Enhancing prenatal care through deep learning

Plotka, S.S.

Publication date
2024

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Fetal Birth Weight Prediction on Fetal Multimodal Data

Accurate prediction of fetal weight at birth is essential for effective perinatal care, particularly in the context of antenatal management, which involves determining the timing and mode of delivery. The current standard of care involves performing a prenatal ultrasound 24 hours prior to delivery. However, this task presents challenges as it requires acquiring high-quality images, which becomes difficult during advanced pregnancy due to the lack of amniotic fluid. In this paper, we present a novel method that automatically predicts fetal birth weight by using fetal ultrasound video scans and clinical data. Our proposed method is based on a Transformer-based approach that combines a Residual Transformer Module with a Dynamic Affine Feature Map Transform. This method leverages tabular clinical data to evaluate $2D + t$ spatio-temporal features in fetal ultrasound video scans. Development and evaluation were carried out on a clinical set comprising 582 2D fetal ultrasound videos and clinical records of pregnancies from 194 patients performed less than 24 hours before delivery. Our results show that our method outperforms several state-of-the-art automatic methods and estimates fetal birth weight with an accuracy comparable to human experts. Hence, automatic measurements obtained by our method can reduce the risk of errors inherent in manual measurements. Observer studies suggest that our approach may be used as an aid for less experienced clinicians to predict fetal birth weight before delivery, optimizing perinatal care regardless of the available expertise. The source code of BabyNet++ is available at https://github.com/SanoScience/BabyNetPlusPlus.

### 4.1 Introduction

Ultrasound (US) is a widely used prenatal imaging method due to its safe, non-invasive, real-time data acquisition properties. Fetal US is the most common and essential tool for assessing fetal growth and detecting fetal abnormalities during pregnancy. In clinical settings, fetal US is mainly used for fetal biometric measurements on standard planes of the head, abdomen, and femur. It also requires knowledgeable and skilled sonographers [152]. However, proper detection of standard planes is subject to intra- and inter-observer variabilities, and results depend on the applied measuring technique [170].

Using the aforementioned standard planes, clinicians typically measure the head circumference (HC), biparietal diameter (BPD), abdominal circumference...
ence (AC), and femur length (FL). These clinical parameters are important for proper pregnancy and delivery management [4; 167], and clinicians can rely on them to derive gestational age (GA) and estimated fetal weight (EFW), both of which are important from the point of view of clinical indicators of the newborn’s growth, survival, and future health, such as fetal birth weight (FBW).

FBW is also used for perinatal risk management and to determine the time and type of delivery (vaginal or cesarean). Low birth weight (LBW) (< 2500 g) and macrosomia (> 4000 g) are signs of potential disruption of normal delivery, which may have significant effects on infant health [20]. While macrosomia may cause infant and maternal morbidity, prolonged labor, and various delivery traumas, LBW is one of the leading risk factors of neonatal death [133]. In addition to considering the parameters of the fetus, factors related to the mother’s health, such as obesity or diabetes, are also taken into account.

Despite the common use of the US to estimate fetal weight, standard prediction models have been shown to cause errors of up to 15% in FBW estimation [172]. Measurement techniques and observer variability contribute significantly to systematic and random error. Therefore, medical examination standardization and substantial training are required for accurate FBW estimation. Averaging multiple measurements, improving image quality, refining measuring techniques, and developing fetal weight estimation formulae [59] are the existing mechanisms that can improve the accuracy of standard measurement techniques.

In current clinical practice, FBW is estimated using heuristic formulae [79] including fetal biometric measurements of body organs – HC, BPD, AC, and FL, extracted from fetal US. With this approach, there is, however, a tendency to overestimate weights in smaller pregnancies while underestimating those of bigger fetuses [137]. Recently, automated methods based on regression algorithms [128; 129] and deep neural network-based models [65] have been investigated for more accurate and personalized FBW prediction.

Fetal birth weight has been previously predicted by our research group [151] and by other researchers [13; 153] using only imaging-based information extracted from US images. These predictions were achieved using state-of-the-art methods such as Transformers [208]. However, these approaches have drawbacks, including limited robustness against artifacts and the lack of domain
adaptation capabilities.

In this paper, we aimed to overcome these drawbacks by expanding upon our previous imaging-based approach. We presented, for the first time, an automated FBW prediction method that utilizes multimodal data, along with the visual data processing component of our network, which is called BabyNet \cite{151}. BabyNet is a hybrid model that efficiently combines Transformers and convolutional neural networks (CNNs). It extends the 3D ResNet-18 architecture with a Residual Transformer Module (RTM), which enables both local and global feature representation through residual connections, utilizes convolutional layers, and refines features through the global self-attention mechanism.

In this work, we further refine our network by adding a Dynamic Affine Feature Map Transform (DAFT) module to leverage tabular clinical data in order to improve the FBW estimation process and to make it more robust compared to the preceding approach, as well as to other state-of-the-art methods. We used fetal US video scans conducted within 24 hours preceding delivery, as well as relevant clinical indicators such as fetal biometry parameters and maternal characteristics such as age, to estimate FBW. The actual birth weight post-delivery served as our ground truth.

In this paper, we not only demonstrate the robustness of our approach but also explore its sensitivity to variability in manual biometry measurements provided by clinicians. We evaluated our methodology using a novel, multimodal, multisite clinical dataset, comprised of US videos and tabular clinical data. In an effort to encourage further research in this field, we have made this dataset publicly available. Moreover, we performed studies with human observers to benchmark our model’s performance against the clinical outcomes derived by clinicians employing heuristic formulae.

4.2 Related work

This section presents a brief overview of state-of-the-art algorithms for birth weight prediction, combining clinical and imaging data, and past work on Transformer-based neural networks.
4.2 RELATED WORK

4.2.1 Birth weight prediction

Fetal birth weight is one of the most clinically significant indicators for optimal growth, pregnancy delivery management, and the future health of newborns. Estimation of this important parameter has been studied using classical statistical analysis [172]. In recent years, machine learning-based methods have been investigated to predict fetal birth weight. Tao et al. [196] proposed a hybrid long short-term memory (LSTM)-based classifier that combines multiple electronic medical records with a B-ultrasonic examination of pregnant women. As input data for the model, they used both fetal and maternal-related physical parameters, as well as time-series data showing fetal weight changes over time. Their method performed better than the estimation based on the empirical formula.

In [129] as well as [116], an ensemble model consisting of Random Forest, XGBoost, and LightGBM algorithms is used. They apply the cubic spline function method to fit the functional relationship between fetal biometry parameters (HC, BPD, AC, FL, GA) to predict fetal birth weight at any gestational age. The authors report a reduction in the mean relative error (MRE) by approximately 3% compared to the Hadlock formula [79].

Recently, deep learning-based solutions have also been investigated [13; 153] where the estimation of FBW was based on fetal US images and videos acquired during routine fetal US biometry examinations. These approaches enable the estimation of FBW at various gestational ages. However, these approaches are susceptible to errors and have the potential to introduce additional bias due to the estimation of fetal biometry parameters relying on segmentation results obtained from the model.

4.2.2 Transformer-based neural network

The Transformer [208] is a deep neural network based on a self-attention mechanism that enables large receptive fields and can identify dependencies in sequential input. Due to their popularity in Natural Language Processing applications, Transformers have been adapted for computer vision tasks [7; 54]. Although Transformers achieve state-of-the-art results, they require large amounts of training data and time due to their high computational complexity. For example, for video regression [161] and classification [7], tens
and even hundreds of thousands of video clips are required in the training process.

Consequently, many approaches have been developed to bridge the gap between sample-efficient learning with high inductive bias of convolutional neural networks (CNNs) and the performance of data-inefficient Transformers. Hybrid models based on CNN layers and Transformer blocks have also been proposed for various computer vision tasks, including image segmentation [83; 203; 226], classification [72; 175; 231], regression [161], reconstruction [124; 130], and registration [238].

In this work, we addressed the presented difficulties in the medical imaging domain using Transformers with a relatively small training video dataset. Similarly to [191], we develop a hybrid model that combines a CNN backbone with Multi-Head Self-Attention (MHSA). In contrast to [191], our method is adapted to 3D image data processing. To enable video processing we added temporal position encoding to 3D MHSA in a 3D ResNet-based neural network.

### 4.2.3 Combining clinical and imaging data

Diagnosing diseases or conducting clinical measurements is a multifaceted process that often requires information unavailable in images or videos (for example, the age of the patient or their medical history). Therefore, the injection of clinical data, if done properly, should increase the performance of deep learning models.

The simplest approach to combining clinical data with features extracted by the network is to concatenate them in one of the final fully connected (FC) layers. Multiple previously proposed studies which utilize this approach have been described in the literature, such as the methods presented in [61; 63; 190] for Alzheimer’s disease diagnosis and prognosis. A multi-task, multi-channel network that predicts clinical scores and classifies brain diseases is another example of multimodal data usage [120]. Gene expression and histopathological image data fusion for breast cancer survival prediction have also been proposed [116]. The same data types are utilized by [81] in the PAGE-Net architecture where genomic features are introduced through FC layers and aggregated with features of histopathological images.
4.2. RELATED WORK

Figure 4.1: An overview of the proposed method for fetal birth weight (FBW) estimation. The input to the network is comprised of 16-frame segments of fetal US videos, along with corresponding tabular data that includes measurements such as abdominal circumference (AC), head circumference (HC), biparietal diameter (BPD), femur length (FL), gestational age (GA) and maternal age. These predictions are then averaged to obtain the final estimation of fetal birth weight.

On the other hand, Duanmu et al. [58] proposed an Interactive model which uses channel-wise multiplication with tabular data-derived features at multiple intermediate layers of the CNN network. Tabular features serve as input at multiple levels of the CNN, which increases the performance of the model at a low computational cost.

A different approach to combining clinical and imaging data was used in [154; 220]. Instead of concatenating tabular data in final FC layers when imaging features are already computed, they propose the DAFT module, which transforms visual-related feature maps conditionally on the basis of tabular data. DAFT is integrated into the earlier layers of the network and impacts the feature extraction process inside the CNN.
4. FETAL BIRTH WEIGHT PREDICTION ON FETAL MULTIMODAL DATA

4.2 An overview of the BabyNet++ method for birth weight estimation directly from fetal US video scans. In BabyNet++, we replace two Residual Modules of 3D ResNet-18 with two Residual Transformer Modules (RTM) containing 3D Multi-Head Self-Attention (MHSA). The Dynamic Affine Feature Map Transform (DAFT) is added before MHSA. RM and RMD stand for Residual Module and Residual Module with Downsampling, respectively. The network takes 16 consecutive US frames as the input to make a single-segment prediction. All frames for a given patient are divided into non-overlapping 16-frame segments and a patient-level prediction is obtained by averaging segment predictions.

4.3 Methods

An overview of our method, BabyNet++, is presented in Figure 4.1. US videos are divided into non-overlapping 16-frame segments and passed to our model. BabyNet++ makes multiple predictions based on each of the segments, which are then averaged to provide the final estimation.

4.3.1 Feature extraction

For the purpose of extracting high-level $2D + t$ spatio-temporal US feature representations, we employed the 3D ResNet-18 [200] as the base network (Table 4.2). The input samples used for training the network consist of a US video sequence $S_{US} \in \mathbb{R}^{T_0 \times 1 \times H_0 \times W_0}$, where $H_0$, $W_0$, and $T_0$ (with 1 indicating the number of channels) corresponds to the height, width, and frame number of the sequence, respectively.

Obtaining reliable data from US images was difficult due to noise and limited data volume. To mitigate this, we divided the fetal US video into shorter...
4.3. METHODS

segments and averaged them, similar to an ensemble method which is often beneficial. This reduces the impact of noise and improves information accuracy by better representing fetal features. However, hardware constraints limit the maximum segment length during network training. Our study found the optimal segment length to be 16 frames, denoted $T_0$, given our network’s batch size.

This video sequence was processed using convolutional residual modules, resulting in a low-resolution feature map sequence $S'_{US} \in \mathbb{R}^{T_1 \times D_1 \times H_1 \times W_1}$, where $T_1 = \frac{T_0}{4}$, $D_1 = 512$, $H_1 = \frac{H_0}{8}$, and $W_1 = \frac{W_0}{8}$.

Following this, the low-resolution, multi-channel feature map sequences were fed into the RTM for final fine-grained feature extraction before being passed to the fully connected classification layer (Figure 4.2).

4.3.2 3D Multi-Head Self-Attention

We generated multiple attention representations at different positions using MHSA, which employs several self-attention heads that were trained jointly, and their outputs are concatenated [208].

To include positional information, we added positional encoding $r$, which was permutation-invariant. We used Relative Positional Encodings (RPE) [177], which were better suited for vision tasks [222] than absolute encoding (e.g., sinusoidal).

To process $2D + t$ US videos, we added temporal positional encoding to the 2D RPE. We computed the positional encoding $r$ as the sum of $R_h \in \mathbb{R}^{1 \times D \times H \times 1}$, $R_w \in \mathbb{R}^{1 \times D \times 1 \times W}$, and $R_t \in \mathbb{R}^{T \times D \times 1 \times 1}$, the height, width, and temporal positional encodings, respectively.

Finally, we computed the 3D MHSA output of $S'''_{US} \in \mathbb{R}^{T \times D \times H \times W}$ input as:

$$MHSA\left(S'''_{US}\right) = \text{concat} \left[ \text{softmax} \left( \frac{Q_i (K_i + r)^T}{\sqrt{d}} \right) V_i \right],$$

(4.1)

where $T = \frac{T_1}{2}$, $D = D_1$, $H = \frac{H_1}{2}$, $W = \frac{W_1}{2}$, $Q_i$, $K_i$, $V_i$ are queries, keys, and values for the $i$th attention head calculated from $W_Q(S'''_{US})$, $W_K(S'''_{US})$, and $W_V(S'''_{US})$ $1 \times 1 \times 1$ 3D convolutions performed over input $S'''_{US}$. Here, $d$ was $D$ divided by the number of heads, and $r$ was a positional encoding.
4. FETAL BIRTH WEIGHT PREDICTION ON FETAL MULTIMODAL DATA

Figure 4.3: An overview of the Dynamic Affine Feature Transform Maps (DAFT) module. The DAFT module adjusts feature maps using tabular data. Once passed to the DAFT module, the feature maps undergo global-average pooling and are combined with a single row of clinical data. This merged representation is then processed through a bottleneck FC layer with \( \tau \) neurons (\( \tau = 7 \)), which is followed by the computation of \( \alpha \) and \( \beta \) values within the second FC layer. Finally, the \( \alpha \) and \( \beta \) values are utilized to scale and shift the feature maps, respectively.

4.3.3 Dynamic Affine Feature Map Transform

Clinical data often contains valuable information that can be extracted from medical images, such as measurements obtained from MRI, tumor size on CT scans, or fetal biometrics captured on US videos. By incorporating this type of data into neural network training, the network’s understanding and perception of the importance of various features in the feature maps can be increased.

The conventional method of utilizing tabular data within the CNN involves concatenating it with clinical features in the final fully connected layers. However, this approach may not have a significant impact on the features learned in earlier layers.

To address this limitation, we used DAFT to embed clinical data within the model’s interior and dynamically rescale and shift the generated feature maps.
4.3. METHODS

These affine transformations, which were conditioned on the clinical dataset, help the network focus on essential features.

DAFT’s goal was to generate $T$ scales $\alpha$ and $T$ offsets $\beta$, where $T$ was the number of input feature maps. The new feature maps are then computed by:

$$S''_{US_t} = \alpha_t \times S''_{US_t} + \beta_t.$$  \hspace{1cm} (4.2)

Figure 4.3 provides a detailed overview of the DAFT architecture. To begin with, we take the input feature maps $S''_{US} \in \mathbb{R}^{T \times D \times H \times W}$, where $T = T_1^2$, $D = D_1$, $H = H_1^2$, $W = W_1^2$ and compute a global average pool. Next, we concatenate the pooled features with the clinical data corresponding to the specific patient. This concatenated data is then passed through two fully connected layers to generate the final feature maps, as given by Eq. 4.2.

4.3.4 Residual Transformer Module

The residual modules used in our study follow a typical structure, starting with a convolutional layer followed by a rectified linear unit (ReLU) activation function and Batch Normalization. This composition is repeated twice, and a skip connection was added to add the output of the previous layers, element-wise, to the input of the residual module [87].

To process the global, low-resolution feature map context, we designed the RTM similarly to the Bottleneck Transformer (BoT) [191], including a self-attention mechanism. RTM is based on the BoT architecture and extended to the 3D space by incorporating temporal position encoding (TPE) [177] into 3D MHSA.

In BoT, MHSA replaces the $3 \times 3$ convolution in the residual bottleneck module to reduce the computational complexity in deeper ResNet architectures. Since 3D ResNets used for video processing were often shallower and do not have bottleneck blocks, we replaced the last convolutional layer in the residual module with MHSA (similar to BoT) to utilize the self-attention mechanism.

We assumed that the self-attention mechanism enhances features better when the input has previously been refined. Therefore, we processed the feature maps with DAFT before passing them to MHSA. DAFT allowed for affine transformations conditioned on clinical data, which were tabular mea-
surements that represented the aggregation of all the features in the feature maps. Using DAFT, we directed the feature maps toward important characteristics of the US video.

Clinician-developed measures transform the feature maps so that MHSA would focus on the correct features. We added DAFT to RTM only when tabular data was available. Finally, we defined RTM as follows:

$$y = BN(MHSA(DAFT(\sigma(BN(Conv(x)))))) + x,$$

where $x$ and $y$ are the input and output of RTM, respectively, $Conv$ denotes the convolutional layer, $BN$ is Batch Normalization, $\sigma$ stands for ReLU, and DAFT is an optionally added module for feature refinement conditioned on tabular data.

### 4.4 Experimental design

This section presents our novel multimodal clinical dataset obtained from multiple centers, consisting of fetal US videos and tabular data. Our approach is described in detail, including implementation specifics and the metrics utilized to evaluate our proposed method.

### 4.4.1 Data

Ethical approval was granted by the Ethics Committee of the Medical University of Warsaw (Reference KB.195/2021). To develop and evaluate our method, we used a multimodal dataset that consists of fetal US video scans and clinical tabular data. We used a multi-center clinical dataset supplied by four centers, namely:

1. Center A: First Department of Gynecology and Obstetrics, University Centre of Mother and Child’s Health of the Medical University of Warsaw, Warsaw, Poland,

2. Center B: Department of Reproduction of the Poznan University of Medical Science, Poznan, Poland,

3. Center C: Department of Obstetrics, Perinatology and Neonatology, Centre of Postgraduate Medical Education, Warsaw, Poland,
Figure 4.4: A histogram showing the distribution of fetal weight in the dataset. The weight in grams is represented on the x-axis, while the number of pregnancies is shown on the y-axis. Most fetuses weigh between 2,500 and 4,000 grams, resulting in an approximately normal distribution. However, some outliers had extremely low or high weights, which may require additional investigation. Overall, this distribution of fetal weight aligns with typical population trends.

4. Center D: Department of Obstetrics, Holy Family Hospital, Warsaw, Poland.

We collected data from a total of 79, 16, 52, and 47 pregnant women from Centers A, B, C, and D, respectively.

4.4.1.1 Data acquisition

The fetal datasets were acquired following a pre-defined protocol pursuant to international standards approved by the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG) [167]. 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. While recording, sonographers did not freeze the video to perform any measurements.

To minimize inter-observer variability, the fetal biometry measurement (HC, AC, BPD, and FL) of the training data was provided by two clinicians with at least 15 years of clinical experience. Throughout the annotation process, the two clinicians were in regular contact and reached a consensus
on any ambiguous or challenging cases. Prior to further use, the dataset was anonymized in accordance with the ethical standards listed in the Helsinki Declaration.

The data was provided by four different clinical sonographers with 20, 15, 10, and 8 years of experience, respectively. The data represented 194 pregnant women aged 21 to 45, with a mean of $31.58 \pm 4.57$ years, and acquired through routine US examinations less than 24 hours prior to delivery. Our dataset included only singleton pregnancy cases in Caucasians representing local demographics.

As the ground truth, we used the true fetal weight measured at birth. Figure 4.4 and Figure 4.5 show the distribution of gestational age and fetal weight, respectively. The ground truth values were between 1,850 and 4,995 grams [g], with a mean of $3,439.8 \pm 503.9$ [g].
4.4. EXPERIMENTAL DESIGN

**Figure 4.6:** The figure presents sample US frames from fetal US videos used in this study for fetal birth weight estimation. The frames depict standard planes of the fetal head, abdomen, and femur, displayed from left to right. However, it is important to note that the quality of US images can vary depending on factors such as gestational age, maternal body habitus, and fetal position. In particular, images obtained closer to the time of delivery may be of lower quality due to the lack of amniotic fluid, which can limit the clarity and visibility of fetal structures. The presented sample frames illustrate the types of images included in the dataset and the potential challenges associated with using them for accurate fetal birth weight estimation.

### 4.4.1.2 Fetal ultrasound video dataset and preprocessing

The dataset consists of 582 (194 patients, 3 videos of different body parts per patient) 2D fetal US video scans in standard plane view of the fetal head, abdomen, and femur. The dataset came from a single US device manufacturer (General Electric Healthcare, Zipf, Austria) of several models (Voluson E6, S8, and S10) with corresponding standard transabdominal convex transducers: 2-5 MHz RAB2-6-RS, and 2-8 MHz RAB6-D.

Each US video scan was stored in the DICOM file format, captured in two resolutions: $960 \times 720$ and $852 \times 1,136$ pixels. For each video, we resampled pixel spacing to $0.2 \text{ mm} \times 0.2 \text{ mm}$. The number of frames was between 463 and 1,480, with a mean of 692. In total, the dataset contained 134,264 US video frames. The US video scans were obtained in the sector scan sweep mode with frame per second (FPS) counts of between 24 and 37.

Figure 4.6 shows sample fetal US video frames of the fetal head, abdomen, and femur, respectively.
4. FETAL BIRTH WEIGHT PREDICTION ON FETAL MULTIMODAL DATA

Table 4.1: Summary statistics for clinical features in the dataset, including median, mean with standard deviation (std), and p-values. The features include head circumference (HC), biparietal diameter (BPD), abdominal circumference (AC), femur length (FL), gestational age (GA), age, and weight.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Median</th>
<th>Mean ± std</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>HC [cm]</td>
<td>33.29</td>
<td>33.31 ± 1.43</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>BPD [cm]</td>
<td>9.39</td>
<td>9.37 ± 0.44</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>AC [cm]</td>
<td>34.43</td>
<td>34.26 ± 2.31</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>FL [cm]</td>
<td>7.27</td>
<td>7.34 ± 0.46</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>GA [weeks]</td>
<td>39</td>
<td>38.53 ± 1.56</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Age [years]</td>
<td>30</td>
<td>31.58 ± 4.57</td>
<td>0.98</td>
</tr>
<tr>
<td>Weight [g]</td>
<td>3,445</td>
<td>3,439.8 ± 503.9</td>
<td>-</td>
</tr>
</tbody>
</table>

4.4.1.3 Fetal clinical parameters

The dataset consisted of 194 tabular data samples for each patient. Each sample contains the following clinical markers: head circumference (HC), biparietal diameter (BPD), abdominal circumference (AC), femur length (FL), gestational age (GA), and mother’s age (Age). The ”outer-to-inner” method was used to obtain the BPD measurements [221]. The rest of the parameters were measured according to ISUOG standards [167].

Our dataset does not contain any missing clinical data. We normalized the tabular data by linear scaling.

Table 4.1 shows the features and their corresponding statistics, such as mean, median, standard deviation, and p-value. The null hypothesis states that the coefficient of the feature in a linear regression of fetal body weight equals zero.

4.4.2 Implementation details

We adopted 3D ResNet-18 [200] as our base neural network. Table 4.2 presents the architectural details of BabyNet++. BabyNet++ comprises a 3D convolutional stem followed by four conv stages: three with two residual modules each, and one final stage implemented with two DAFT modules and two RTMs. The output of the final RTM is global-average pooled (GAP) and fed to the FC
4.4. EXPERIMENTAL DESIGN

Table 4.2: This table compares the ResNet3D-18 and BabyNet++ architectures. The ResNet3D-18 architecture was modified by replacing the last two residual modules with two Residual Transformer Modules (RTMs) containing a 3D multi-head self-attention (MHSA) mechanism instead of the second 3 × 3 3D convolution. Additionally, a Dynamic Affine Feature Map Transform (DAFT) module was included within the RTMs to integrate clinical data into the network.

<table>
<thead>
<tr>
<th>stage name</th>
<th>output size</th>
<th>3D ResNet-18</th>
<th>BabyNet++</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv1</td>
<td>$T_0 \times \frac{H_0}{2} \times \frac{W_0}{2}$</td>
<td>$3 \times 7 \times 7, 64$, stride $1 \times 2 \times 2$</td>
<td></td>
</tr>
<tr>
<td>conv2</td>
<td>$T_0 \times \frac{H_0}{2} \times \frac{W_0}{2}$</td>
<td>$\left[\begin{array}{c}3 \times 3 \times 3, 64 \ 3 \times 3 \times 3, 64 \end{array}\right] \times 2$</td>
<td>$\left[\begin{array}{c}3 \times 3 \times 3, 64 \ 3 \times 3 \times 3, 64 \end{array}\right] \times 2$</td>
</tr>
<tr>
<td>conv3</td>
<td>$T_0 \times \frac{H_0}{4} \times \frac{W_0}{4}$</td>
<td>$\left[\begin{array}{c}3 \times 3 \times 3, 128 \ 3 \times 3 \times 3, 128 \end{array}\right] \times 2$</td>
<td>$\left[\begin{array}{c}3 \times 3 \times 3, 128 \ 3 \times 3 \times 3, 128 \end{array}\right] \times 2$</td>
</tr>
<tr>
<td>conv4</td>
<td>$T_0 \times \frac{H_0}{8} \times \frac{W_0}{8}$</td>
<td>$\left[\begin{array}{c}3 \times 3 \times 3, 256 \ 3 \times 3 \times 3, 256 \end{array}\right] \times 2$</td>
<td>$\left[\begin{array}{c}3 \times 3 \times 3, 256 \ 3 \times 3 \times 3, 256 \end{array}\right] \times 2$</td>
</tr>
<tr>
<td>conv5</td>
<td>$T_0 \times \frac{H_0}{16} \times \frac{W_0}{16}$</td>
<td>$\left[\begin{array}{c}3 \times 3 \times 3, 512 \ 3 \times 3 \times 3, 512 \end{array}\right] \times 2$</td>
<td>$\left[\begin{array}{c}3 \times 3 \times 3, 512 \ 3 \times 3 \times 3, 512 \end{array}\right] \times 2$</td>
</tr>
<tr>
<td></td>
<td>$1 \times 1 \times 1$</td>
<td>Global Avg Pooling, FC layer</td>
<td></td>
</tr>
</tbody>
</table>

layer with one neuron (512 input weights) for fetal birth weight prediction.

We implemented our model with PyTorch and trained it using an NVIDIA RTX 2080 Ti 24GB GPU with a mini-batch size of 16 and an initial learning rate of $1 \times 10^{-4}$ with a cosine annealing learning rate scheduler [126] which is defined as:

$$
\eta_t = \eta_{\text{min}} + \frac{1}{2} (\eta_{\text{max}} - \eta_{\text{min}}) \left(1 + \cos \left(\frac{T_{\text{cur}}}{T_{\text{max}}} \pi\right)\right),
$$

(4.4)

where $\eta_{\text{min}}$, $\eta_{\text{max}}$ are ranges for the learning rate, $T_{\text{cur}}$ account for how many epochs have been performed since the last restart, and $T_{\text{max}}$ is the max number.
of epochs, which is set to 250. To minimize the loss function $L$, we employed an Adam optimizer \cite{Kingma2014AdamAM} with L2 regularization of $1 \times 10^{-4}$. For the loss function $L$, we used the Mean Squared Error (MSE), which was defined as:

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

(4.5)

where $N$, $y_i$, and $\hat{y}_i$ are data points, observed values, and predicted values, respectively.

During training, we applied data augmentation including rotation ($\pm 25^\circ$, $p = 0.5$), 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. We retained the height-to-width ratio and resized video frames to $128 \times 128$ ($H_0 \times W_0$) with padding.

The number of attention heads was empirically set to 8, while the temporal sequence length $T_0$ was 16. Prior to training, we randomly chose 20% of cases from each center as the test set. We split the remaining 80% of the training set at the patient level and performed five-fold cross-validation (CV) to compare and verify the robustness of the regression algorithm. We ensured that data from a single patient appeared only in a single fold. In total, we utilized 156 patients for the purpose of training and validation of our method, while 38 patients were employed to test our approach.

We used a 16-frame segment length of $T_0$ due to hardware limitations, and experiments showed that this was the maximum number of frames that the network could handle with a suitable batch size. Additionally, given the limited amount of data and the presence of noise, it was advantageous to divide the entire video into shorter sequences and average them, similar to an ensemble method. Thus, BabyNet++ transforms a US input sequence $S_{US} \in \mathbb{R}^{16 \times 1 \times 128 \times 128}$ to the output $O_{S_{US}} \in \mathbb{R}^1$ of predicted fetal birth weights.

### 4.4.3 Evaluation metrics

To evaluate the performance of the regression algorithms, we applied the following metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). We calculated the standard deviation for both MAE and MAPE since these metrics were avail-
4.4. EXPERIMENTAL DESIGN

able for each individual patient.

For most experiments, we performed a one-way analysis of variance (ANOVA) to assess the significance of differences among the performance metrics of FBW prediction methods. The methods under investigation included BabyNet++ and other state-of-the-art approaches. Additionally, we performed Tukey’s Honestly Significant Difference (HSD) test to perform pairwise comparisons between groups following an ANOVA to identify which specific groups differ significantly from each other.

To compare predictions provided by BabyNet++ and clinicians, we performed statistical analysis using paired t-tests. We used $p < 0.05$ to distinguish a statistically significant difference.

We performed an inference speed test on a 60-second video sequence. To ensure a fair comparison of inference time, synchronization between the host and device (i.e., CPU and GPU) was utilized. This means that time recording began only after the process running on the GPU had been completed. Additionally, a GPU warm-up of 300 iterations was performed to stabilize the final results. The inference was tested on an NVIDIA Clara AGX equipped with RTX 6000 Quadro 24GB GPU.

4.4.4 Experimental design

We conducted multiple experiments to measure the performance of our approach and to prove its advantage over existing methods. At first, we analyzed the impact of the introduction of clinical data on the efficiency of FBW estimation. Then, we introduced noise to fetal biometry to measure the stability of the model’s predictions. Further, we inspected outlier predictions and compared them to the estimations provided by clinicians. We examined the importance of each fetal body part in FBW estimation. Finally, we compared regression metrics achieved by BabyNet++ with results of state-of-the-art approaches and clinically utilized heuristics.
4. FETAL BIRTH WEIGHT PREDICTION ON FETAL MULTIMODAL DATA

Figure 4.7: Comparison of fetal birth weight (ground truth) with clinician estimations and predictions provided by BabyNet++ for the test set. BabyNet++ predictions are averaged over five folds. Clinician estimations based on heuristic formulae tend to underestimate the FBW value. The model-based prediction error is greater in cases of low birth weight and macrosomia due to the normal data distribution.

4.5 Results

4.5.1 Fetal birth weight prediction results

In Figure 4.7, the differences between the absolute error made by clinicians and BabyNet++ are presented. In twenty-four cases (63%) our model’s error was lower than the error resulting from the use of heuristic formulae by clinicians. Only five prediction errors (13%) were greater than 100 g compared to the corresponding errors made by clinicians. We recorded only four outliers with a high absolute error above 300 g (three of those were recorded in our dataset based on measurements performed by clinicians).

4.5.2 Comparison with state-of-the-art algorithms

We compared our method with several state-of-the-art spatio-temporal approaches $2D + t$ including 3D ResNet [200], ViViT [7], Interactive [58], DAFT [154; 220] and BabyNet [151]. To ensure a fair comparison, we used the same computational settings and environment.

Quantitative evaluation of regression results across all five folds was presented in Table 4.3. The proposed BabyNet++ outperforms all competing approaches across all regression metrics. Specifically, BabyNet++ outperformed the closest competing approaches by a mean of 22 g, 44 g, and 0.6%
Table 4.3: The table presents a quantitative comparison of fetal birth weight prediction using several state-of-the-art methods and clinicians. For each method, the table shows the mean absolute error (MAE) in grams, root mean squared error (RMSE) in grams, mean absolute percentage error (MAPE) in percentage points, number of parameters in millions, and inference time in seconds. All methods underwent evaluation using the test set with an image size of 128 × 128 pixels. The p-value indicates the pairwise comparison of the significance between BabyNet++ and each method. (*) denotes statistical significance provided by Tukey’s HSD test. The results of five-fold cross-validation are averaged. The table is sorted by overall regression performance, and the best results are highlighted in boldface.

<table>
<thead>
<tr>
<th>Method</th>
<th>Image</th>
<th>Tabular</th>
<th>MAE [g]</th>
<th>RMSE [g]</th>
<th>MAPE [%]</th>
<th>Parameters [M]</th>
<th>Inference time [s]</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinicians (this work)</td>
<td>✓</td>
<td>✓</td>
<td>188 ± 24</td>
<td>238</td>
<td>5.4 ± 0.5</td>
<td>-</td>
<td>-</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>ViViT [7]</td>
<td>✓</td>
<td>×</td>
<td>390 ± 52</td>
<td>508</td>
<td>12.0 ± 1.4</td>
<td>3.92</td>
<td>1</td>
<td>&lt; 0.05 (*)</td>
</tr>
<tr>
<td>3D ResNet-18 [200]</td>
<td>✓</td>
<td>×</td>
<td>302 ± 45</td>
<td>390</td>
<td>8.9 ± 1.2</td>
<td>33.14</td>
<td>1</td>
<td>&lt; 0.05 (*)</td>
</tr>
<tr>
<td>BabyNet [151]</td>
<td>✓</td>
<td>×</td>
<td>285 ± 39</td>
<td>374</td>
<td>8.5 ± 1.1</td>
<td>20.58</td>
<td>1</td>
<td>&lt; 0.05 (*)</td>
</tr>
<tr>
<td>Gradient Boost Regressor</td>
<td>✓</td>
<td>✓</td>
<td>215 ± 38</td>
<td>260</td>
<td>6.0 ± 0.7</td>
<td>-</td>
<td>1</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>✓</td>
<td>✓</td>
<td>210 ± 28</td>
<td>251</td>
<td>5.8 ± 0.8</td>
<td>-</td>
<td>1</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Interactive [58]</td>
<td>✓</td>
<td>✓</td>
<td>209 ± 32</td>
<td>251</td>
<td>5.8 ± 0.7</td>
<td>0.06</td>
<td>1</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>DAFT [154]</td>
<td>✓</td>
<td>✓</td>
<td>201 ± 27</td>
<td>247</td>
<td>5.7 ± 0.7</td>
<td>0.06</td>
<td>1</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>BabyNet++</td>
<td>✓</td>
<td>✓</td>
<td>179 ± 19</td>
<td>203</td>
<td>5.1 ± 0.6</td>
<td>20.61</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>

for MAE, RMSE and MAPE, respectively.

The ANOVA revealed a significant effect of the prediction methods on the performance metrics (F(1, N) = 11.000, p < 0.0000001), indicating that there were significant differences in the prediction outcomes among the methods. Further examination of the individual F-values showed that the within-group variability accounted for approximately 1.966 of the overall variability, suggesting that method-specific differences contributed significantly to the observed variations in the performance metrics. Specifically, BabyNet++ performed better compared to the other state-of-the-art methods.

Tukey’s HSD test revealed a statistically significant difference between BabyNet++ and BabyNet, ViViT, and 3D ResNet-18 (p < 0.05), indicating that their predictions significantly differ. The detailed pairwise comparisons for all groups are presented in Table A.1 in the supplementary material. All of the models that were compared exhibited an inference time that was less than one second.
4. FETAL BIRTH WEIGHT PREDICTION ON FETAL MULTIMODAL DATA

Table 4.4: This table presents a quantitative comparison of fetal birth weight prediction on the test set by four clinicians and BabyNet++ across four different centers. At each center, a single clinician provided fetal biometry measurements. Exp. denotes the years of experience of clinicians at each center. The mean values of the mean absolute error (MAE) in grams, root mean squared error (RMSE) in grams, and mean absolute percentage error (MAPE) in percent, along with the standard deviation (for BabyNet++) per patient for Centers A through D, are displayed in the table. The p-value indicates the level of significance between BabyNet++ and clinicians. Additionally, the overall mean for all centers is reported.

<table>
<thead>
<tr>
<th>Center</th>
<th>Exp. [years]</th>
<th>MAE [g] (\downarrow) BabyNet++</th>
<th>MAE [g] (\downarrow) Clinicians</th>
<th>RMSE [g] (\downarrow) BabyNet++</th>
<th>RMSE [g] (\downarrow) Clinicians</th>
<th>MAPE [%] (\downarrow) BabyNet++</th>
<th>MAPE [%] (\downarrow) Clinicians</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>8</td>
<td>215 ± 14</td>
<td>211</td>
<td>246</td>
<td>236</td>
<td>5.9 ± 0.4</td>
<td>5.8</td>
<td>0.09</td>
</tr>
<tr>
<td>B</td>
<td>15</td>
<td>129 ± 45</td>
<td>173</td>
<td>159</td>
<td>202</td>
<td>3.6 ± 1.2</td>
<td>4.8</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>C</td>
<td>20</td>
<td>143 ± 34</td>
<td>161</td>
<td>182</td>
<td>177</td>
<td>4.4 ± 1.0</td>
<td>5.0</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>D</td>
<td>10</td>
<td>178 ± 21</td>
<td>183</td>
<td>225</td>
<td>208</td>
<td>5.1 ± 0.7</td>
<td>5.3</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>All</td>
<td>13.25 ± 4.66</td>
<td>179 ± 19</td>
<td>188 ± 24</td>
<td>203</td>
<td>213</td>
<td>5.1 ± 0.6</td>
<td>5.4 ± 0.5</td>
<td>&lt; 0.05</td>
</tr>
</tbody>
</table>

4.5.3 Comparison with clinicians

We assessed the accuracy of fetal birth weight predictions made by BabyNet++ and experienced clinicians. The study involved four gynecologists with 20, 15, 10, and 8 years of experience, respectively, who provided measurements at Centers A through D. Fetal biometric measurements were obtained using a US device during a routine US examination within 24 hours before delivery. The estimated fetal weight (EFW) was computed using the Hadlock IV formula, a heuristic formula utilized by clinicians and defined as:

\[
\log_{10} EFW = 0.3596 + (0.00061 \times BPD \times AC) \\
+ (0.0424 \times AC) + (0.174 \times FL) \\
+ (0.0065 \times HC) - (0.00386 \times AC \times FL),
\]

where HC, BPD, AC, and FL are head circumference, biparietal diameter, abdominal circumference, and femur length defined in centimeters [cm].

Table 4.4 shows variations among the four clinicians and BabyNet++ at each Center. Our study highlights the variation in fetal birth weight estimations among clinicians with different levels of experience. The most favorable
4.5. RESULTS

Table 4.5: This table compares the fetal birth weight prediction for different body parts and the entire fetus. The mean absolute error (MAE) in grams, root mean squared error (RMSE) in grams, and mean absolute percentage error (MAPE) in percent for each fetal body part and the entire body are displayed in the table. The results were obtained using the test set, and the best results are highlighted in boldface.

<table>
<thead>
<tr>
<th>Fetal body part</th>
<th>MAE [g]</th>
<th>RMSE [g]</th>
<th>MAPE [%]</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head</td>
<td>218 ± 19</td>
<td>262</td>
<td>6.2 ± 0.6</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Femur</td>
<td>204 ± 28</td>
<td>250</td>
<td>5.8 ± 0.8</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Abdomen</td>
<td>175 ± 17</td>
<td>200</td>
<td>5.0 ± 0.5</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>All</td>
<td>179 ± 19</td>
<td>203</td>
<td>5.1 ± 0.6</td>
<td>-</td>
</tr>
</tbody>
</table>

outcomes were provided by the most experienced clinicians (Center B and Center C), whereas the least accurate estimations were made by the least experienced clinicians (Center A). Nevertheless, we illustrate that BabyNet++ outperforms all clinicians, except for clinician A, for all the metrics that were evaluated.

Using the MAPE metric, paired t-tests revealed a statistically significant difference between BabyNet++ and clinicians from Centers B, C, D, and All (when all Centers were merged) (p < 0.05). However, for Center A, the p-value was 0.09, indicating that the difference between experts and automatic measurements was not statistically significant.

4.5.4 Impact of fetal body part on birth weight prediction

We performed statistical analysis of results achieved for all fetal body parts and results acquired from fetal US videos of each body part separately (i.e. head, femur, and abdomen) to demonstrate how the given body part affects birth weight prediction. The regression results presented in Table 4.5 demonstrated that the fetal head produced the largest errors in terms of predicting birth weight, while the abdominal video scans had the lowest impact. Specifically, the MAE, RMSE, and MAPE for the fetal head were 218 ± 19 g, 262 g, and 6.2 ± 0.6%, respectively, while the corresponding values for the abdominal
Table 4.6: Experimental results of ablation study with different configurations of key components of BabyNet++. The results of five-fold cross-validation on the test set are presented. The best results are shown in boldface.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>MAE [g] ↓</th>
<th>RMSE [g] ↓</th>
<th>MAPE [%] ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>302 ± 45</td>
<td>390</td>
<td>8.9 ± 1.2</td>
</tr>
<tr>
<td>Baseline + RTM</td>
<td>285 ± 39</td>
<td>374</td>
<td>8.5 ± 1.1</td>
</tr>
<tr>
<td>Baseline + DAFT</td>
<td>198 ± 40</td>
<td>245</td>
<td>5.7 ± 0.7</td>
</tr>
<tr>
<td>BabyNet++</td>
<td>179 ± 19</td>
<td>203</td>
<td>5.1 ± 0.6</td>
</tr>
<tr>
<td>(Baseline + RTM + DAFT)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Video scans were 175 ± 17 g, 200 g, and 5.0 ± 0.5%, respectively.

The one-way ANOVA revealed that there was a statistically significant difference between "All" and each body part separately. These findings suggest that analyzing fetal abdominal video scans can lead to more accurate birth weight prediction.

4.5.5 Effect of introducing clinical data upon FBW estimation

We conducted ablation studies to investigate the effectiveness of major components in the regression performance of BabyNet++. We implemented four configurations with five-fold cross-validation:

1. We trained a baseline network (3D ResNet-18) only on fetal US video scans.
2. We added the RTM with temporal position encoding to the baseline network and trained the network only on fetal US video scans (as before).
3. We added only the DAFT to the baseline network and combined clinical data with fetal US video scans.
4. We added both RTM and DAFT to complete our framework and trained the network with both fetal US video scans and clinical data.

The one-way ANOVA revealed that there was no statistically significant difference between “Baseline” and “Baseline+RTM”. However, when considering “Baseline”, “Baseline+DAFT”, and our proposed method, a statistically
significant difference ($p < 0.05$) was detected. The results for MAE, RMSE, and MAPE are presented in Table 4.6.

“Baseline” refers to the 3D ResNet-18 neural network. It was observed that the addition of the RTM module can help the model predict FBW and decrease MAE, RMSE, and MAPE. We can see that adding the DAFT module to the “Baseline” significantly outperformed both “Baseline” and “Baseline+RTM”. MAE and MAPE errors decreased by 104 g, 87 g, and 3.2%, 2.8% compared to “Baseline” and “Baseline+RTM”, respectively. This was due to the introduction of clinical data through the DAFT module which can efficiently embed clinical information in the model’s interior and dynamically rescaled and shift-generated feature maps, demonstrating the effectiveness of multimodal data. However, the DAFT module alone cannot guarantee the highest performance due to the lack of the MHSA mechanism that helps retain focus on the fetal body structure in US video scans.

Finally, our proposed model outperformed both standalone key components in all regression metrics, proving the significance and effectiveness of introducing the RTM module. It showed that the MHSA mechanism’s global learning features can direct attention to the multimodal feature maps, both from image and tabular data representations.

4.5.6 Impact of variability in fetal biometry measurements

This experiment aimed to investigate the impact of changes in fetal biometry measurements, such as head circumference (HC), biparietal diameter (BPD), abdominal circumference (AC), and femur length (FL), on fetal birth weight estimation by clinicians, BabyNet++, and the next best method, DAFT. To compare the impact of changes in fetal biometry parameters on fetal birth weight estimation we calculated the differences in estimations when modifying the parameters within a range of -5% to 5%. This involved artificially modifying fetal biometry measurements by adding or subtracting a small percentage (e.g., 1%) from the original measurements to assess the impact on fetal birth weight prediction.

In this study, we calculated the average difference in fetal birth weight estimation using DAFT, BabyNet++, and clinical methods across a sample of 194 independent patients. Clinicians used the Hadlock IV formula (see
Figure 4.8: The figure illustrates the impact of changes in fetal biometry measurements carried out in the course of fetal birth weight prediction upon performance loss (increase in Mean Absolute Percentage Error), with a comparison between the results obtained by clinicians (represented by the blue line), BabyNet++ (represented by the orange dotted line), and DAFT (represented by the green dotted line). Starting in the top left corner: a) Head Circumference (HC), b) Biparietal Diameter (BPD), c) Abdominal Circumference (AC), and d) Femur Length (FL). The biometry parameters varied between -5% to 5%, and the presented values are the mean results obtained from 38 independent patients. The plot highlights that BabyNet++ produces more accurate estimations of fetal birth weight than clinicians, as evidenced by the narrower and flatter slopes of the orange dotted lines. The blue line, representing the clinician’s estimations, shows a steeper slope, indicating greater sensitivity to changes in biometry measurements and suggesting a higher potential for errors. According to other results, measurements made on the abdominal plane have the greatest impact on performance loss in fetal birth weight estimation.
Eq. 4.6), while DAFT and BabyNet++ employed their prediction models. Figure 4.8 shows the differences between clinicians, DAFT, and BabyNet++. The findings indicate that BabyNet++ exhibits a lower degree of sensitivity in its predictive outcomes compared to the Hadlock heuristic formula, currently considered the gold standard approach used by clinicians, as well as DAFT, which is considered the next best method.

4.6 Discussion

In this work, we presented a Transformer-based framework for direct fetal birth weight prediction from clinical multimodal data acquired less than 24 hours before delivery, referred to as BabyNet++. Compared with previous deep learning methods, BabyNet++ achieves superior performance and can efficiently and automatically analyze multimodal data, including 2D+t spatio-temporal fetal US video scans and tabular clinical data.

To the best of our knowledge, we took the first step toward predicting fetal birth weight less than 24 hours before delivery, using fetal US video scans and clinical tabular data. The applied DAFT-based module was efficient in embedding clinical tabular data in the model’s interior and dynamically rescaling and shifting generated feature maps. The corresponding affine transformations extracted essential, imbalanced feature map representations from image sequences and tabular data. As shown in the ablation study summarized in Table 4.6, BabyNet++ took advantage of the 3D MHSA mechanism provided by the RTM. Combining DAFT and 3D MHSA provided by RTM gave rise to further improvements; therefore, both modules could be jointly used to improve fetal birth weight prediction performance.

The use of video analysis in fetal US examinations offered several advantages over static images, one of which was the ability to utilize 2D+t spatio-temporal feature representations that could enhance performance. However, conducting a fetal US examination 24 hours before delivery presented unique challenges due to changes in the fetal position, making it difficult to capture the standard plane of a particular fetal body part. The utilization of short video sequences in our method might obviate the need for expert knowledge and skills for the precise estimation of fetal birth weight, as it did not fully depend on a reference standard plane.
Our framework for fetal weight prediction has the potential to provide clinically relevant information. During validation, the presented algorithm reduced the mean error of expert clinicians by 21 g and 0.6% in MAE and MAPE, respectively. Specifically, our predictions constituted a notable improvement over the $5.1 \pm 0.6\%$ MAPE reported in literature [172; 180]. Moreover, we investigated the impact of each fetal body part on fetal birth weight prediction. Table 4.5 supported the hypothesis that the best results were obtained using video recordings of the fetal head, abdomen, and femur, along with their corresponding clinical tabular data. However, the difference between the fetal abdomen versus all body parts was relatively small. This demonstrated that fetal abdominal video scans alone might be sufficient to predict fetal birth weight within expert clinician error margins ($175 \pm 17$ g vs. $188 \pm 24$ g MAE, and $5.0 \pm 0.5\%$ vs. $5.4 \pm 0.5\%$ MAPE). These results were consistent with previously published reports on fetal birth weight prediction in [26; 96; 187]. This was noteworthy because it implied that a simpler screening method might be utilized with equal effectiveness. We will investigate this further in the course of our future work.

In the common practice of using the US to estimate fetal birth weight, standard prediction models have been shown to introduce errors of up to 15% [172]. This is consistent with other reported studies [180], demonstrating a MAPE of $10.1 \pm 7.1\%$. The clinicians’ estimations in the current study yielded a MAPE of $5.4 \pm 0.5\%$, with a maximum error of 11.5% which is significantly lower. A key distinction is the setup of the respective studies. While [172] data came from clinicians outside of tertiary referral centers, [180] research involved sonographers within tertiary hospitals. Uniquely, our centers integrate the expertise of both clinicians and the specialized environment of tertiary referral centers, overseeing about 4,000 deliveries annually, potentially contributing to the enhanced accuracy observed.

The measurement of abdominal circumference, which primarily represented soft tissue mass, could be affected by the status of the fetus [59]. Our hypothesis is that fat, which was less dense than the average fetal density, might affect birth weight predictions provided by clinicians who used available commercial tools. Several studies [37; 109] showed the effect of various combinations of fat and lean masses as the main causes of inaccuracies in birth weight prediction.

Figure 4.7 presented the FBW predictions provided by clinicians and
BabyNet++ for each of the patients. The fact that model-based prediction error was greater in cases of low birth weight and macrosomia might be due to the relatively small number of such cases in our training dataset, which resulted from the normal data distribution (see Figure 4.4). On the other hand, heuristic formulae used by clinicians seemed to underestimate FBW. We did not observe increased errors for higher FBWs, which were often reported in literature [137; 172]. The overestimation and underestimation errors followed the same pattern as in the case of clinicians’ predictions. This might be due to the application of the DAFT module inside our network, utilizing tabular data supplied by clinicians, which might have resulted in a similar impact on the ultimate predictions.

We demonstrated the impact of changes in biometry parameters on fetal birth weight estimations. To compare the stability of our approach with that of clinicians, we conducted experiments on a set of 38 independent patients. We computed performance loss while modifying biometry parameters within a range of -5% to 5% and compared our results with those provided by clinicians and their gold standard method, the Hadlock heuristic formula, as well as the next best method - DAFT. Our findings, as shown in Figure 4.8, indicated that our approach was less sensitive to changes than both the clinicians’ and DAFT methods. This demonstrated that our method not only relies on tabular (clinical) data of fetal biometry measurements but also on image feature representation, making it a valuable support tool for less experienced clinicians. Moreover, we demonstrated that the abdominal plane exhibited the most significant influence on performance loss due to variations in fetal biometry measurements in the estimation of fetal birth weight. Hence, a consistent pattern emerged when considering our other findings.

We presented selected frames from samples with the lowest and highest prediction errors within the test set, in Figure 4.9. Videos with the lowest FBW estimation error provided a clear view of the fetal body planes, which resulted in less noisy features extracted by BabyNet++ and more accurate biometry. As a result, these videos led to a more precise prediction. Conversely, the sample with the highest error exhibited noise and poor visibility of the anatomical structures.

In order to demonstrate BabyNet++’s capacity to generalize effectively to external datasets, we conducted an additional analysis. This assessment was
4. FETAL BIRTH WEIGHT PREDICTION ON FETAL MULTIMODAL DATA

Figure 4.9: This figure provides a visual representation of the cases in the test set that had the lowest (top) and highest (bottom) prediction error. The fetal head, abdomen, and femur are positioned from left to right, respectively. In the lowest error sample, the images provide a clear view of the fetal body planes. The underlying texture patterns of the image make the anatomical structures of the fetus easily identifiable, highlighting the precision and accuracy of the prediction model. However, in the bottom row, especially in the views of the abdomen and femur, poorly acquired images obstruct the analysis of the anatomical structures. Such limitations may result in inaccurate biometry acquisition, which, in turn, could decrease the performance of the model.

designed to unveil the model’s performance across a spectrum of data split configurations, thus elucidating its adaptability to various scenarios. The outcomes of this investigation have been reported in Table B.1 in the supplementary material, where we assessed the model’s performance by employing multiple training, validation, and test splits. These findings highlight the robustness of BabyNet++ and its potential to generalize. According to expectations, the settings in which test data from each center is represented in the training outperform settings in which this is not the case. Moreover, including large and diverse training data will likely further increase the robustness with respect to differences in the centers.

One of the limitations of the current work is that the method was developed and evaluated on a relatively small dataset. Diversifying training data by including additional device types and manufacturers, patient races, physical maternal characteristics, or twin pregnancies is also foreseen. In the future, a
larger dataset containing more operators from different skill groups and more scans per operator would be used. Furthermore, we developed our approach using short video clips, and we plan to test generalization on full-length routine US video scans. Due to computing hardware limitations, we had to use a 16-frame length segment of the entire sequence, which could potentially result in poorer outcomes. However, in the future, we intend to use the entire patient sequence as input to the neural network by employing a more lightweight model and improving the hardware setup.

Our method required good quality recordings: the short videos should not contain many frames with background artifacts or noise. Results suggested that the method displayed robustness against high levels of noise during training. However, further analysis was required to determine its efficacy in the presence of noise samples that had not been encountered before. We plan to address these limitations in the future. We also intend to extend the dataset by adding standard plane annotations as binary masks, as well as classification labels at the frame level. This could help increase the explainability of the proposed model. While our current work may have some limitations in terms of explainability, we plan to address them in future work to ensure a more transparent and understandable outcome.

4.7 Conclusions

In this paper, we introduced a novel network, BabyNet++, designed to predict FBW using multimodal data. Specifically, we used a custom RTM by adding DAFT to efficiently embed clinical data within the model structure.

We conducted experiments using a novel clinical fetal dataset, showing that the proposed method outperforms existing state-of-the-art approaches. Moreover, our method also outperformed expert clinicians who used commercial tools to estimate FBW. We demonstrated that the use of standalone abdominal video scans results in the best FBW prediction performance.

Additionally, BabyNet++ proved to be less sensitive to clinical data variabilities compared to heuristic formulae used currently as a standard of care making it a potentially supportive tool for less experienced clinicians. Our method has the potential to be applied in the clinical environment to assist in the selection of the safest type of delivery for both the mother and the child.
Supplementary material

Table A.1: This table presents the results of pairwise MAPE comparisons using Tukey’s HSD test for various groups. The groups are compared based on their mean differences, adjusted p-values, confidence intervals, and the decision to reject or not reject the null hypothesis of equal means. The ”Reject” column indicates whether the null hypothesis is rejected (True) or not rejected (False) based on the adjusted p-value ($p < 0.05$). LR and GBR denote Linear Regression, and Gradient Boost Regressor, respectively.

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>Mean diff</th>
<th>p-adj</th>
<th>Lower</th>
<th>Upper</th>
<th>Reject</th>
</tr>
</thead>
<tbody>
<tr>
<td>BabyNet++</td>
<td>BabyNet</td>
<td>3.3579</td>
<td>0.014</td>
<td>0.385</td>
<td>6.3308</td>
<td>True</td>
</tr>
<tr>
<td>BabyNet++</td>
<td>Clinicians</td>
<td>0.2827</td>
<td>1.0</td>
<td>-2.6903</td>
<td>3.2556</td>
<td>False</td>
</tr>
<tr>
<td>BabyNet++</td>
<td>DAFT</td>
<td>1.0664</td>
<td>0.9709</td>
<td>-1.9066</td>
<td>4.0393</td>
<td>False</td>
</tr>
<tr>
<td>BabyNet++</td>
<td>Interactive</td>
<td>0.6969</td>
<td>0.9983</td>
<td>-2.276</td>
<td>3.6699</td>
<td>False</td>
</tr>
<tr>
<td>BabyNet++</td>
<td>LR</td>
<td>0.7034</td>
<td>0.9982</td>
<td>-2.2695</td>
<td>3.6764</td>
<td>False</td>
</tr>
<tr>
<td>BabyNet++</td>
<td>3D ResNet-18</td>
<td>3.7555</td>
<td>0.0031</td>
<td>0.7826</td>
<td>6.7285</td>
<td>True</td>
</tr>
<tr>
<td>BabyNet++</td>
<td>ViViT</td>
<td>6.7565</td>
<td>0.0021</td>
<td>3.7836</td>
<td>9.7295</td>
<td>True</td>
</tr>
<tr>
<td>BabyNet++</td>
<td>GBR</td>
<td>0.9198</td>
<td>0.9886</td>
<td>-2.0531</td>
<td>3.8928</td>
<td>False</td>
</tr>
</tbody>
</table>

Table B.1: This table presents the evaluation results for various data split configurations. The ”Split” column indicates the specific split used for training and validation, while the ”Test” column indicates the corresponding test split.

<table>
<thead>
<tr>
<th>Split</th>
<th>Training and validation</th>
<th>Training and validation size</th>
<th>Test</th>
<th>Test size</th>
<th>MAE [g] (\downarrow)</th>
<th>RMSE [g] (\downarrow)</th>
<th>MAPE [%] (\downarrow)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configuration 1</td>
<td>A, B, C</td>
<td>147</td>
<td>D</td>
<td>47</td>
<td>186 ± 20</td>
<td>210</td>
<td>5.3 ± 0.7</td>
</tr>
<tr>
<td>Configuration 2</td>
<td>A, B, D</td>
<td>142</td>
<td>C</td>
<td>52</td>
<td>190 ± 21</td>
<td>215</td>
<td>5.4 ± 0.6</td>
</tr>
<tr>
<td>Configuration 3</td>
<td>A, C, D</td>
<td>178</td>
<td>B</td>
<td>16</td>
<td>184 ± 23</td>
<td>207</td>
<td>5.2 ± 0.8</td>
</tr>
<tr>
<td>Configuration 4</td>
<td>C, B, D</td>
<td>115</td>
<td>A</td>
<td>79</td>
<td>204 ± 24</td>
<td>231</td>
<td>5.8 ± 1.0</td>
</tr>
<tr>
<td>Configuration 5</td>
<td>0.8×(A, B, C, D)</td>
<td>156</td>
<td></td>
<td>38</td>
<td>179 ± 19</td>
<td>203</td>
<td>5.1 ± 0.6</td>
</tr>
</tbody>
</table>