performances on the three biometric structures and the heatmap plots show the DICE (left) and IoU (right) scores per structure. The R_50_DC5_3x model achieves the best performance with a DICE score of 0.89 and IoU score of 0.82. The segmentation of the brain achieved the best performance with a DICE score of 0.95 and IoU score of 0.91.

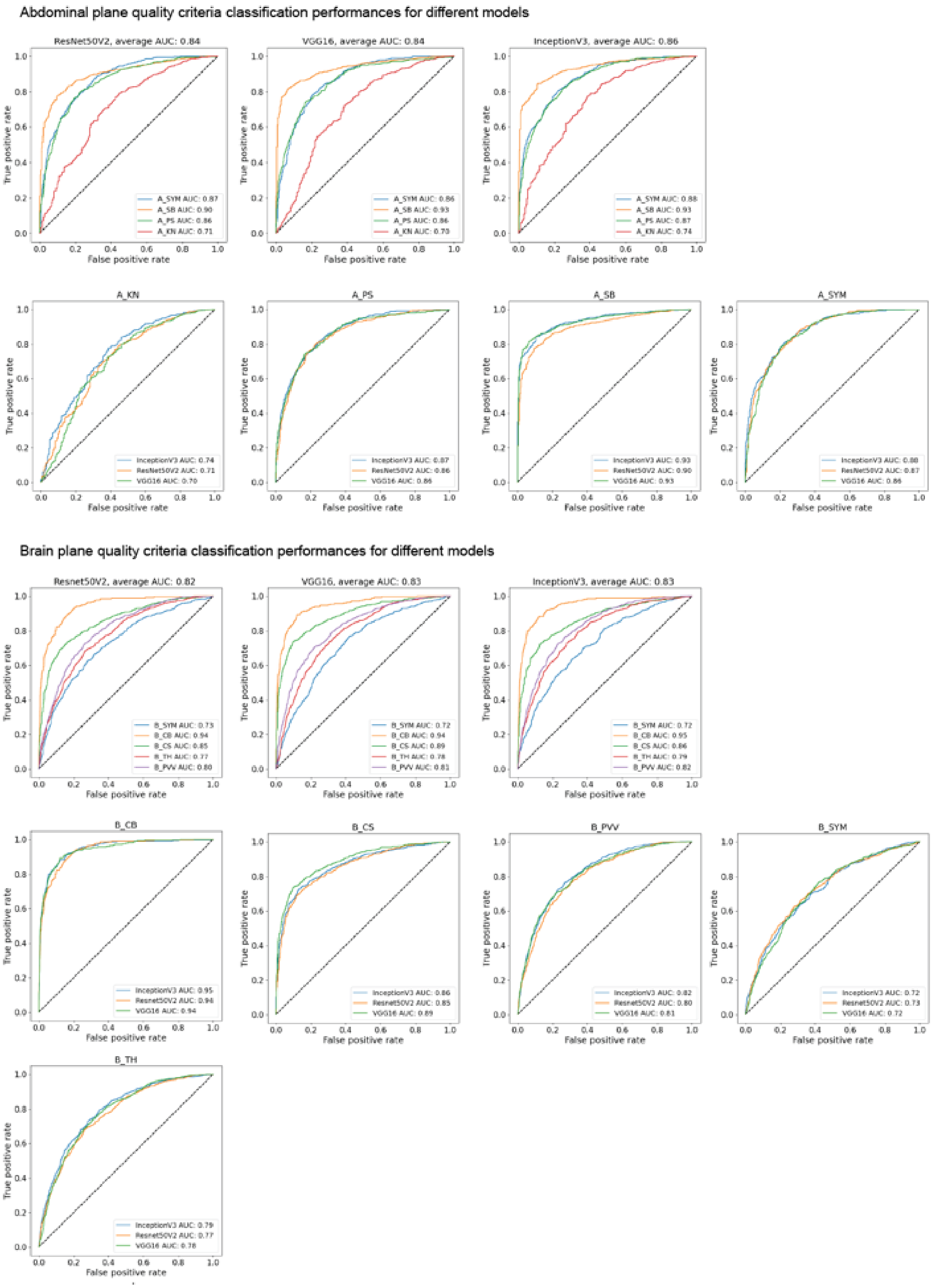


Figure 3

Comparison of the receiver operating characteristics (ROC) curves of three finetuned models (INCEPTIONV3, RESNET50V2 and VGG16) for the brain and abdominal planes classification using their respective quality criteria (kidneys not visible (A_KN), portal sinus visible (A_PS), stomach bubble visible (A_SB), abdominal plane symmetry (A_SYM), brain plane symmetry(B_SYM), cerebellum not visible (B_CB), cavum septum visible (B_CS), posterior horn of lateral ventricle visible (B_PVV) and thalami visible (B_TH)) on the retrospective test set. The top row shows the classification per model and the bottom row shows the results per quality criteria. Overall, the three models show similar results for the cephalic plane quality criteria and INCEPTIONV3 shows the best results for the abdominal criteria with an average AUC of 0.86.

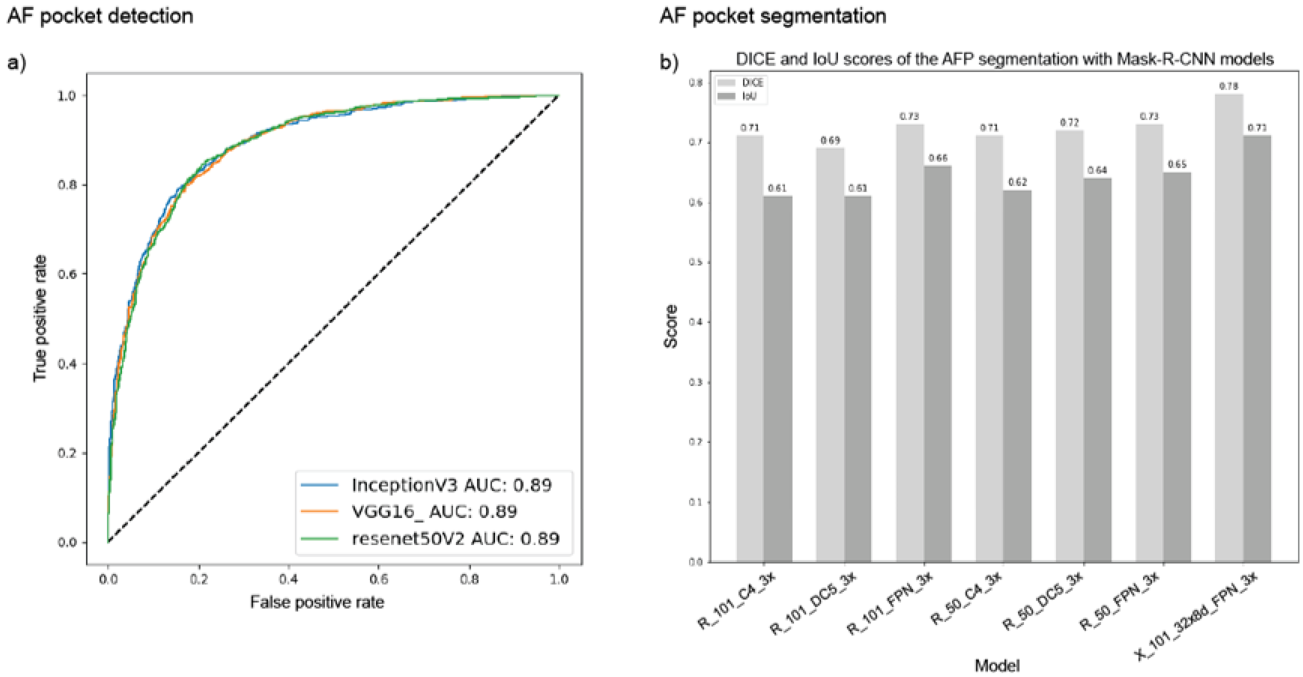


Figure 4

ROC curve and bar plot of the AFP classificationand segmentation respectively on the retrospective test set. a) AUC of three finetuned models for the AFP classification. The results show equivalent AUC scores of 0.89. b) DICE and IoU scores of seven finetuned MASK-RCNN models for the AFP segmentation. The results show that 'X_101_32x8d_FPN_3x' achieved the best performance with a DICE score of 0.78 and an IoU of 0.71.

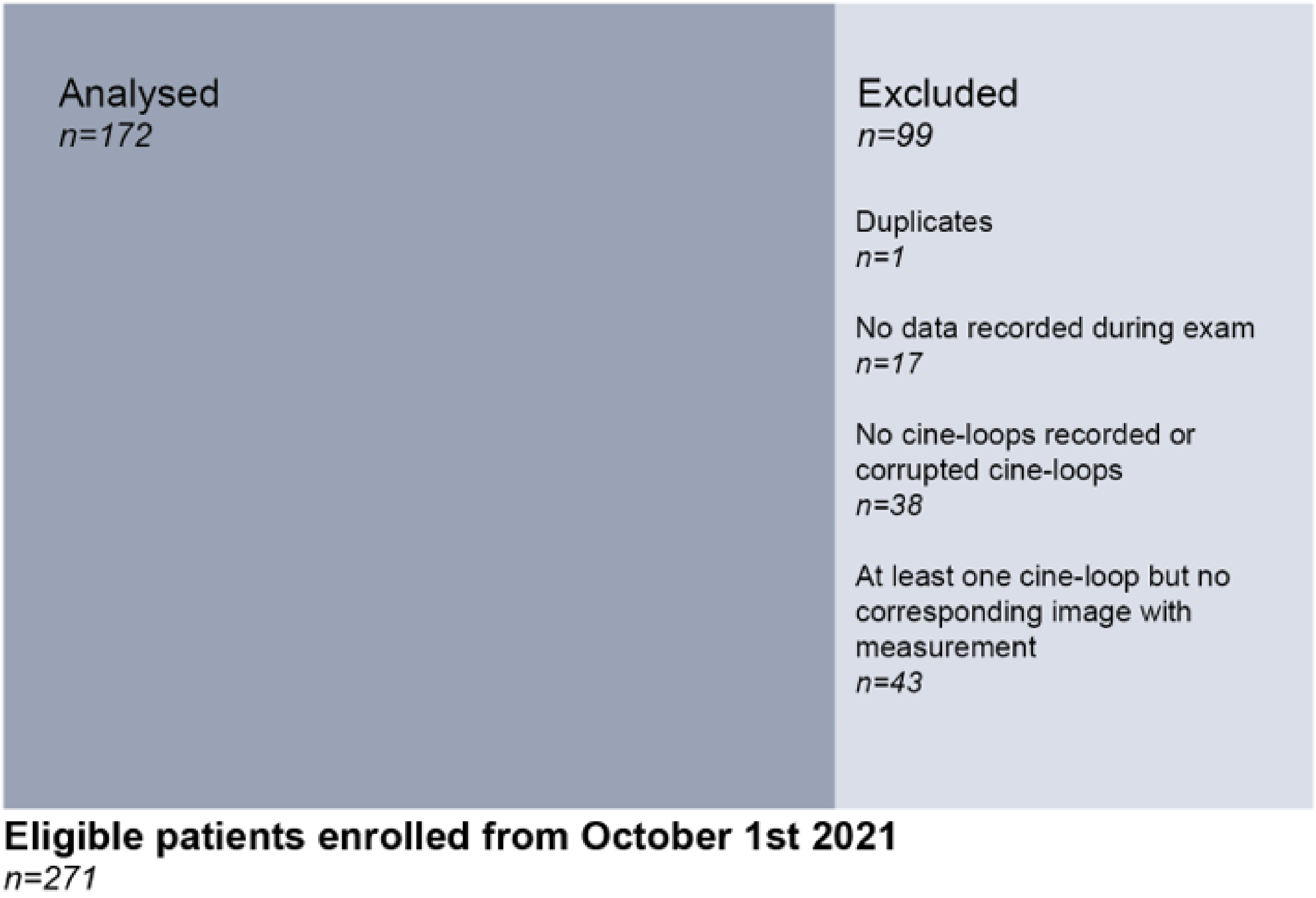

Analysed
n=172
Excluded
n=99
Duplicates
n=1
No data recorded during exam
n=17
No cine-loops recorded or corrupted cine-loops
n=38
At least one cine-loop but no corresponding image with measurement
n=43
Eligible patients enrolled from October 1st 2021
n=271

Figure 5

Study Flow Chart

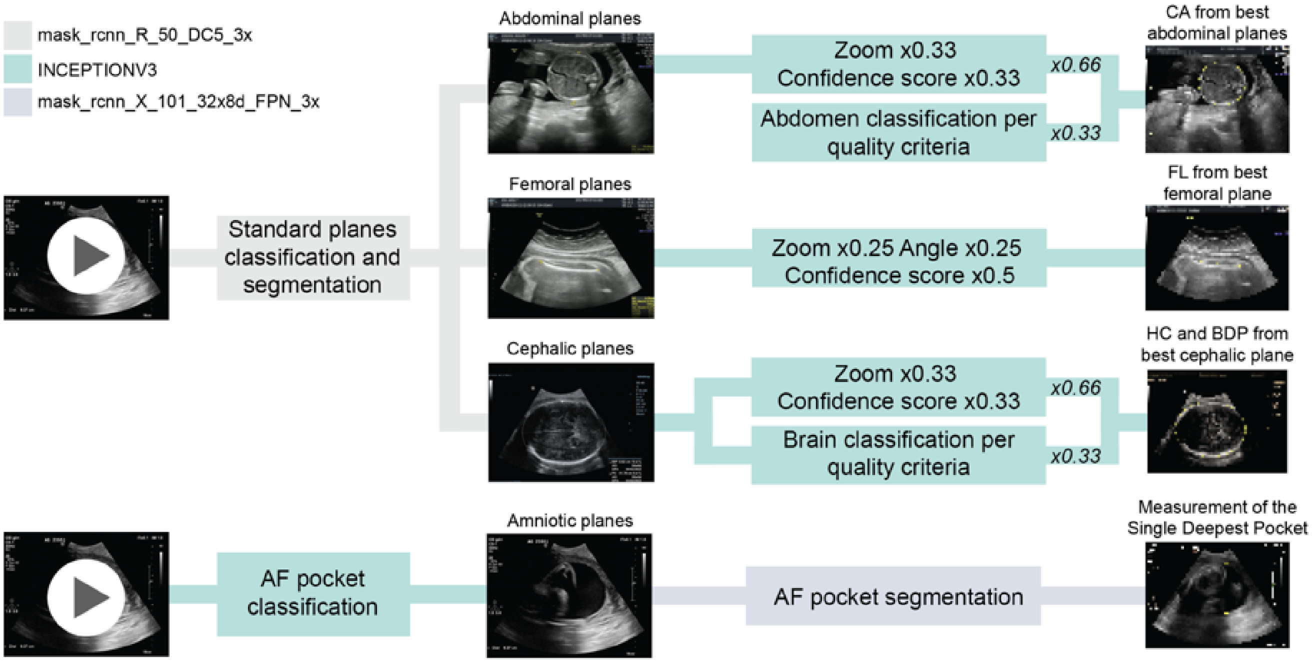

mask_rcnn_R_50_DC5_3x
INCEPTIONV3
mask_rcnn_X_101_32x8d_FPN_3x
Abdominal planes
Zoom x0.33
Confidence score x0.33
Abdomen classification per quality criteria
x0.66
x0.33
CA from best abdominal planes
Standard planes classification and segmentation
Femoral planes
Zoom x0.25 Angle x0.25
Confidence score x0.5
FL from best femoral plane
Cephalic planes
Zoom x0.33
Confidence score x0.33
Brain classification per quality criteria
x0.66
x0.33
HC and BDP from best cephalic plane
AF pocket classification
Amniotic planes
AF pocket segmentation
Measurement of the Single Deepest Pocket

Figure 6

Flow chart of the end-to-end automated extraction of biometric parameters from ultrasound cine-loops. In every cine-loop, all standard biometry planes are detected, the relevant anatomical structures segmented, then the quality criteria of each plane are assessed and the highest scoring plane is selected. There is no quality assessment in the case of the AF volume assessment, the AF pocket with the larger depth is selected from the cine loop.

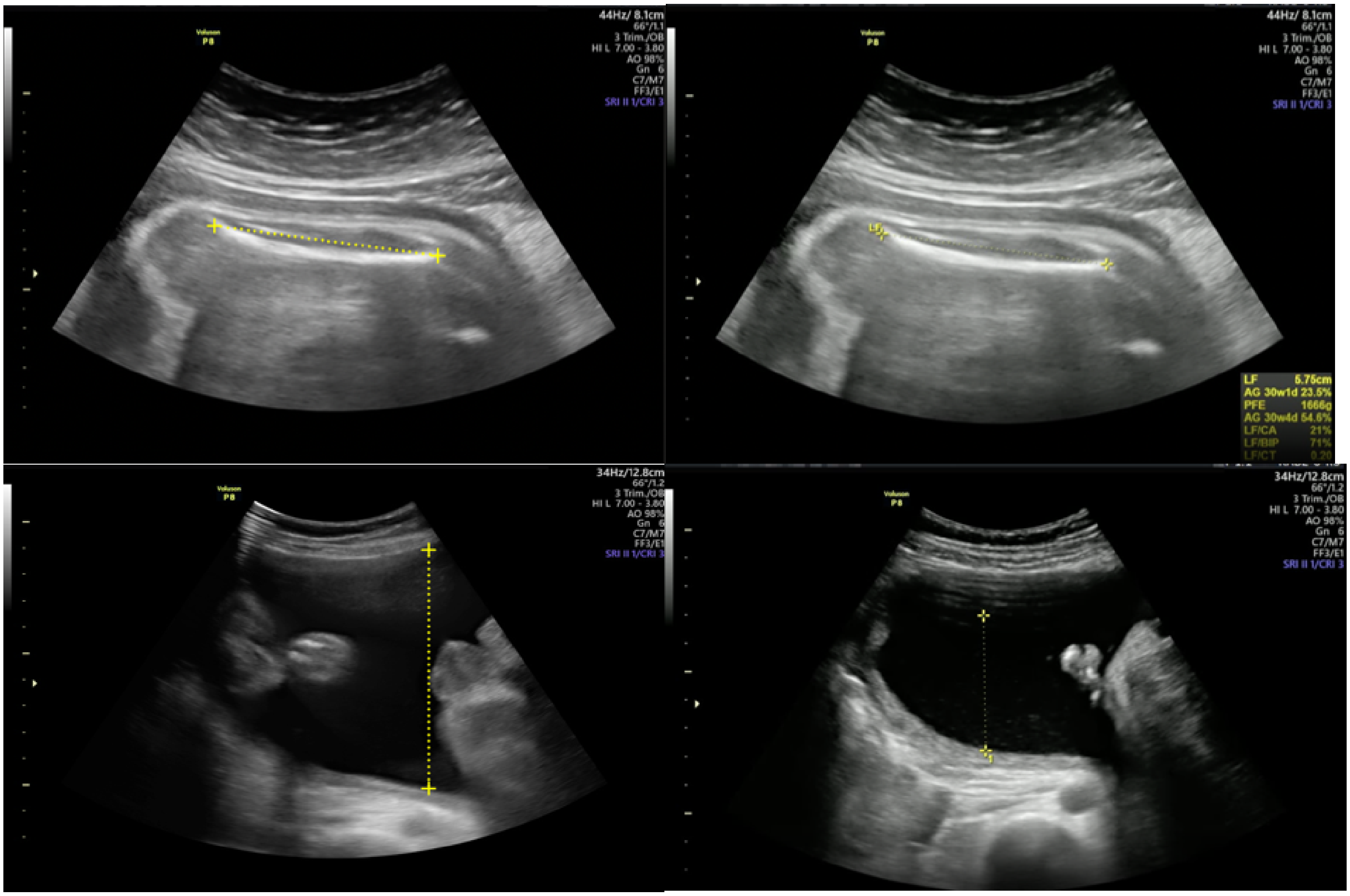


Figure 7

Examples of larger predicted (left) than measured (right) Femur Lengths (FL) and Single Deepest Pockets (SDP) on the same patient.

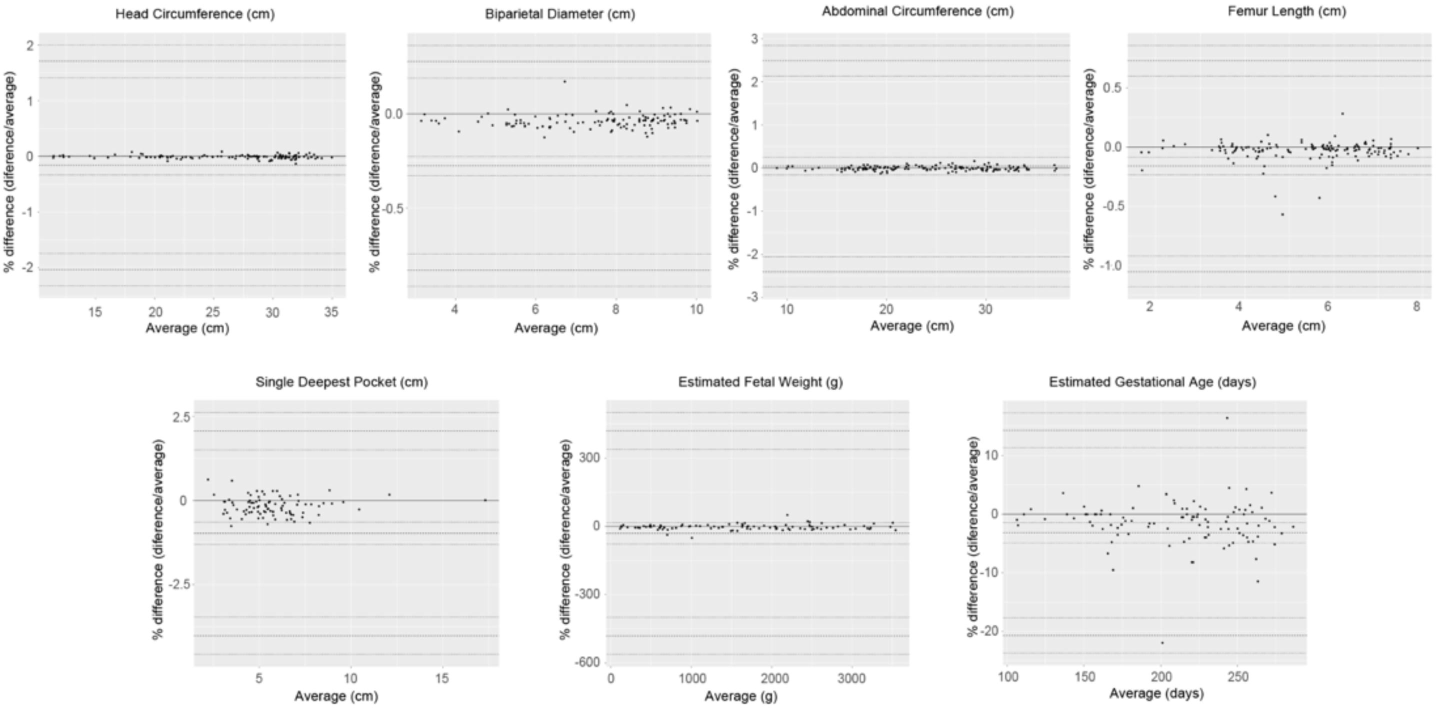


Figure 8

Bland-Altman plots showing the variability between the models and the doctors HC, FL, EGA, AC, AF (Single Deepest Pocket), BDEFW

Supplementary Files

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

- CLAIMSheet1.pdf
- clinicalstudyresultdata.csv
- nreditorialpolicychecklist.pdf
- nrreportingsummary.pdf
- Videoclinicalstudydemo.mp4
- Graphicalabstract.png