AI-Based Normative Modeling of Fetal Heart Rate for Gestational Age Prediction and Detection of Developmental Deviations

Giulio Steyde[1], Chiara Gorla[1], Giovanni Magenes[2], Maria G Signorini[1]

[1] Politecnico di Milano, Milano, Italy, [2] Università di Pavia, Pavia, Italy

Abstract

Cardiotocography (CTG) is the continuous recording of the fetal heart rate (FHR) using ultrasound transducers placed on the mother's abdomen. The signal obtained exhibits characteristic patterns that allow for the assessment of the fetus's well-being, behavioral states, and nervous system development. Information on the gestational age (GA) is crucial in fetal state and development evaluation.

The objective of this work is to train a model to predict the GA from the FHR in healthy pregnancies, and assess whether the same model can identify deviations from normal development due to Intra Uterine Growth Restrictions (IUGR). The chosen AI model is based on a modified ResNet architecture and was trained with FHR signals from healthy subjects. The model is then tested on both Healthy and IUGR fetuses. This paradigm offers advantages such as eliminating the need for class balancing and opens up potential applications in various pathological conditions. The use of these AI methods was enabled by the NAPAMI database, from which we used 63342 signals over more than 80000 available.

Preliminary results (R^2: 0.50 in healthy and R^2: 0.16 in IUGR) show that the proposed method is capable of detecting deviations from normal fetal development. We are working to improve robustness, interpretability, and performance to contribute to its applicability in real-world clinical settings.

1. Introduction

Cardiotocography (CTG) is a standard clinical exam used to monitor the fetal well-being during pregnancy by recording the continuous fetal heart rate (FHR) trace [1]. The interpretation of this signal allows understanding fetal well-being states, possible signs of pathological complications, and the level of autonomic maturation. Intrauterine growth restriction (IUGR) is a common pregnancy complication defined by the failure of the fetus to fulfill its growth potential. Clinically, the IUGR term refers to fetuses whose estimated fetal weight is lower than the 10th percentile for the identified Gestational Age (GA) and show alterations in the flow of the umbilical artery [2]. The antepartum diagnosis of IUGR is inherently uncertain, and requires confirmation at birth.

Research conducted in recent years in cardiotocography has led to the development of a series of quantitative indices derived from the FHR, which have demonstrated good ability to classify the condition of the fetus in utero at various gestational weeks [3, 4]. Several of these features have been shown to correlate with the GA [4–7].

Other works tried to estimate the GA from FHR traces [8,9]. Among them, Hoyer and colleagues [8] used magnetocardiographic recordings, which require very expensive and cumbersome instrumentation inadequate for clinical practice, and manually annotated fetal behavioral states, thus severely limiting the model's applicability. On the other hand, Gu et al. [9] attempted to predict GA from standard CTG recordings but, despite using over 7,000 CTG traces for model training, only a moderate correlation between real and predicted GA was found (Pearson r: 0.434, mean absolute error: 10.91 days).

Being able to predict the GA from the FHR has one logical application [8, 9]: use the prediction (and its deviation from the actual GA) to assess whether the fetus is lagging behind the *normal* development. This would provide a functional biomarker for the presence of Intra-Uterine Growth Restriction and its severity.

Our work exploits the availability of a very large semi-annotated database of antenatal CTG recordings, called NAPAMI, representing a promising opportunity for the use of Deep Learning (DL) techniques to predict the GA from the CTG signals, and to infer possible developmental delays.

The intent is to evaluate whether the model is able to characterize the normal development and detect deviations in abnormal cases. The DL technique is expected to learn to predict the GA of healthy fetuses and consequently provide a measure of possible growth impairments. This could increase the support given to clinicians to make better-informed decisions.

2. Methods

2.1. Data

NAPAMI is a very large database of antepartum CTG recordings, which has been described in its preliminary form in [10]. Currently, NAPAMI contains 81531 unique recordings from 33208 distinct pregnancies. Exams were performed in Naples, Italy as part of routine clinical practice. All recordings consist of the FHR trace, along with uterine contractions and fetal movements (which are not used in this work), sampled at 2 Hz and obtained using cardiotocographs from the Philips Avalon family. Moreover, recordings are associated with a free-text note, which is used by clinicians to annotate fetal and maternal pathologies and other relevant clinical information.

CTG signals underwent basic preprocessing [7] and all signals with more than 20% of low-quality samples were excluded. Based on clinical annotations, the following groups were extracted:

- Healthy. N = 12433
- Confirmed IUGR. N = 1285
- Unknown. N = 49624

Pregnancies are denoted as *Healthy* when no fetal or maternal pathologies were identified by the clinical team, and the weight of the newborn was higher than the 10th percentile for the GA. *Confirmed IUGR* presented an estimated fetal weight lower than the threshold in the anterpartum period, had alterations in the umbilical artery flow and were born with a weight below the 10th percentile for the GA. *Unknown* presented no clinical annotation. Twin pregnancies and those not belonging to the specified groups were excluded from the analysis.

A hold-out test set composed of 150 *Healthy* and 150 *Confirmed IUGR* pregnancies with the same, and approximately uniform, distribution with respect to the GA was first sampled. More in detail, 9 bins uniformly spaced between 160 and 310 days of gestation were defined, and from each, the same number of signals for healthy and IUGR was sampled. Any additional recordings associated with pregnancies represented in the hold-out test set were excluded from the training set.

Since the final goal is to develop a *normative* model which can identify deviations from expected development, only Healthy and Unknown cases have been used in training (with the latter being assigned a lower weight, as will be discussed in the next paragraph). These were divided into train and validation using a 90:10 split.

All signals were then divided into windows of 30 minutes with a maximum overlap of 50%. In total, 65130 30-minute windows are used for training and 6931 for validation. In testing, predictions performed on different windows of the same signal were averaged to obtain a unique prediction per recording.

2.2. Model Development

A ResNet-based architecture (hereafter simply called "ResNet"), similar to the one described in [11], was used. The model is composed of three residual blocks with convolutional kernels of sizes 8, 5, and 3, respectively. The initial block is configured with 32 feature maps, while the subsequent two blocks utilize 64 feature maps. Each block integrates batch normalization, ReLU activation, a residual shortcut connection, and a squeeze-and-excitation recalibration block, which models interdependencies across channels [12]. The convolutional encoder is followed by a global average pooling layer and a dropout layer ($p = 0.3$). The resulting feature vector is processed by a coarse-to-fine regression head that first estimates logits over four GA bins and subsequently predicts a within-bin residual, yielding the final continuous GA estimate. This strategy was adopted to prevent regression towards the mean.

During training, each window was assigned a composite sample weight to (1) emphasize reliable windows, (2) prioritize healthy cases, (3) down-weight unknown cases, and (4) increase the contribution of under-represented GA ranges.

Thus, the sample weight is equal to the product of three terms: (1) a quality weight, (2) a status weight, and (3) a GA density weight. The quality weight (1) was defined as the fraction of samples in the window marked as good quality. The status weight (2) was used to down-weight uncertain labels, assigning weight 1.0 to healthy windows and 0.2 to unknown-status windows. As previously discussed and illustrated in Figure 1, the target variable GA exhibits a heavily left-skewed distribution. To mitigate this imbalance, we applied an inverse-density weight (3) calculated across four uniformly spaced bins. The per-bin probability p was converted to a weight proportional to $w_{i\_\text{raw}} = (p_b(i) + 10^{-8})^{-0.6}$, normalized to unit mean, and clipped to the range [0.5, 2.5].

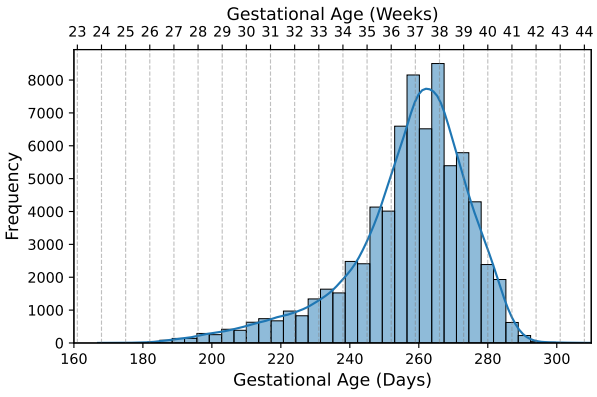

Figure 1. Gestational age (GA) distribution in the NAPAMI database.

3. Results

Figure 2 shows the scatterplot of the true versus predicted GA values for both healthy and IUGR cases in the hold-out test set.

Figure 3 presents the Bland-Altman plot, and relevant numerical values are summarized in Table 1.

As can be seen in Figure 2, the ResNet model identified a moderate relationship between the FHR and the GA target, with an R^2 of 0.501 and a Pearson correlation coefficient r of 0.743. The mean absolute error was 13.6 days.

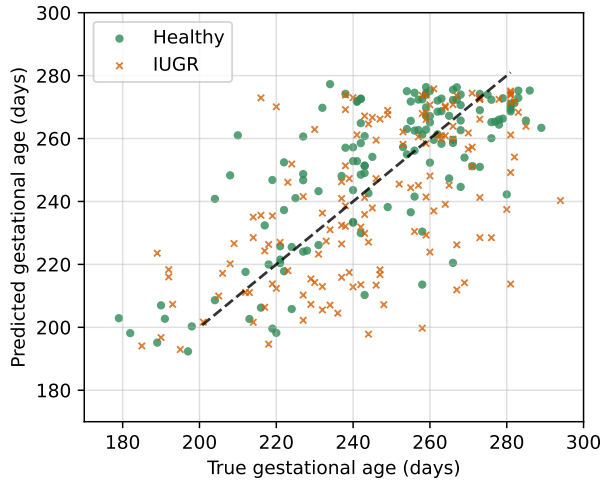

Figure 2. Scatterplot illustrating the relationship between true and predicted gestational age (GA) for Healthy (green dots) and IUGR (orange crosses) cases using ResNet.

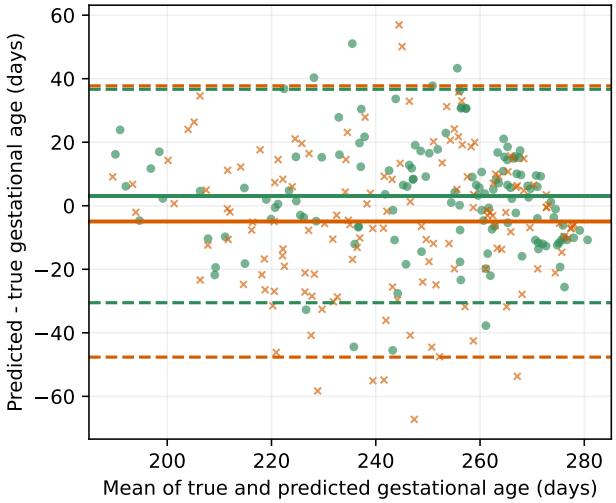

Figure 3. Bland-Altman plot assessing the prediction bias and limits of agreement between true and predicted gestational age (GA) using ResNet.

Table 1. ResNet performance metrics.

Condition	Pearson r	R^2	MAE	p-value
Healthy	0.743	0.501	13.6	0.0171
IUGR	0.601	0.161	17.3	

As reported in Figure 3, the Bland-Altman plots reveal that the prediction errors are consistently distributed, even though the Healthy cohort presents a small positive bias. This is likely due to the training set being more severely skewed than the test set, which was sampled to be as uniform as possible with respect to the target.

The model showed higher prediction error on the IUGR cohort. As can be seen in Fig 3, the ResNet model shows a tendency to predict lower GA values for the IUGR cohort, which aligns with the expectations. The distributions of the prediction errors (not normalized) were also compared between Healthy and IUGR by means of the Wilcoxon rank-sum test, resulting in a p-value of 0.0019.

4. Discussion

The correlation values obtained clearly show that the model has learned a relationship between FHR characteristics and GA in the healthy population, with a Pearson r of 0.743 and a mean absolute error of 13.6 days.

While the error value can appear high, it must be considered that the reference for the gestational age itself has an intrinsic level of uncertainty, since in NAPAMI the GA is based on the last menstrual period. At present, there are no robust references in the literature regarding what a reasonable predictive value might be. Therefore, the model's indication should be considered a data-driven starting point.

The ResNet model shows a significant tendency to predict lower GA values for IUGRs, even though the magnitude of this difference is relatively small. This may be a consequence of sub-optimal model performance, or it could be related to the fact that in many cases IUGR is poorly described as a simple *delay* in maturation, but includes more complex pathophysiological adaptations which may differ radically according to the etiology of the condition.

Comparing the results with other works is complicated by the heterogeneity of the data used in fetal monitoring. In particular, magnetocardiography seems to provide cleaner signals and better results [8], but is infeasible in clinical practice and limits the usefulness of the results. Additionally, Shirazi and colleagues [13] addressed the task of predicting GA using as input to the model the 1-dimensional Doppler ultrasound signal recorded with a low-cost device and achieved impressive results, with a MAE lower than a week, which is in line with the error of

conventional biometry-based methods in early pregnancy. While promising, this type of data is not part of clinical practice.

To the best of our knowledge, the only work that used standard CTG recordings with a comparable number of cases is the one by Gu and colleagues [9], who nonetheless achieved lower performances than those presented in this work.

This suggests that, for this application, the quality of the input signal is crucial to the final performance. It is also possible that the obtained estimation could be improved only by adding additional information from different sources. For example, Shirazi et al [13] used the ultrasound signal directly, which includes information on the maternal heart rate and respiration, thus providing other inputs that are not typically considered in standard fetal monitoring.

Based on the results presented in this paper, the goal is to strengthen and improve classification by refining the models and incorporating additional information. We believe that integrating different signals and methods is the approach we should pursue and develop.

Acknowledgments

The authors would like to acknowledge the entire team that over the years contributed to the creation of the NAPAMI database and the development of the computational methods for FHR quantitative analysis.

References

[1] ACOG Practice Bulletin, Number 229. Antepartum fetal surveillance. Obstetrics Gynecology June 2021; 137(6):e116–e127.

[2] Gordijn SJ, Beune IM, Thilaganathan B, Papageorghiou A, Baschat AA, Baker PN, Silver RM, Wynia K, Ganzevoort W. Consensus definition of fetal growth restriction: a delphi procedure. Ultrasound Obstet Gynecol 9 2016;48(3):333–339.

[3] Signorini M, Pini N, Malovini A, Bellazzi R, Magenes G. Integrating machine learning techniques and physiology based heart rate features for antepartum fetal monitoring. Computer Methods and Programs in Biomedicine 2020;185:105015.

[4] Signorini MG, Fanelli A, Magenes G. Monitoring fetal heart rate during pregnancy: Contributions from advanced signal processing and wearable technology. Computational and Mathematical Methods in Medicine 2014; 2014(1):707581.

[5] Van Leeuwen P, Cysarz D, Edelhäuser F, Grönemeyer D. Heart rate variability in the individual fetus. Auton Neurosci 11 2013;178(1-2):24–28.

[6] Amorim-Costa C, Costa-Santos C, Ayres-de Campos D, Bernardes J. Longitudinal evaluation of computerized cardiotocographic parameters throughout pregnancy in normal fetuses: a prospective cohort study. Acta Obstet Gynecol Scand October 2016;95(10):1143–1152.

[7] Signorini M, Magenes G, Cerutti S, Arduini D. Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings. IEEE Transactions on Biomedical Engineering 2003;50(3):365–374.

[8] Hoyer D, Tetschke F, Jaekel S, Nowack S, Witte OW, Schleußner E, Schneider U. Fetal functional brain age assessed from universal developmental indices obtained from neuro-vegetative activity patterns. PLoS One September 2013;8(9):e74431.

[9] Gu J, Lin Z, Ma J, Wang J, Zhang L, Bai R, Tu Z, Jiang Y, Xie D, Zhou Y, Liu G, Hong S. Artificial intelligence-derived cardiotocography age as a digital biomarker for predicting future adverse pregnancy outcomes, 2025. URL doi.org/10.48550/arXiv.2509.14242.

[10] Spairani E, Daniele B, Magenes G, Signorini MG. A novel large structured cardiotocographic database. In 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2022; .

[11] Spairani E, Steyde G, Spuri Forotti F, Magenes G, Signorini MG. Prediction of iugr condition at birth by means of ctg recordings and a resnet model. Computers in Biology and Medicine 5 2025;190:110123.

[12] Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-excitation networks, 2017. URL doi.org/10.48550/arXiv.1709.01507.

[13] Motie-Shirazi M, Ramos E, Chandrasekaran S, Sameni R, Rohloff P, Clifford G, Katebi N. Self-supervised learning for gestational age estimation from low-cost doppler ultrasound in low-resource settings. In NeurIPS 2025 Workshop on Learning from Time Series for Health. 2025; .

Address for correspondence:

Maria G Signorini, Ph.D.
Politecnico di Milano, Department of Electronics, Information and Bioengineering - DEIB. Piazza Leonardo da Vinci, 32, 20133 Milano MI, Italy
mariagabriella.signorini@polimi.it