Enhancing prenatal care through deep learning

Płotka, S.S.

Publication date
2024

5

Estimation of fetal weight throughout the pregnancy from fetal abdominal ultrasound

BASED ON: Płotka, S., Grzeszczyk, M. K., Szenejko, P., Żebrowska, K., Szymecka-Samaha, N., Łęgowik, T., Lipa, M., Kosińska-Kaczyńska, K., Brawura-Biskupski-Samaha, R., Išgum, I., Sánchez, C. I., and Sitek, A. (2023). Deep Learning for Estimation of Fetal Weight throughout the Pregnancy from Fetal Abdominal Ultrasound. American Journal of Obstetrics & Gynecology Maternal-Fetal Medicine, 101182.

Fetal weight is currently estimated from fetal biometry parameters using heuristic mathematical formulas. Fetal biometry requires measurements of the fetal head, abdomen, and femur. However, this examination is prone to intra- and inter-observer variability due to factors such as the experience of the operator, image quality, maternal characteristics, or fetal movements. We hypothesized that a deep learning (DL) method can estimate fetal weight based on a video scan of the fetal abdomen and gestational age (GA) with similar performance to the full biometry-based estimations provided by clinical experts. The aim is to develop a DL method to automatically estimate fetal weight from fetal abdominal ultrasound video scans. A dataset of 900 routine fetal ultrasound examinations was used. Among those 800 retrospective ultrasound video scans of the fetal abdomen from 700 pregnant women between 15+6 and 41+0 weeks of gestation were used to train the DL model. Following the training phase, the model then was evaluated on an external prospectively acquired test set of 100 scans from 100 pregnant women between 16+2 and 38+0 weeks of gestation. The DL model was trained to directly estimate fetal weight from ultrasound video scans of the fetal abdomen. The DL estimations were compared with manual measurements on the test set made by six human readers with varying levels of expertise. Human readers used standard three measurements made on the standard planes of the head, abdomen, and femur and heuristic formula to estimate fetal weight. Bland-Altman (B-A) analysis, mean absolute percentage error (MAPE), and intraclass correlation coefficient (ICC) were used to evaluate the performance and robustness of the DL method and compare it to human readers. B-A analysis did not show systematic deviations between readers and DL. The mean and standard deviation of MAPE between six human readers and DL approach was 3.75 $\pm$ 2.00%. Excluding junior readers (residents), the MAPE between four experts and DL approach was 2.59 $\pm$ 1.11%. The ICCs reflected excellent reliability and varied between 0.9761 and 0.9865. This study reports the use of DL to estimate fetal weight using only ultrasound video of the fetal abdomen from fetal biometry scans. Our experiments demonstrated similar performance of human measurements and DL on prospectively acquired test data. DL is a promising approach to directly estimate fetal weight using ultrasound video scans of the fetal abdomen.

5.1 Introduction

Fetal ultrasound biometry is a crucial examination that evaluates the growth and health of both the fetus and the mother. The standard procedure involves measuring various parameters such as head circumference (HC), biparietal diameter (BPD), abdominal circumference (AC), and femur length (FL). These parameters are employed as inputs for several existing and standardly used formulas [79] that utilize a unique combination of fetal biometry parameters to estimate fetal weight. The estimated fetal weight (EFW) serves as a significant indicator in monitoring fetal growth and is considered when planning for delivery.

The standard procedure requires the operator to find standard planes for three body parts in which the measurements are performed by hand. These measurements suffer from intra- and inter-observer variability [147; 170], and are time-consuming [57] and dependent on the position of the fetus in the uterus or the quality of the ultrasound image [38].

Recent studies have shown that abdominal circumference, which is a measurement done on abdomen standard plane, has a high correlation with EFW [32] and can be used as a reliable marker for assessing fetal growth [46], identifying small-for-gestational-age or large-for-gestational-age fetuses [26; 99], and determining the appropriate method of delivery (Cesarean or vaginal) [96]. Here, we extended this idea and tested the hypothesis that ultrasound frames of the fetal abdomen are sufficient to reliably estimate the fetal weight and the other scans (head and femur) may not be necessary. We used artificial intelligence (AI) to accomplish this task.

In healthcare, AI has shown the potential to provide new ways of diagnosing, treating, and managing diseases [111]. Deep learning (DL), a rapidly growing field within AI, has seen significant advancements in natural language processing or computer vision [107]. Convolutional neural networks [78] and Transformers [208] are key DL algorithms, and have become popular due to their speed and performance, and are utilized in tasks such as classification [29], segmentation [205], reconstruction [90], and regression [8]. DL-based methods have already been applied to the analysis of fetal ultrasound imaging for fetal diagnosis [57; 225], automatic fetal biometry measurements [13; 152; 153], or estimation of fetal weight at birth [128; 151].

In this study, we aim to directly estimate the fetal weight at various gestational ages of pregnancy using an ultrasound video scan of the fetal abdomen and GA. We evaluate the performance of our method against human experts who use standard-of-care methods. This study demonstrates the feasibility of using solely ultrasound video scans of the fetal abdomen and gestation age to estimate fetal weight using a deep learning algorithm.

This study demonstrates that there is no systemic deviation between the fetal weight estimates made by human readers and those obtained through our deep learning-based method using solely ultrasound video scan of the fetal abdomen. Conventional methods employ measurements from standard ultrasound plane views of the fetus's head, abdomen, and femur to estimate fetal weight. However, we illustrate that data solely derived from the fetal abdomen may be adequate for this purpose when utilizing artificial intelligence. This approach could considerably reduce the time required for biometry and potentially minimize discrepancies that arise from inter-observer variability.

5.2 Materials and Methods

This study was approved by the Ethics Committee of the Medical University of Warsaw (Reference KB.195/2021). Before usage, the dataset was thoroughly anonymized following the ethical standard listed in the Declaration of Helsinki. All patients provided written informed consent to use ultrasound video scans for research purposes.

5.2.1 Data

Here, fetal ultrasound scans were performed between November 2021 and February 2023 at two medical centers, namely, Center A (retrospective set): First Department of Obstetrics and Gynecology, Medical University of Warsaw, Warsaw, Poland, and Center B (prospective set): Department of Obstetrics, Perinatology and Neonatology, The Medical Centre for Postgraduate Education, Warsaw, Poland.

The fetal datasets were acquired following a predefined protocol following international standards approved by the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG) [167]. Consecutive uncomplicated singleton pregnancies were used. Sonographers who performed the examinations

were instructed to record short video clips (10-20 seconds) during which fetal head, abdomen, and femur standard planes were captured. The ultrasound video scans were obtained using sector scan sweep mode, with frame per second ranging between 24 and 37. Sonographers acquired videos without pauses to perform measurements. The video scans were recorded by ten different clinical sonographers with experience ranging between 5 and 40 years. Specifically, the retrospective training set comprised recordings from seven sonographers, whereas the remaining three sonographers (with 20, 15, and 10 years of experience) contributed to the prospective test set.

The full fetal biometry was used exclusively by human readers to manually measure and then compare it with a DL-based algorithm. Here, we trained and evaluated our algorithm solely on ultrasound video frames of the fetal abdomen and GA. The sweeps of the fetal abdomen were intended to capture the appropriate cross-section of the fetal abdominal circumference following the ISUOG guidelines. These sweeps oscillated slightly above and below the correct section. The criteria for the correct biometric image included a symmetrical plane of the abdomen showing the stomach bubble and portal sinus, preferably with one unsegmented rib and no visible kidneys. The abdomen should occupy more than half of the entire image. The abdominal circumference itself was not used to train the model, but rather specific landmarks of the abdomen.

As the training set, we used a clinical, retrospectively collected dataset from Center A. The training set consisted of 800 fetal ultrasound examinations and was obtained from 600 pregnant women aged 18 to 42 years and acquired through routine prenatal ultrasound examinations. The dataset included only singleton pregnancies of White individuals, reflecting the local demographics. The GA was between 15+6 and 41+0 weeks + days with a mean and standard deviation (SD) of 26.9 $\pm$ 5.9 weeks. The training set was acquired using scanners from a single manufacturer. Several models were used, which included P8, S6, and S8 models of GE Voluson. The training data were acquired with 1-5 MHz C-1-5-D, and 2-8 MHz RAB-6-RS standard transabdominal convex transducers. We used a total of 142,837 ultrasound video frames of the fetal abdomen to train the model. The image resolution varied between 975 $\times$ 742 to 1,100 $\times$ 960 pixels. To standardize the data for analysis by our DL-based algorithm, we resized the pixel spacing to 0.2 mm $\times$ 0.2 mm. We chose 0.2

Table 5.1: Basic statistics of minimum, maximum, and mean (with SD) values of HC, BPD, AC, FL, GA, and Hadlock III for the training set. HC, BPD, AC, and FL values are provided in millimeters [mm], while GA and Hadlock III are in weeks, and grams [g], respectively.

Measurement	Minimal value	Maximum value	Mean ± std
Head circumference (HC)	122.1	370.3	249 ± 54.7
Biparietal diameter (BPD)	33.6	103.2	67.9 ± 15.8
Abdominal circumference (AC)	96.7	404.9	229.1 ± 63.6
Femur length (FL)	19.3	82.6	49.9 ± 13.6
Gestational age (GA)	16+2	41+0	26.92 ± 5.91
Hadlock III	133	4737	1283.8 ± 989.4

mm as it was the closest value to the mean pixel size in our dataset. Table 5.1 shows the basic statistics of the training set.

As a test set, we used a prospectively acquired clinical dataset from Center B. The data were acquired from October 2022 to January 2023. The test set consisted of 100 fetal ultrasound examinations and was obtained from 100 women aged 19 to 43 years. The dataset includes only singleton uncomplicated pregnancies of White individuals, reflecting the local demographics. The data were acquired as a part of routine fetal care using the same protocol as for the training set. The GA ranged between 16+2 and 38+0 weeks with a mean and SD of 26.3 ± 5.7 weeks. The test set was acquired by a single ultrasound device, GE Voluson S10 with the 2-8 MHz RAB6-D standard transabdominal convex transducer. Each ultrasound video scan was stored in the DICOM file format, captured in one resolution of 852 × 1,136 pixels which we resized to 0.2 mm × 0.2 mm pixel size. Overall, the test set contained 13,310 ultrasound video frames of the fetal abdomen. Table 5.2. shows the basic statistics for the test set.

5.2.2 Training of deep learning model

As true fetal weights are not available for training the algorithm in a supervised manner, we used an expert-defined estimate of the fetal weight as the

Table 5.2: Basic statistics of minimum, maximum, and mean values of HC, BPD, AC, FL, GA, and Hadlock III for the test set. For every biometry measurement, we present the standard deviation, except gestational age which has a constant minimum and maximum value. HC, BPD, AC, and FL values are provided in millimeters [mm], while GA and Hadlock III are in weeks, and grams [g], respectively.

Measurement	Minimal value	Maximum value	Mean ± std
Head circumference (HC)	129.38 ± 3.76	347.87 ± 7.18	243.34 ± 4.41
Biparietal diameter (BPD)	34.83 ± 0.81	99.28 ± 1.40	66.64 ± 0.80
Abdominal circumference (AC)	107.45 ± 0.49	352.28 ± 8.52	223.88 ± 2.76
Femur length (FL)	20.77 ± 0.33	73.88 ± 0.71	48.37 ± 0.34
Gestational age (GA)	16+0	38+0	26.25 ± 5.77
Hadlock III	156.83 ± 2.26	3,372.92 ± 220.90	1,199.07 ± 33.34

reference standard. These were obtained from HC, BPD, AC, and FL measurements. Each video clip in the training set was annotated by two experienced clinicians, both with more than 15 years of experience by measuring the HC, BPD, AC, and FL. The protocol required reviewing the video clip, selecting the optimal standard plane, and performing measurements using OsiriX MD DICOM viewer software [164]. Of note, one sonographer performed the measurements, which were then reviewed by the other. If any objections were raised about the correctness of the measurement, both sonographers reviewed the study and corrected the measurement as necessary. Once the measurements on the training data set were obtained, the EFW was calculated using a heuristic formula.

A recent systematic review [80] suggested that the most accurate EFW, regardless of whether fetal weight [80] was the standard Hadlock III formula defined as:

$$\begin{aligned} \log_{10} EFW = (&1.326 - 0.00326 \times AC \times FL \\ &+ 0.0107 \times HC + 0.0438 \times AC \\ &+ 0.158 \times FL), \end{aligned} \tag{5.1}$$

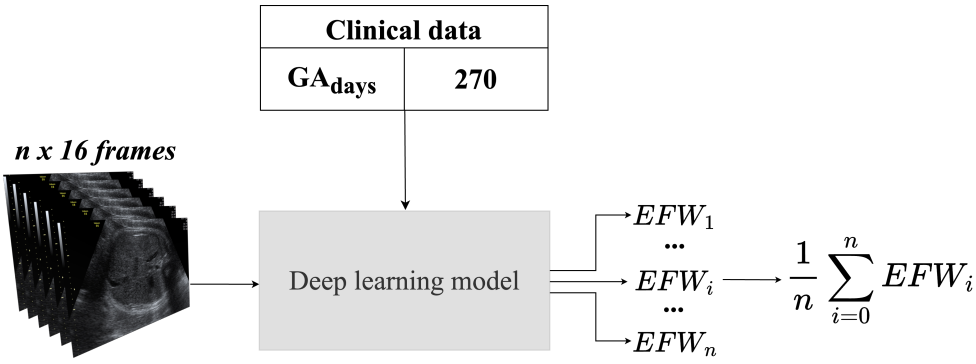

Figure 5.1: The proposed method aims to estimate fetal weight directly from abdominal data using a deep learning-based model. The input to the model consists of US fetal abdominal video scans and clinical data, with each batch sample consisting of 16 video frames and corresponding patient-level clinical data. The final prediction for each patient is obtained by averaging the model's outputs.

where HC, BPD, AC, and FL are head circumference, biparietal diameter, abdominal circumference, and femur length defined in centimeters [cm]. We used this formula in our work.

The architecture of the DL model was based on BabyNet [150; 151] that combines the strengths of both convolutional and transformer architectures. The convolutional component was designed to capture local image patterns and interactions in the input data, whereas the transformer component was used to model long time-range dependencies and relationships. In addition, in the deeper layers of BabyNet, a module was used to shift feature maps conditionally on non-imaging data [154], GA in our case.

The proposed DL model was implemented using PyTorch [146], and a framework, MONAI, developed for medical image processing by NVIDIA [34]. The hyperparameters of the neural network can be found in the supplementary material. In the supplementary material, we have provided data distribution for both the training and test sets, reproducible neural network parameters, and an additional experiment on late pregnancy data.

The neural network was trained in a supervised manner using as input the images from the abdominal video clip and the GA. The output constituted the EFW. Note that as input only, the abdominal scan was used, and no informa-

tion from head or femur scans was used. Each video frame was resized to 128 × 128 pixels. If necessary, the image was cropped or padded. We partitioned the training dataset at the patient level and performed five-fold cross-validation to optimize and verify the robustness of the model. We ensured that all data from a patient appeared only in a single fold. More information regarding the neural network architecture and training process of the neural network can be found in the Appendix. After hyperparameter optimization, the model was retrained on full training data and tested on a prospective test set.

5.2.3 Reader study

We performed a reader study on the test set of 100 prospective fetal ultrasound data samples and later compared them with the results from the DL model. Of note, six readers participated in the study, including three senior gynecologists with 15 (R1), 15 (R2), and 10 (R3) years of experience as completion of residency training, two junior clinicians (residents in obstetrics and gynecology) (R4 and R5), and a senior radiologist (R6) with more than 20 years of experience in the general medical ultrasound. The adopted reading protocol followed the recommendations approved by the ISUOG guidelines for ultrasound assessment of fetal biometry and growth [167].

Readers had no access to any clinical information from the patients. Each reader performed fetal biometry measurements on all 100 cases. First, the reader estimated the standard plane in head, abdomen, and femur video clips and then performed HC, BPD, AC, and FL measurements as per ISUOG guidelines. All readers used the same software, Weasis DICOM medical viewer [229].

For each case in the test set, there were seven estimations of EFW: one estimated measurement provided by the DL algorithm and six estimated measurements derived from the reader measurements of AC, FL, and HC using the Hadlock III formula. We used Bland-Altman [25] plot analysis to investigate the agreement among the estimations of EFW by humans and the DL algorithm.

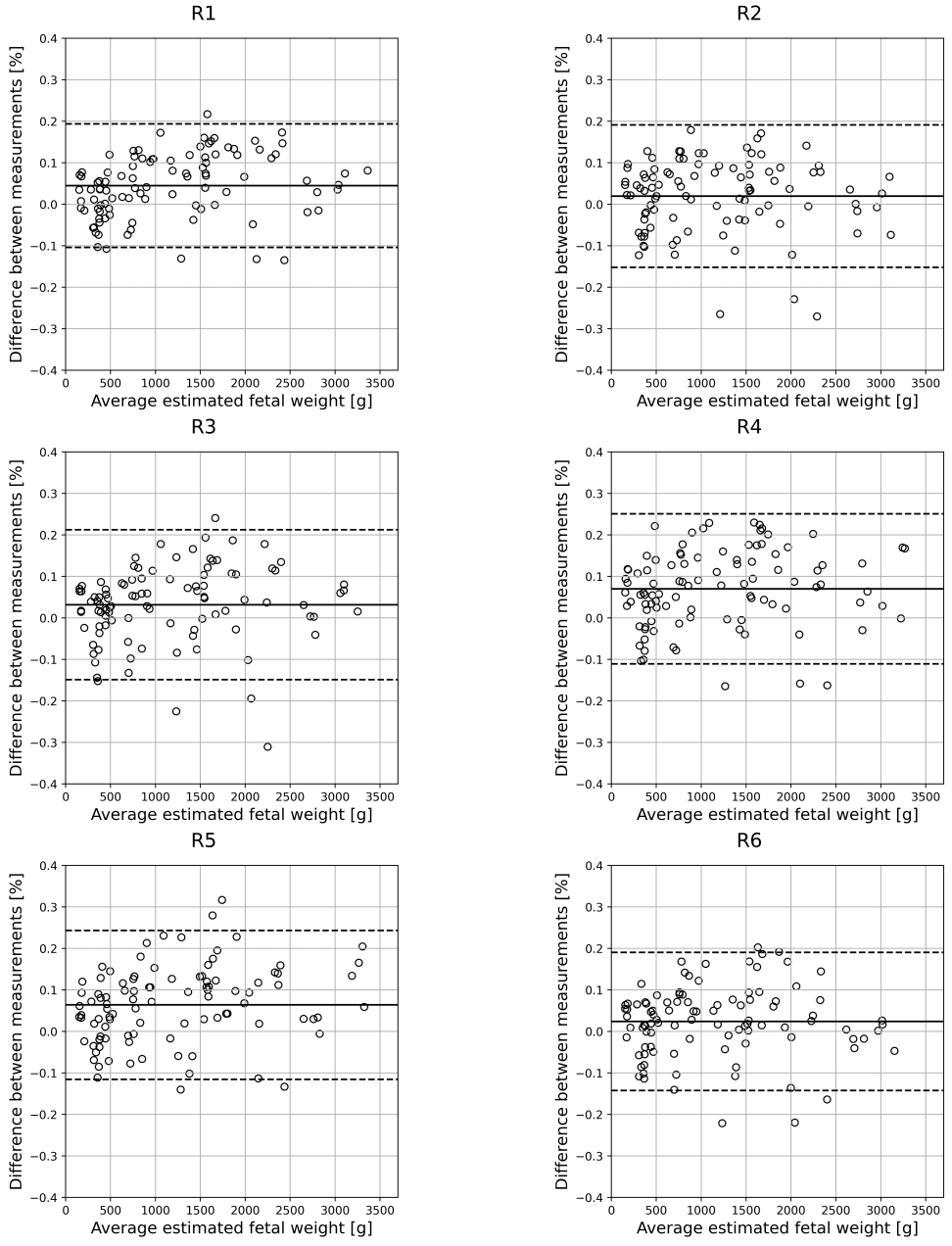

Figure 5.2: Bland-Altman plots show inter-observer differences between estimated fetal weight obtained by six human readers (R1-R6) using the Hadlock III formula and those obtained by deep learning-based method. Solid lines indicate bias, and dotted lines indicate 95% confidence interval.

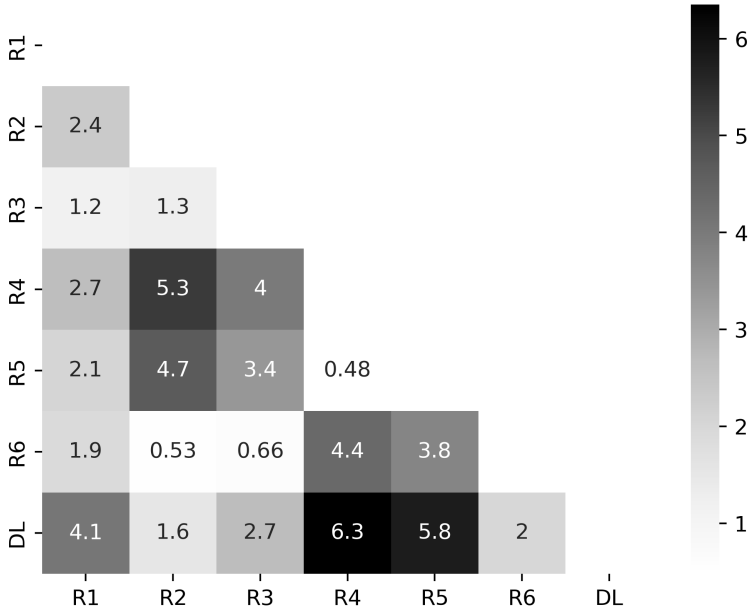

Figure 5.3: Correlation map of mean absolute percentage error (MAPE) for estimated fetal weight for six human readers (R1-R6), and prediction provided by deep learning-based method (DL). The lowest values represent the most accurate estimations of fetal weight alignment between two methods, with the highest values indicating the greatest disparity.

5.2.4 Statistical analysis

We performed statistical analysis using Python and packages for data analysis, which included NumPy, Pandas, and Pingouin. We calculated the mean absolute percentage error (MAPE) between six human readers and DL predictions, which is defined as follows:

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_h - y_{DL}|}{0.5(y_h + y_{DL})} \times 100\%, \tag{5.2}$$

where N, y_h, y_{DL} were the number of patients, fetal weight predicted by human readers, and fetal weight predicted by the DL model, respectively.

Intraclass correlation coefficient (ICC) [24] was calculated between six hu-

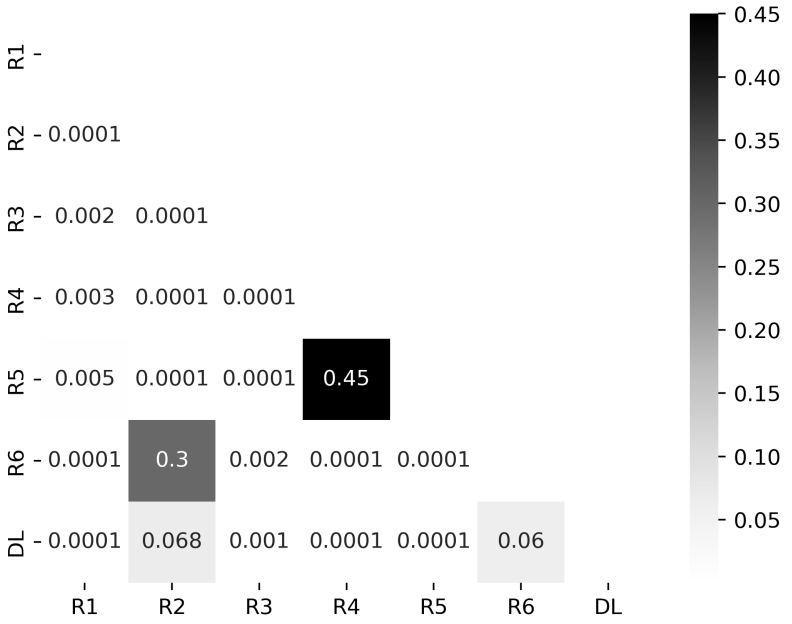

Figure 5.4: Correlation map of the p-value for estimated fetal weight for six human readers (R1-R6), and prediction provided by deep learning-based method (DL). The $p < 0.05$ values indicate significant statistical differences between both methods.

man readers and DL approach. We considered an ICC < 0.50 poor, 0.50 to 0.75 moderate, 0.75 to 0.90 good, and 0.91 to 1.00 excellent reliability [104]. We performed Bland-Altman analysis to calculate the mean difference between two methods of measurement and 95% limits of agreement as the mean difference [25]. The t-tests were performed to assess differences. Differences were considered statistically significant for a p-value of < 0.05.

5.3 Results

Fetal biometry was successfully performed in all 100 patients in the test dataset by six human readers and DL. Figure 5.2 shows the Bland-Altman plots for the estimation of fetal weight by each human reader based on head, abdomen, and femur measurements and DL-based predictions using only abdomen data.

This study presents the differences between measurements in percentages [%] and average EFW in grams [g].

The MAPE for measurements made by six human readers and the DL approach are 4.10%, 1.55%, 2.69%, 6.35%, 5.79%, and 2.00% with a mean ± SD of 3.75 ± 2.00% for R1 to R6, respectively. If we exclude resident (R4, R5) measurements mean of MAPE by four experts (R1, R2, R3, and R6) and the DL approach is 2.59 ± 1.11%.

Figure 5.3 presents an average absolute percentage difference for EFW for human readers (R1-R6), and the prediction (DL) provided by the DL-based method. The average difference (in percentage [%]) among each observer, including R1 to R6 and DL, ranged from a minimum of 0.48% to a maximum of 6.35%, with a mean ± SD of 2.91 ± 1.75%. The ICCs between four senior expert readers and the DL approach were 0.9854 for R1, 0.9865 for R2, 0.9824 for R3, and 0.9872 for R6. The ICCs between two residents and the DL approach were 0.9769 for R4 and 0.9761 for R5.

Figure 5.4 presents the p-value difference for EFW for human readers (R1-R6), and prediction (DL) provided by the DL-based model.

5.4 Discussion

5.4.1 Principal findings

This study investigates the feasibility of using a DL-based algorithm for fetal weight estimation at various stages of pregnancy using fetal abdominal ultrasound video scans and GA. Our study findings, as depicted in the Bland-Altman plots in Figure 5.2, demonstrated that there is no systemic deviation between the fetal weight estimates calculated by human readers and those calculated by our DL-based method using only ultrasound video scan of the fetal abdomen. The plots effectively illustrate the compatibility between the two methods and reveal that the discrepancies in the estimates are evenly distributed around the mean difference, signifying that the DL-based method is comparable with manual biometry performed by human readers.

The visualization of differences between methods in Figure 5.3 shows a pattern in the distribution of the differences, suggesting that DL-based methods have lower MAPE and are more consistent with senior experts as opposed to

junior residents. This correlation can likely be attributed to the training of the network on senior measurements.

5.4.2 Results

Our study demonstrates that the DL-based method estimates the fetal weight within, on average, 3.75% deviation compared with human expert readers using the Hadlock III formula (Figure 5.2 and Figure 5.3). The measurement of fetal weight is not a very consistent process in clinics as many formulas exist, which can differ in weight estimation even by more than 30%. Considering this, our 3.75% range deviation compared with methods used by clinicians is clinically insignificant.

Intra- and inter-observer variabilities for human readers and DL measured by ICCs were excellent and did not differ statistically.

5.4.3 Clinical implications

Our approach boasts a significant advantage as it only necessitates abdominal ultrasound video scans to calculate fetal weight. In contrast, other methods necessitate two or three manual measurements that demand expert knowledge, leading to inefficiency, longer image acquisition and analysis time, and elevated risk of measurement errors.

The advantage of our proposed method is further enhanced by the fact that obtaining the fetal abdominal plane is easier and quicker than other biometric measurements, such as the head or femur, as it can be more accurately captured on ultrasound images. Moreover, the stability in size and shape of the fetal abdomen ensures a more precise estimation of fetal weight, compared with other biometric measurements.

5.4.4 Research implications

Future research focusing on evaluating the method on the larger set of normal and complex cases coming from different scanner manufacturers and populations is warranted. Moving forward, we have set our sights on enhancing the efficiency and user-friendliness of our method.

One of our key objectives is to improve the interpretability of our model. To accomplish this, we intend to extend our method that can execute end-to-

end fully automated fetal weight estimation, including the segmentation of the fetal abdomen, directly from abdominal scans. This will enable us to deliver more accurate, speedy, and completely automated results to our users.

5.4.5 Strengths and Limitations

The strengths of our study include the proposal of a novel DL-based method for fetal weight estimation using only fetal abdominal video scans and GA. Although our method can be a valuable support tool for less experienced clinicians, it is important to note that extensive evaluation and consideration of legal and ethical aspects of the application are warranted. Furthermore, we evaluated our approach on prospective data acquired throughout the entire pregnancy and compared it with the assessments of six clinical experts with varying levels of experience.

It is important to acknowledge the limitations of this study. The sample size of 100 patients used in the evaluation, although consecutive, is modest and may not include cases with severe anomalies that could result in larger errors for either method. In addition, both the retrospective and prospective datasets were obtained using ultrasound scanners from a single manufacturer, which could limit the generalizability of the results to other ultrasound devices. The study population is exclusively composed of White individuals, which reflects the local demographic, and, therefore, may not be representative of other populations.

Similar to other works in fetal biometry, our study faces the challenge of lacking a definitive reference or "ground truth." Our evaluation is based on demonstrating that our method performs with a small margin of error compared with the current standard. Although this margin of error is minor compared with the larger errors that may be associated with the current standard, it remains a limitation of our study that the evaluation is indirect.

In some cases, relying solely on abdominal data for fetal biometry may not provide adequate assessments of fetal growth, such as in cases of congenital anomalies [33], multiple pregnancies [103], intrauterine growth restriction [66; 106], and maternal conditions such as preeclampsia or gestational diabetes mellitus [33; 133]. These cases were not included in our test set, and it is currently unknown how our method would perform in these scenarios.

5.5 Conclusions

Our study compared the results of manual biometry performed by human readers and a DL-based approach for fetal weight estimation and found that both methods produced comparable results with no significant difference between them. The DL approach relied solely on ultrasound video scans of the fetal abdomen, which not only saves time, unlike the traditional full fetal biometry but also minimizes the risk of critical measurement errors that can occur with manual biometry. Our findings suggest that using the ultrasound video scans of the fetal abdomen only with a DL-based method could be a viable option as an alternative to manual biometry.

Supplementary material

Hyperparameters of the neural network

The model was trained using 2 × NVIDIA A100 80GB GPUs with a mini-batch size of 16 and an initial learning rate of 1×10^{-4} with a cosine annealing learning rate scheduler for 200 epochs. For the loss function L, the Mean Squared Error was employed which was defined as:

$$L = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{5.3}$$

where N, y_i, and $\hat{y}_i$ were the number of patients, fetal weight predicted by the Hadlock formula, and DL-model predicted fetal weight, respectively. To minimize the loss function, we employed an ADAM optimizer with L2 regularization of 1×10^{-4}.

During training data augmentation was applied including rotation (± 25 degrees, $p = 0.5$), random brightness and contrast ($p = 0.5$), horizontal flip ($p = 0.5$), image compression ($p = 0.1$), and one of the following: motion blur, median blur, or gaussian blur ($p = 0.5$) for each mini-batch.

Evaluation on data acquired less than 24 hours prior to delivery and ground truth

We conducted additional experiments to demonstrate the generalization and performance of our proposed method on data from the late third trimester. We collected retrospective data from 80 pregnant women aged 23 to 42, with a mean of 31.32 ± 4.21 years. The data were acquired through routine ultrasound (US) examinations less than 24 hours prior to delivery. Our dataset exclusively comprised singleton pregnancy cases within the White population, reflecting local demographics. The data were collected by a clinician with over 15 years of experience at Center B, using a single US device (GE Healthcare Voluson S8) with a 1-5 MHz C-1-5-D transabdominal convex transducer. The acquisition followed the same study protocol as other data in this study. The ground truth for evaluation was the actual birth weight post-delivery in grams [g] (min: 1,850 g, max: 4,399 g, mean: 3,429 ± 484 g).

For evaluation, we utilized 80 ultrasound video scans of the fetal abdomen, totaling 27,440 ultrasound video frames with an average of 343 frames per video. Gestational age in days was also considered during the evaluation.

We calculated the Mean Absolute Percentage Error (MAPE) between predictions provided by the deep learning-based algorithm and the actual birth weight measured post-delivery. On the test set, our method achieved a 4.7% MAPE, indicating a low error in predictions compared to the actual birth weights.

Estimations provided by six human readers

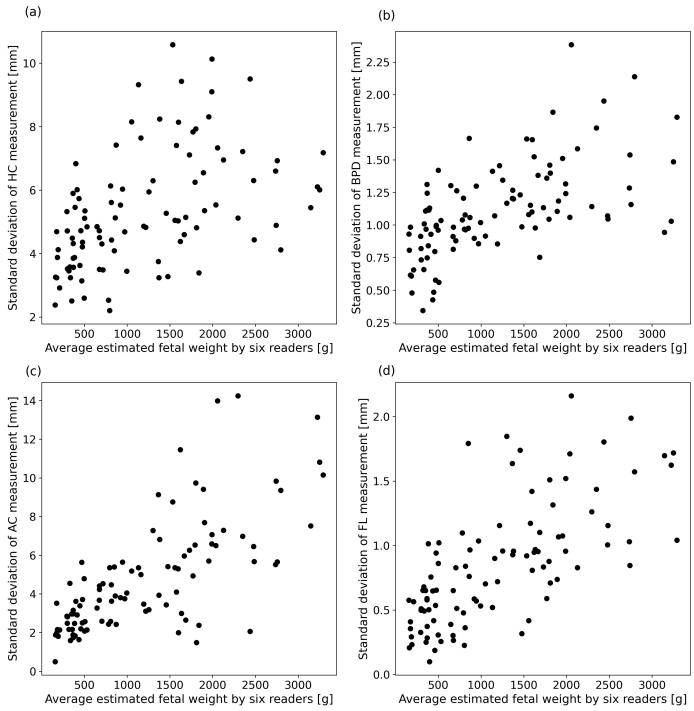

Figure A.1: Scatter plot of average estimated fetal weight by six human readers and standard deviation of measurements a) HC, b) BPD, c) AC, and d) FL.

Distribution of estimated fetal weight using Hadlock III formula in the training and test sets

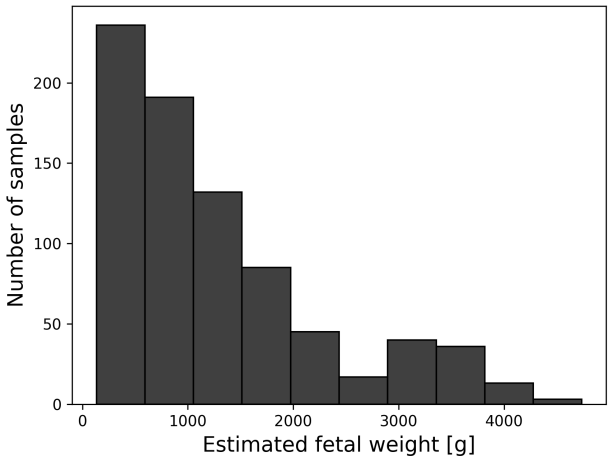

Figure A.2: Histogram - distribution of estimated fetal weight using Hadlock III formula in the training set.

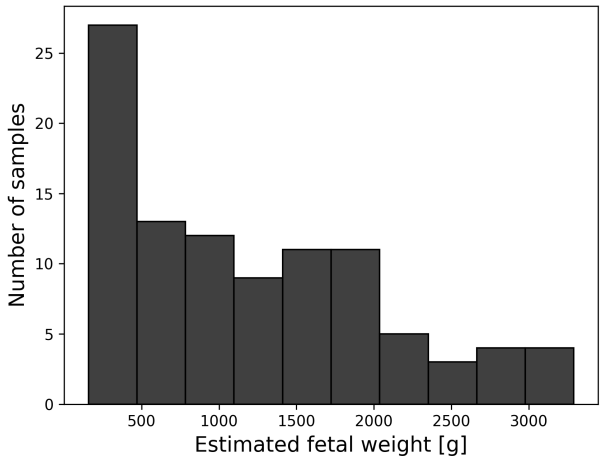

Figure A.3: Histogram - distribution of estimated fetal weight using Hadlock III formula in the test set.

Distribution of gestational age in the training and test sets

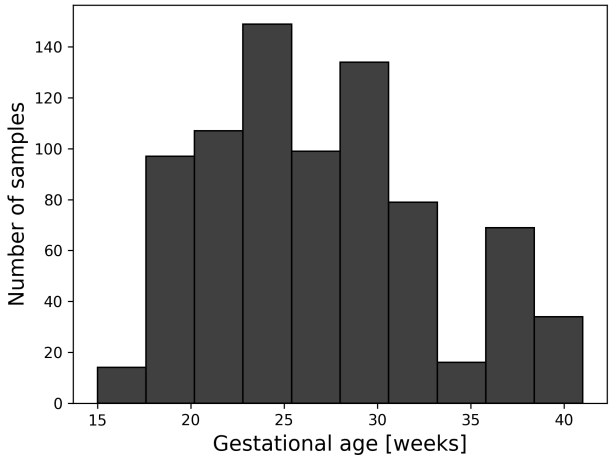

Figure A.4: Histogram - distribution of gestational age in the training set.

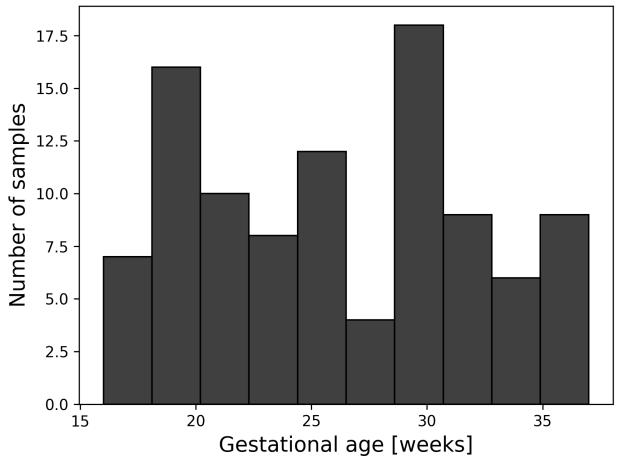

Figure A.5: Histogram – distribution of gestational age in the test set.