Article

Use of Deep Learning to Detect the Maternal Heart Rate and False Signals on Fetal Heart Rate Recordings

Samuel Boudet[1,*], Agathe Houzé de l'Aulnoit[2], Laurent Peyrodie[3], Romain Demailly[4] and Denis Houzé de l'Aulnoit[5]

[1] Faculty of Medicine and Midwifery, ETHICS EA 7446, Lille Catholic University, F-59000, Lille, France
[2] Lille Catholic Hospital (Obstetrics Department), Lille Catholic University, F-59020, Lille, France
[3] Junia Haut de France, F-59000, Lille, France
* Correspondence: samuel.boudet@univ-catholille.fr

Abstract: We have developed deep learning models for automatic identification of the maternal heart rate (MHR) and, more generally, false signals (FSs) on fetal heart rate (FHR) recordings. The models can be used to preprocess FHR data prior to automated analysis or as a clinical alert system to assist the practitioner. Three models were developed and used to detect (i) FSs on the MHR channel (the FSMHR model), (ii) the MHR and FSs on the Doppler FHR sensor (the FSDop model), and (iii) FSs on the scalp ECG channel (the FSScalp model). The FSDop model was the most important because FSs are far more frequent on the Doppler FHR channel. All three models were based on a multilayer symmetric gated recurrent unit and were trained on data recorded during the first and second stages of delivery. The FSMHR and FSDop models were also trained on antepartum recordings. The training dataset contained 1030 expert-annotated periods (mean duration: 36 min) from 635 recordings. In an initial evaluation of routine clinical practice, 30 fully annotated recordings for each sensor type (mean duration: 5 h for MHR and Doppler sensors, and 3 h for the scalp ECG sensor) were analyzed. The sensitivity, positive predictive value (PPV) and accuracy were respectively 62.20%, 87.1% and 99.90% for the FSMHR model, 93.1%, 95.6% and 99.68% for the FSDop model, and the 44.6%, 87.2% and 99.93% for the FSScalp model. We built a second test dataset with a more solid ground truth by selecting 45 periods (lasting 20 min, on average) on which the Doppler FHR and scalp ECG signals were recorded simultaneously. Using the scalp ECG data, the experts estimate the true FHR value more reliably and thus annotated the Doppler FHR channel more precisely. The models achieved a sensitivity of 53.3%, a PPV of 62.4%, and an accuracy of 97.29%. In comparison, two experts (blinded to the scalp ECG data) achieved a sensitivity of 15.7%, a PPV of 74.3%, and an accuracy of 96.91% for expert 1 and a sensitivity of 60.7%, a PPV of 83.5% and an accuracy of 98.24% for expert 2; hence, the model performed better than one expert and worse than the other. Hence, the models performed at expert level, although a well-trained expert with good knowledge of FSs could probably do better in some cases. The models and datasets have been included in the Fetal Heart Rate Morphological Analysis open source MATLAB toolbox and can be used freely for research purposes.

Keywords: fetal heart rate; maternal heart rate; cardiotocogram; gated recurrent unit; deep learning

1. Introduction

The fetal heart rate (FHR) is a key parameter for monitoring fetal well-being during pregnancy, labor, and delivery. Accurate interpretation of the FHR is important for avoiding unnecessary cesarean sections and instrumental delivery (forceps or suction cup and reducing the risk of fetal acidosis. In France, the FHR is always recorded during delivery and (for at-risk pregnancies) before delivery (antepartum).

The FHR signal is analyzed by midwives and obstetricians for abnormalities such as decelerations, low variability, bradycardia, tachycardia, and sinusoidal patterns. The International Federation of Gynecology Obstetrics (FIGO) has issued guidelines on FHR analysis [1].

The FHR is measured with a Doppler sensor or a scalp electrocardiogram (ECG) sensor. The Doppler is noninvasive, whereas the ECG sensor requires an incision on the fetus'

doi:10.20944/preprints202207.0131.v1

scalp (associated with a small degree of risk) and can be slightly uncomfortable for the mother. Moreover, a Doppler

sensor can be used at any time during pregnancy, whereas a scalp ECG electrode can only be used after rupture of the membranes and often fall during second stage of delivery. However, the Doppler sensor is less accurate – particularly when the FHR is highly variable – and is more subject to missing signal (MS) and false signals (FSs). This is why a Doppler sensor is used first and then replaced by a scalp ECG sensor when FS ambiguities or MS are present or when FHR variability must be measured accurately [2]. In our maternity hospital, a scalp ECG electrode is used in around 10% of deliveries, although this proportion will vary from one institution to another. The FHR can be also measured through abdominal ECG sensors, although this technique falls outside the scope of the present study [3].

Using the raw data from the Doppler sensor or the scalp ECG sensor, the FHR is calculated on the cardiotocograph (CTG) monitor by applying a proprietary algorithm based on auto-correlation whose details are hidden. Those algorithms can often output false values corresponding to a harmonic of the true FHR value: double the rate, half the rate, or (more rarely) the triple the rate. Moreover, the Doppler FHR sensor often records the maternal heart rate (MHR) or a harmonic of the latter rather than the FHR, which is then considered to be an FS. Lastly, noisy raw data can give rise to random values but only for a few seconds. All these signals will be referred to as FSs here. Although scalp ECG sensors can reportedly sometimes measure the MHR (particularly in cases of fetal death) [4], we have not found any examples in on our dataset. FSs on scalp ECGs mainly correspond to rare, short periods of random values.

In most cases, an expert can easily identify FSs. However, this process is critical because some FSs look like pathologic FHR signals: for example, a switch from an FHR value to an MHR value may look like a deceleration or bradycardia in the fetus. Moreover, maternal tachycardia can produce a signal that looks like the FHR, and so the true FHR might not be analyzed for several hours. Misinterpretation of the MHR as the FHR is relatively common. According to Reinhard et al. [5], MHR periods are found in up to 90% of intrapartum recordings and account for 6.2% of the recording. This problem is particularly frequent during the second stage of delivery and can be particularly dangerous for the fetus [6,7]. For example, France's Melchior classification used to describe the second stage of delivery [8] proposed erroneously the type 3 FHR pattern (corresponding to bradycardia plus accelerations synchronized with the uterine contractions (UC), and for which expulsive efforts should be less than 15 min) whereas all those cases (representing ≈ 4% of deliveries) are in fact FS of MHR [9,10]. This shows the extend of the problem.

To help identify FSs, modern cardiotocographs have an MHR sensor: either an ECG sensor on combined with the tocometer sensor on the belt, or an oximeter on the mother's finger. Superposition of the MHR and the FHR suggests that the latter is an FS, although natural coincidences can occur. Moreover, MHR channel has often periods of MSs and FSs (generally harmonics double, triple or half the true rate), particularly during the second stage of delivery.

To avoid ambiguity, we defined signal loss or a missing signal (MS) as a period during which the CTG device did not send MHR or FHR values. Furthermore, we defined an FS as any measured signal that did not correspond to the sensor's target heart rate (i.e. periods of MHR recorded by a Doppler FHR sensor, signal harmonics, or other aberrant values). We did not consider that an inaccurate heart rate was an FS, even though this distinction is not always evident.

The practitioner can use several clues to differentiate between FSs and true signals (TSs), as summarized in Appendix A. Many practitioners are apparently unaware of some of these clues, and so their ability to differentiate between FSs and TSs could probably be improved.

FSs impede also the automatic analysis of FHR recordings. Several automatic methods for FHR analysis have been developed in the last few years, notably with a view to preventing acidosis during delivery. Most of these methods include a preprocessing step in which FSs are partially removed [11]. This generally consists in detecting short periods during which the measured rate is significantly higher or lower (by more than 25 bpm, typically) than in the preceding period. We are not aware of any publications on the accurate identification of long periods of MHR – particularly when the MHR signal is at least partly missing. The closest reference to this problem was made by Pinto et al. [12]], who considered there is MHR-FHR ambiguities when the difference between MHR and FHR is less than 5 bpm and then removed those periods. However, this method cannot work when MHR is missing, when the FHR coincides with MHR and when the FS are harmonics of FHR or MHR. Pinto et al.'s results nevertheless emphasize the significant bias in automated FHR feature detection introduced by periods of MHR recording.

The software in CTG monitors comprises also sometime an alert system based on coincidence between MHR and FHR channels[7]. However, these coincidences have to be checked manually. Moreover, no alarm is given if the MHR sensor does not record a signal during this period. This problem is often neglected in literature [13]) because it merely constitutes a preprocessing step. However, FS is known to be a major source of error in FHR analysis, whether visual [14,15] or automatic [12].

Deep learning (DL) models have recently emerged in which the FHR is directly used as input [16]. There were no prior feature extraction steps, and features could emerge automatically from the models' architectures and the clinical outcome (often the arterial umbilical cord blood pH) used as output. Thus, one could imagine that the concept of an FS could emerge from a DL model. However, given (i) the relatively weak link between the FHR and the clinical outcome and (ii) the complexity of the problem, we suspected that even several tens of thousands or hundreds of thousands of delivery recordings might not be enough to trigger the emergence of such complex features. We therefore sought to design models for these specific tasks. In particular, we sought to determine a solid ground truth for FSs (i.e. better than an expert could produce using only the same information that the models) and thus increase the models' efficiency.

The objectives of the present study were to develop the first intelligent methods for the automatic recognition of FSs and to present this problem to other signal processing researchers. We introduced three models for the detection of FSs on CTG recordings. Firstly, the FSMHR model detected FS periods on MHR channel (either from ECG sensor on the tocometry belt or from the finger oximeter). Secondly, the FSDop model detected MHR and FS periods on Doppler FHR channel. Thirdly, the FSScalp model detected FS periods on scalp ECG channel. All three models have been established for the first and second stages of delivery, and the FSMHR and FSDop models have also been established for antepartum recordings. FSDop is the most complicated and clinically important model, in view of the several possible sources of FSs (harmonics, the MHR, and other aberrant signals).

The work described below was based on recordings from Philips CTG monitors. Although our methods should apply to other brands, some of the preprocessing steps are probably brand-specific and would have to be adjusted for use with other monitors.

Application of these models might help to (i) improve automatic FHR analysis methods (particularly for acidosis prediction during delivery) (ii) develop a smart alert system that tells the practitioner to reposition the FHR sensor or replace it with a scalp ECG electrode, and (iii) indicate when the FHR sensor is in fact measuring the MHR (thus avoiding potentially dangerous misinterpretations for the fetus).

Below, we describe the models, their training, the two evaluation methods, and the results.

2. Description of the models

2.1. Data acquisition

The three models were trained and evaluated on data recorded at the Saint Vincent de Paul Maternity Hospital (Lille, France) and stored in the "Bien Naître" data warehouse (registered with the French National Data Protection Commission; reference: REG 077)). At

doi:10.20944/preprints202207.0131.v1

the time of writing, this data warehouse contained 22,000 delivery recordings (recorded from 2011 onwards) and 5,000 antepartum recordings (recorded from 2019 onwards) [17]. The database's research objectives and procedures were approved by the local institutional review board (CIER GHICL, Lille, France) on August 4th, 2016 (reference: 2016-06-08). Each woman was informed about the inclusion of her newborn's data in the data warehouse and gave her written consent to the storage and use of these data.

The CTG monitors (Avalon FM30 and FM20®, Philips Medical Systems, Best, The Netherlands) sent signals to a central, dedicated research server via an Ethernet or WiFi connection, using in-house solution [18].

All heart rate signals were acquired at a frequency of 4 Hz and a resolution of 0.25 bpm (or 1 bpm, for MHRs recorded with an oximeter). The tocometer signal was recorded at 4 Hz and at a resolution of 0.5 mmHg. MHR signals have only been recorded in our maternity hospital since April 2015; given the importance of the MHR to detect FS, we did not include data recorded before this date.

2.2. Selection of the training dataset

Recordings were extracted from the Bien Naître database and used to train and validate the models. These recordings were also annotated by experts, constituting the ground truth. The training/validation dataset was composed of periods from 635 recordings (dating from April 1 st, 2015, to December 31st, 2019) selected as follows:

A 94 perpartum recordings, corresponding to all the recordings containing periods of at least 10 min during which signals from scalp ECG and Doppler sensors were recorded simultaneously. These recordings were used to train FSDop, FSMHR and FSScalp. In the event of doubt, solid ground truth can be determined by annotating experts using the other sensor's signal as indicator (i.e. the scalp sensor for FSDop and the Doppler sensor for FSScalp). These recordings are also of value for the current problem because if practitioners has positioned the two sensors, it is probably because the recordings show FS ambiguities.

B 38 antepartum recordings presenting marked variations (either FHR/MHR switches or decelerations), for training the FSDop and FSMHR models.

C 96 routine perpartum recordings annotated by the practitioner as having major FSs and MSs (for training the FSDop, FSScalp and FSMHR models).

D 107 perpartum recordings (all recorded in 2016) with data from scalp ECG electrode but (in contrast to dataset A) that lacked simultaneously recorded Doppler data. Nonetheless, these recordings usually incorporated a first part of Doppler signals with often MSs or FS ambiguities and were also analyzed. Thus, this dataset was for training the FSDop, FSScalp and FSMHR models.

E 300 perpartum recordings with Doppler signal, selected for their interest by experts, from among the 915 recordings made in 2016 but not already included in the datasets A to D (for training the FSDop and FSMHR models).

On each recording, the experts selected at least one periods ranging from 5 min to 4 h in duration. These periods were used to train the FSDop and/or FSMHR models on one hand or the FSScalp model on the other. Within these selected periods, the experts annotated segments with clearly TSs and segments with clearly FSs. A lack of annotation meant that either the expert was unsure or the period was not difficult enough for the problem and so was not worth annotating.

A high proportion of the selected period might not have been annotated. The selected period was long enough to include all the information required for interpretation. For example, we assumed that an MHR captured by a Doppler FHR sensor for one minute was an FS. During this minute and the preceding 30 min, the MHR sensor was not positioned. The selected period had to include all the signals starting a few minutes before the MHR sensor was removed, so that the model could estimate the range of possible MHRs during the annotated period.

On datasets (A), (B) and (C), the selected periods were fully annotated by experts blinded to the models' outputs. Datasets (D) and (E) were annotated after the initial models had been trained and thus only parts that appeared to be more difficult or where the models were wrong (or not confident enough) were annotated. Thus, the dataset grew progressively as the models' complexity increased.

Recordings were randomly attributed to the training dataset (80% of the total duration) or the validation dataset (20%). This attribution was performed for the FSDop/FSMHR datasets and then separately for the FSScalp datasets. The validation dataset was used for early stopping of training avoiding overfitting, and comparing the various models and hyper-parameters. The final evaluation was carried out with the datasets described in section 3.

2.3. The interface for expert annotation

We created a dedicated MATLAB® (R2021b) interface (Figure 1), based on the viewer from the Fetal Heart Rate Morphological Analysis (FHRMA) toolbox [19]) to display results and enable the experts to annotate TSs and FSs. The expert drew a rectangular window to precisely select the beginning and end of FS and TS periods. The interface could display an interpellated signal during MS as a lighter line and thus could emphasize the short FS periods, which would be only a few pixels on the screen and could easily have been overlooked. The interface shown in the figure is for the MHR and Doppler FHR annotation at the same time. Another interface was dedicated to the analysis of FHR Scalp ECG channel which have selection periods independent from the Doppler/MHR ones.

The training datasets were analyzed by one or two (consensually) experts: an engineer and a senior obstetrician, both of them performed research on and gave lectures on FHR analysis. The role of engineer was important for a faster and more accurate interface using and for a better understanding of (i) how an FS can arise and (ii) what is important for machine learning.

The interface could also display a model's results by using a color gradient for the FHR (TS=blue, FS=red) and the MHR (TS=violet, FS=cyan). Thus, the experts could annotate according to where the first models did mistake.

The interface is available as open-source code in the FHRMA MATLAB® toolbox [19]. The FHRMA toolbox also included the model recoded in MATLAB ® because the training was performed with Python® 3.0 and TensorFlow® 2.4 (see section 2.8.7).

2.4. Manufacturer-specific preprocessing

The CTG monitor measures the FHR via a Doppler sensor or a scalp ECG electrode and measures the MHR via an ECG sensor combined with the tocometer on the belt or an oximeter on the finger. In all cases, the CTG monitor applies a proprietary auto-correlation algorithm to determine the heart rates from the raw signals. The auto-correlation algorithm's time window depends on the sensor, thus creating different time lag. For Philips® monitors, we have determined that the time lag between the scalp ECG FHR and the Doppler FHR is 1 s, the time lag between the Doppler FHR signal and the MHR measured with the tocometry belt is 5 s, and the time lag between the Doppler FHR signal and the MHR measured with the finger oximeter is 12.5 s.

We had to compensate for these time lags before comparing the heart rates. To estimate the size of the error generated by these time lags, we measured the mean difference in the absolutes value before and after compensation between (a) the signal measured during the MHR FS period with the Doppler FHR sensor and (b) the MHR measured with finger oximeter. For six recordings and a total of 60 min without accelerations/decelerations, the mean difference was 5.0 bpm before correction and 3.5 bpm afterwards. Moreover, the difference was much greater when measured during accelerations or decelerations, which are often critical period for FSs. Thus, this correction is very important.

Unfortunately, we did not record the type of MHR sensor in our database until May 2020. The type of sensor could be easily determined retrospectively because the MHR

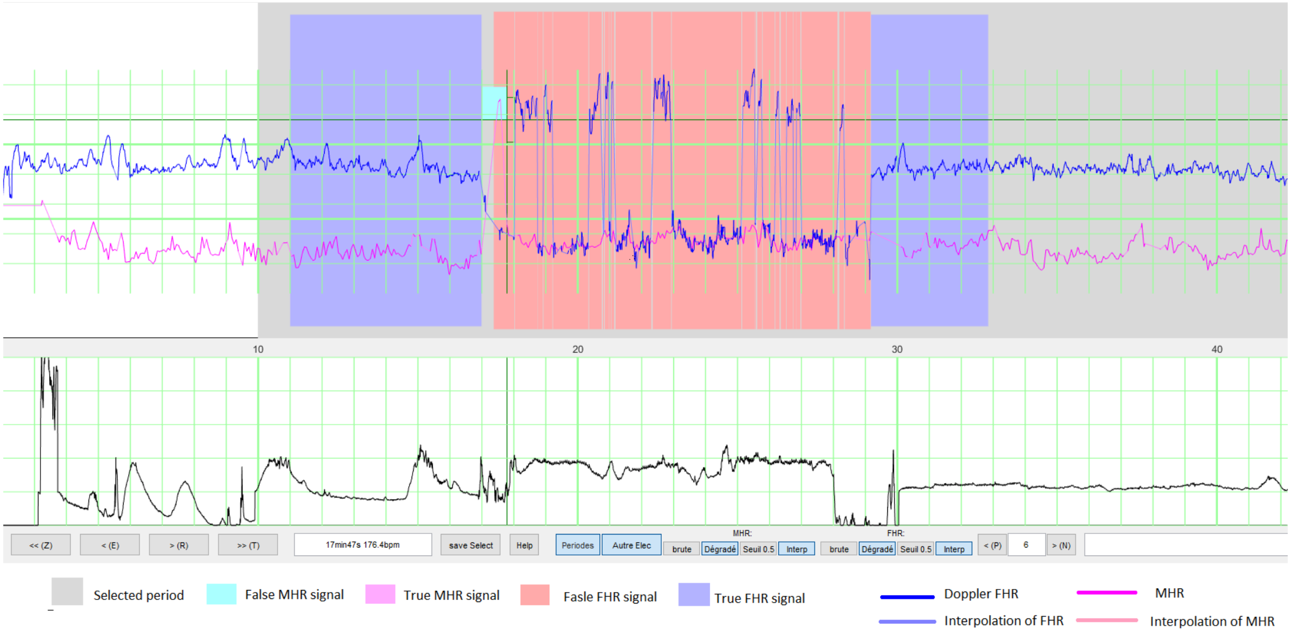


Figure 1. Illustration of the interface for annotating signals during a period with an epidural injection. Both the FHR (in blue) and MHR (in purple) are shown, together with their interpolations (lighter colors) during MS periods for better visualization of isolated samples. The user has selected a window for training the model (in grey) and has annotated two periods with a true FHR signal (in blue) and one period with a false FHR signal (in red). The false signals are either the MHR or the MHR × 2). The user is selecting in a rectangle few a false MHR signal (in cyan).

measured with the tocometry belt had a resolution of 0.25 bpm and that measured with the oximeter had a resolution of 1 bpm. Thus, periods with only whole number corresponded to an oximeter and periods containing numbers with a decimal point corresponded to the ECG sensor combined with the tocomeer.

A second preprocessing step (mostly for the scalp ECG electrode channel) concerned the fact that when a CTG monitor loses the signal, the previous FHR value is sometimes repeated for up to 30s before the absence of a signal is displayed. We detected "hold" periods as periods of more than 12 consecutive samples (each lasting 3 s) with the same value. We estimated that with this threshold, the average number of false detections of "holds" (periods during which the FHR had coincidentally the exact same value over a period 3 s) was <1 for 6 h of recordings, even when the FHR variability was low. These "holds" were replaced by MSs.

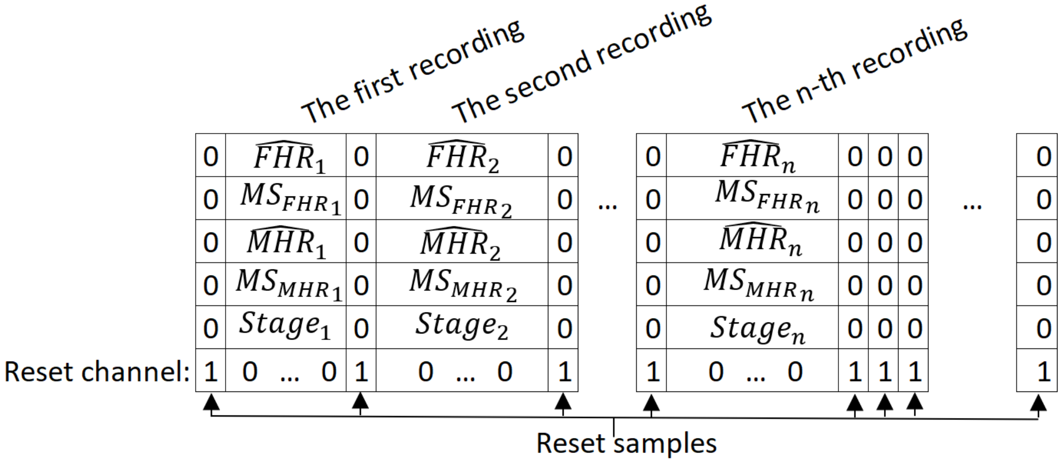
2.5. Preparation of the input matrix

We first normalized each signal (MHR or FHR) in beats per minute (bpm) by doing $\widehat{HR} = \frac{HR-120}{60}$. Next, we coded MSs on an independent channel as MS_{FHR} or MS_{MHR}: no signal was scored as 0 and the presence of a signal was scored as 1. During MSs, the corresponding $\widehat{FHR}$ or $\widehat{MHR}$ was set to 0.

The FSDop model's input comprised the Doppler FHR signal and the MHR signal, whereas the FSHMHR and FSScalp model's inputs comprised only the single, corresponding signal. Since there were no FS of MHR on the scalp ECG channels in our dataset, we did not input the MHR into the FSScalp model.

We added a channel to code the stage of delivery: for each sample, 0 correspond to first stage, or antepartum and 1 to second stage. The second stage of delivery features more MHR accelerations, more FHR decelerations, more MSs, and much more FSs; hence, the delivery stage is important for adjusting the probabilities to the period.

Before deep learning can occur, the data have to be formatted in tables $[batch\ size \times time\ samples \times channels]$. Unfortunately, the recordings differ in their duration. Zero padding is the conventional way of handling differences in duration but our training


The first recording
The second recording
The n-th recording
Reset samples

	0	$\widehat{FHR}_1$	0	$\widehat{FHR}_2$	0		0	$\widehat{FHR}_n$	0	0	0		0
	0	MS_{FHR1}	0	MS_{FHR2}	0	...	0	MS_{FHRn}	0	0	0	...	0
	0	$\widehat{MHR}_1$	0	$\widehat{MHR}_2$	0		0	$\widehat{MHR}_n$	0	0	0		0
	0	MS_{MHR1}	0	MS_{MHR2}	0		0	MS_{MHRn}	0	0	0		0
	0	$Stage_1$	0	$Stage_2$	0		0	$Stage_n$	0	0	0		0
Reset channel:	1	0 ... 0	1	0 ... 0	1		1	0 ... 0	1	1	1		1

Table 1: The input matrix for a unit in the FSDop model. $\widehat{FHR}$ and $\widehat{MHR}$ are the normalized FHR and the normalized MHR, respectively; MS_{FHR} and MS_{MHR} are respectively binary variables indicating whether the heart rate is an MS; Stage is a binary variable indicating whether the sample is in the first stage or second stage of delivery

dataset was mostly composed of short recordings (5 to 30 min) and a few longer signals (up to 4 h). Zero-padding short signals to 4 h would have created a high proportion (85%) of useless calculations of zeros, and cutting the 4 h recordings into shorter parts would have prevented the models from detecting long-term features. We therefore regrouped the short signals into packages of approximately 4 h, and concatenate the signals inside a package to create a “unit”. A mini-batch was then composed of several units of same duration (4 h). However, we had to tell the model to reset the internal status at each change of recording. To this end, we added a reset sample (containing 0) between two recordings and we added a “reset channel”, which was set to 0 most of time and to 1 on reset samples. The layers’ handling of this channel and these samples are described in section 2.8.5 and Appendix B. Thus, Table 1 shows the decomposed matrix obtained for a unit used as the input for FSDop.

For the FSDop dataset, the input matrix sizes were: 80 × 64800 × 6 (training dataset) and 20 × 64800 × 6 (validation dataset). These sizes enabled training on either TPU v3 (batch size: 40) or on Colab Pro® with GPU T4 or P100 (batch size: 20). The batch size was limited, to avoid saturating the memory.

For FSScalp, the input sizes were 40 × 30600 × 4 for the training and validation datasets, since we worked with a batch size of 40 on either TPU v2 or on Colab Pro® with GPU T4 or P100.

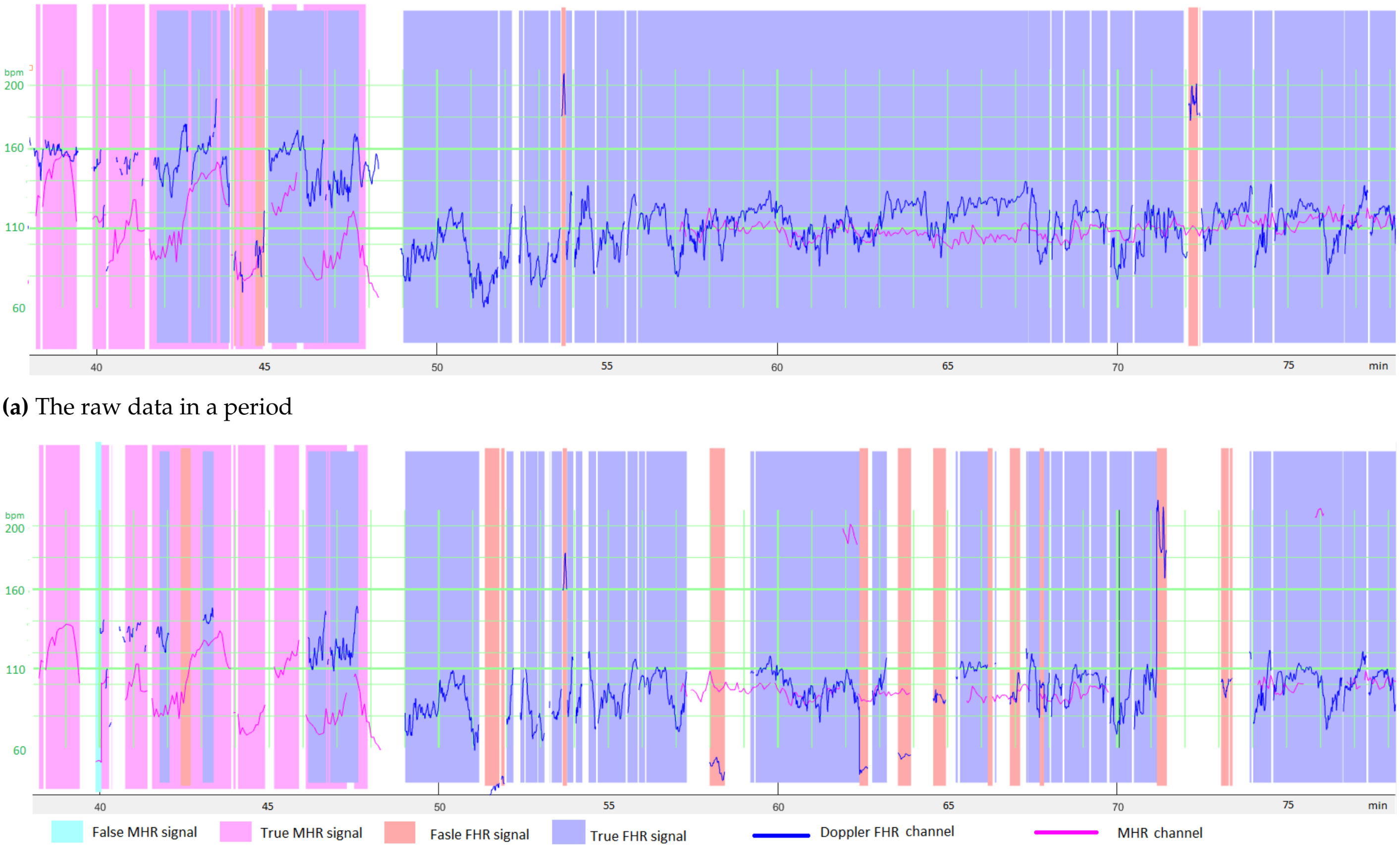
For FSMHR, the input sizes were 22 × 29000 × 4 for the training and validation datasets, since we worked with a batch size of 22 on Colab Pro with GPU T4 or P100.

Our initial trials used a constant window length of 30 min; shorter periods were padded with zeros. Unfortunately, the performance decreased when the method was applied to longer recordings; hence, we have developed this recording concatenation system to train the models on longer periods. We decided to not subsample the recording because we did not know whether the models would be able to extract information from high-frequency signals; for example, a change in variability might suggest a switch between the FHR and the MHR.

2.6. Data augmentation

Even though the number of recordings was high, we reasoned that deep learning could always benefit from more data. We used data augmentation by applying a random transforming the recordings into other realistic recordings:

- Removing the MHR channel completely (P=10%).



(a) The raw data in a period

(b) The period after random transformation

Figure 2. Example of the transformation of a period for data augmentation, with relatively high changes in the Doppler FHR channel (in blue) and the MHR channel (in purple). Expert annotations are shown as colored zones. Data augmentation consisted in adding MS and FS on both the FHR and MHR channels. Moreover, both the MHR and the FHR were multiplied by a $\lambda \approx 0.9$. The final signal corresponds to poor quality recordings but remains realistic. It should be noted that the MHR and FHR were not annotated by the experts for the entire recording (either because they were uncertain, or the period did not contain difficult-to-interpret features

- Adding periods of MS to the MHR or the FHR. We added a random number of MS periods of random duration (from 1 s to 10 min). On average, there was 15% of added FHR MS and 25% of added MHR signal with percentage having a high variance.
- Transforming MHR TSs into MHR FSs by multiplying by 2 or dividing by 2 over a 1 min period, on average. We added 30 s (on average) of MS before and after each period. Overall, 12 periods of MHR FS per hour were added.
- Transforming FHR TS into FHR FS by multiplying or dividing it by 2 or by taking MHR ×2, ×3 or /2 and adding noise $\Delta_{MHR-FHR}$ (for FSDop only). The latter was created by taking periods from other recordings in which the Doppler FHR channel measured the MHR and the maternal sensor also measured the MHR. The noise was defined as the difference between the two signals. We then created 26 min of noise signal and added it randomly to the MHR (so that the FHR was not exactly equal to the MHR multiplied or divided by 2 and thus could be identified easily). MS could be added before or after the generated FS.
- Maintain the MHR channel out of the previously described transformation with a probability of 20% and the same for FHR with probability 15% (so 3% chance of keeping both unchanged).
- Multiplying both the FHR and MHR by a random value, with a Gaussian distribution with an expected value of 1 and a standard deviation of 0.08. This multiplication was applied before normalization, so that the harmonics were still realistic.

- Cutting the recording by adding a reset sample (section 2.5) to the middle of recording (1 reset sample every 10 h, on average).

All the random parameters described above had strongly non-Gaussian distributions and a very high variance. These parameters can be seen with the source code in Python/TensorFlow®. Figure 2 shows a signal after random transformation.

2.7. Cost function

For the FSMHR, FSDop and FSScalp models, the output for each time sample (at 4 Hz) is the probability of being a FS (rather than a TS). This is a binary classification problem and so we used the conventional binary cross-entropy (CE) as cost function. However, we also weighted the samples. The weighted CE is defined in equation 1), where w_i is the weight of sample i, C_i is the label of sample i (1 for FS, 0 for TS) and P_i is the model's estimate of the probability of being a FS. ln corresponds to a natural logarithm.

$$CE = -\frac{\sum_{i,C_i=1} w_i ln(P_i) + \sum_{i,C_i=0} w_i ln(1 - P_i)}{\sum_i w_i} \tag{1}$$

The weights were defined so that:

- Each sample not annotated by an expert (either because he/she was uncertain or the period did not contain difficult-to-interpret features) had a weight of 0. Thus, the cost function is not influenced by the model's output (TS or FS) for these samples.
- Each MS sample had a weight of 0.
- For each specific period, the weight of annotated samples is set to $\sqrt{1/Ratio\ of\ annotated\ samples\ over\ the\ period}$. Thus, if a period is fully annotated, all samples have a weight of 1. If a period of 25 min contains only 1 min of annotation, however, this annotated period is probably more important, and each annotated sample will have a weight of $\sqrt{25}$ (i.e. 5). The total weight of this period is then 5 times lower than that of an entirely annotated period but 5 times greater than that of a selection containing the annotated part only.

We considered that the two types of error (false positives and false negatives) were equally important and thus gave the annotated FS and annotated TS the same weights - even though the classes on the training dataset were highly unbalanced, as shown in section 4.2. If, for example, a few samples are located in a deceleration trough, considering them as FSs would remove the deceleration from the recordings, and so fetal distress might be under-evaluated. In contrast, not removing an MHR period might cause a deceleration to be added to the recording, which would increase the risk of unnecessary intervention.

2.8. The model

2.8.1. Using bidirectional, symmetric gated recurrent units (GRU)

Due to the nature of signal (with a variable length) and the needed for a "synced many-to-many" model [20], we chose to use recurrent neural network (RNN) layers. An RNN layer calculates a state St for a time sample t by using both the input signals I_t at t and the previous state S_{t-1}. Simple RNNs are fully connected ($S_t = f(W_I I_t + W_S S_{t-1})$) where f is an activation function and W_I and W_S are the kernel and recurrent weight matrices, respectively. Simple RNNs are subject to the vanishing gradient problem and have difficulty retaining information for long periods. To solve the vanishing gradient problem, two other RNN architectures have been created: the long short-term memory (LSTM) [21] in 1997 and the GRU [22] in 2014. They add a long-term memory using a gate system. Here, we used a GRU because it requires slightly less weights to train, relative to an LSTM. To create a long-term memory, the GRU uses an update gate corresponding to a set of values of $]0,1[$ for each state; 0 means that the state is updated independently of the previous value, and 1 means that the state is kept as it was at $t-1$. The update gate is a trainable layer. The GRU equations are given in Appendix B.

To optimally analyze a sample at a specific time, it is often necessary to look at what happened before the sample as well as what happened after. We therefore used bidirectional layers (i.e. a GRU applied by moving forward in time and a reverse GRU applied by moving backwards in time). We did not identify direction-specific features and so the FS analysis would be the same in each direction; we therefore constrained the weights to be the same in each direction. Even though TensorFlow® lacks a procedure for this, symmetric RNNs can be effectively produced by concatenating the signal in reverse time order in the batch dimension. The RNN's output for the reversed signals is then re-reversed and concatenated in the channel dimension.

2.8.2. A three-layer GRU

We imagined that the first GRU layer could determine low-level features (such as the mean duration of continuous signal recording, the standard deviation of this duration, and the last FHR value from previous periods), whether the second layer could determine medium-level features (e.g. the expected MHR value and the latter's accuracy), and whether the third and final layer could determine deeper features (e.g. the likelihood of whether the signal was the MHR, the MHR x2, or the FHR x2, etc.). We checked that the GRU was capable of estimating this kind of feature but we did not checked yet whether these features actually emerged in the model.

For accurate estimation, most of the features - even the deepest ones – might need the raw signal on last GRU layer, and so we facilitate their transmission by concatenating them to the previous layer's output. This idea is quite similar to the "shortcuts" used in the famous ResNet network [23] to jump over certain layers.

2.8.3. Sparse kernels

Even though the number of annotated outputs was high ($\approx 3,000,000$ binary values), most were obvious or were highly interdependent. Hence, to avoid overfitting, the number of trainable coefficients must be limited. However, we wanted to keep the numbers of states (e.g. the number of activations of each GRU) high enough to allow the emergence of all the required features. Hence, rather than reducing the number of states, we limited the number of possible interactions between states by setting zeros on recurrent and kernel matrices. For recurrent matrices, we set trainable values on $n \times n$ blocks in the diagonal and set the other matrix elements to 0. For kernel matrices, we also set blocks of trainable values in order to connect some inputs to a small number of outputs (for details, see section 2.8.6). The total number of trainable coefficients was 349 for FSMHR, 7357 for FSDop (in additional to those independently trained with FSMHR), and 3445 for FSScalp.

The same computation could be done by dividing GRU layers into small GRUs and concatenating them thus avoiding several useless multiplications by 0. However, when using a full-size matrix and adding the sparsity constraint, the operations were better parallelized, and the overall process were faster.

2.8.4. Dropout

Since the number of GRU states was low, a strict dropout might erase some primordial features and we preferred Gaussian dropout to add noise but to keep the information. In our few trials, Gaussian dropout performed slightly better than a strict dropout, although the difference was not significant.

2.8.5. Reset constraints

Since several recordings can be concatenated into the same unit, we had to ensure that the state was reset when the GRU switched to another recordings. Unfortunately, TensorFlow does not have a procedure to do this and we could not find a procedure in the literature. By analyzing the GRU equations, we found that by adding a constraint to the kernel matrices and using the reset sample and reset channel (section 2.5), we could force the GRU states to 0 (details in Appendix B).

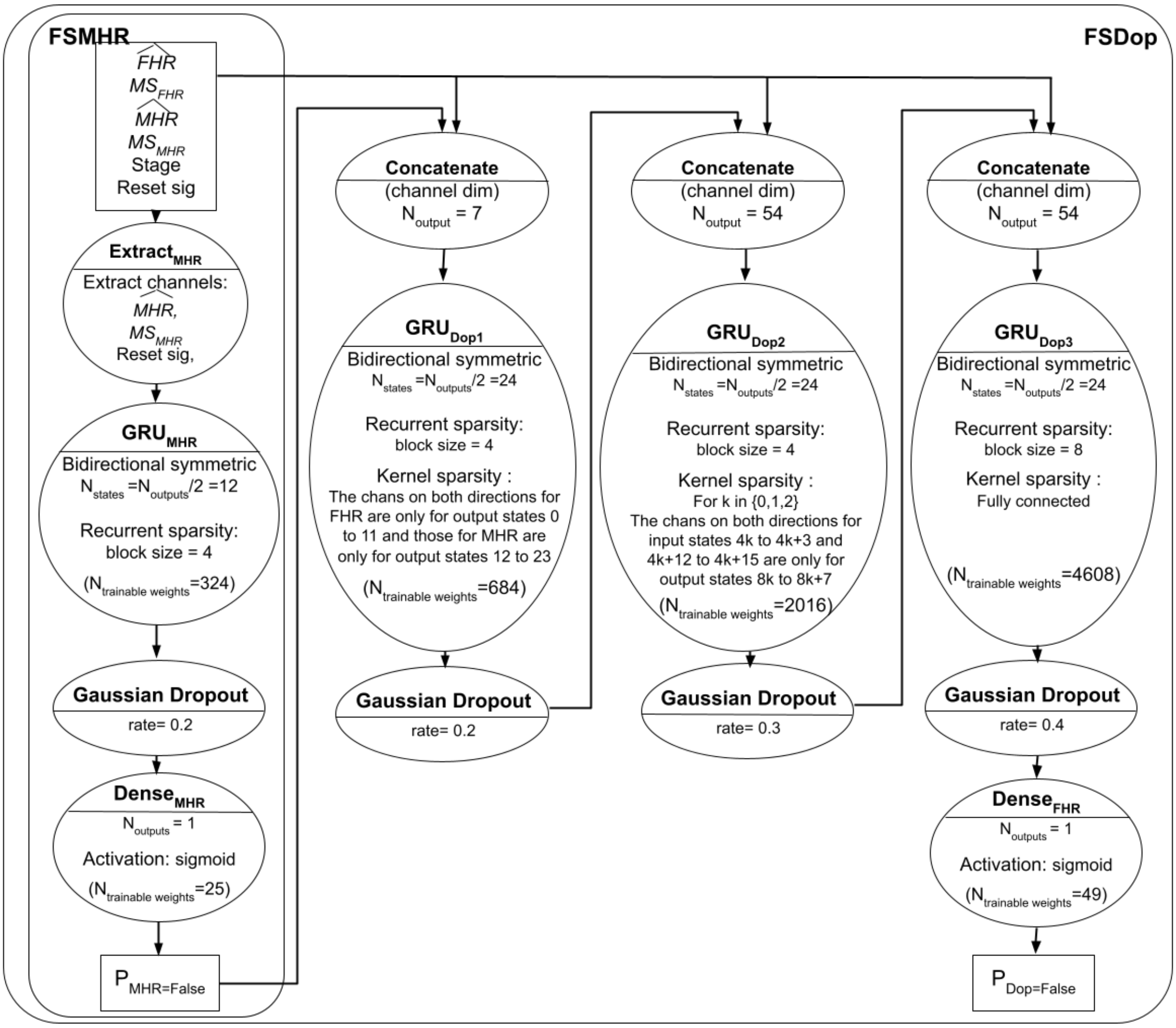


Figure 3. The FSMHR and FSDop models' architectures and hyperparameters.

2.8.6. Overall architecture

The three models' respective architectures are shown in Figures 3 and 4. The FSDop model requires prior computation of the FSMHR model because it is important to know whether the MHR sensor signal is true before comparing to FHR channel and thus we can estimate whether the Doppler FHR signal might be the MHR. We did try training FSDop and FSMHR at the same time but FSMHR overfitted faster than FSDop. Hence, FSMHR was trained first and the weights were fixed during FSDop's training. FSScalp was independent of the other two models.

2.8.7. Training

The models were initially developed in Python® and TensorFlow® 2.4. They were trained using an Adam optimizer and a learning rate that fell progressively from 0.01 to 0.001 after 1000 epochs.

During the developing phase of the project, approximately 15 different models/hyper-parameters were tested for FSMHR, 50 for FSDop and 15 for FSScalp but some of these models contained bugs. For a given architecture and associated hyper-parameters, performance and convergence speed varied greatly from one training session to another. For FSDop (the most complicated model), we selected (according to the validation data) the best of 16 training sessions with the same parameters. For the selected model, the minimum with validation data was reached after 45,000 epochs and then did not improved in the next

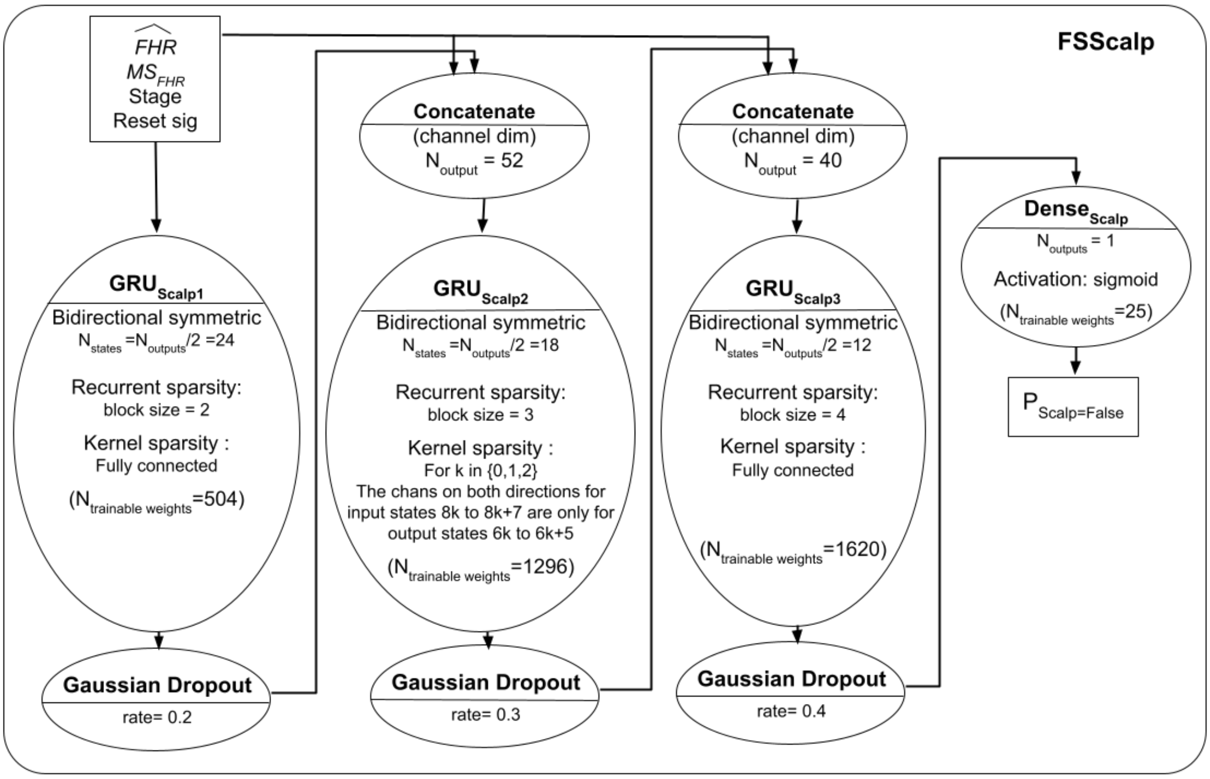

FSScalp
$\widehat{FHR}$
MS_{FHR}
Stage
Reset sig
Concatenate (channel dim) N_{output} = 52
Concatenate (channel dim) N_{output} = 40
GRU_{Scalp1}
Bidirectional symmetric
$N_{states}=N_{outputs}/2=24$
Recurrent sparsity: block size = 2
Kernel sparsity : Fully connected
($N_{trainable\ weights}$=504)
GRU_{Scalp2}
Bidirectional symmetric
$N_{states}=N_{outputs}/2=18$
Recurrent sparsity: block size = 3
Kernel sparsity : For k in {0,1,2} The chans on both directions for input states 8k to 8k+7 are only for output states 6k to 6k+5
($N_{trainable\ weights}$=1296)
GRU_{Scalp3}
Bidirectional symmetric
$N_{states}=N_{outputs}/2=12$
Recurrent sparsity: block size = 4
Kernel sparsity : Fully connected
($N_{trainable\ weights}$=1620)
$Dense_{Scalp}$
$N_{outputs}$ = 1
Activation: sigmoid
($N_{trainable\ weights}$=25)
$P_{Scalp=False}$
Gaussian Dropout rate= 0.2
Gaussian Dropout rate= 0.3
Gaussian Dropout rate= 0.4

Figure 4. The FSScalp model's architectures and hyperparameters.

35,000 epochs. However, some models achieved close results after approximately 3,000 epochs and then stopped improving. The minimum was reached after approximately 5,000 epochs for FSScalp and 10,000 for FSMHR.

For FSDop, the computation time was 40 s per epoch on a computer with an GPU Nvidia® RTX 2080 Ti. Fortunately, we were able to access the Google® TPU Research Cloud program. The computation time with TPU v3 was then 6 s per epoch. The long computation time was due to poor parallelization on the GRU because the parallelizable dimensions were small (number of states: 55; batch size: 40 × 2 for symmetry) and non-parallelizable dimensions were large (sample number: 64800 × 4 GRU layers × 2 batches per epoch).

For testing, we recoded the model in MATLAB® and included it in the FHRMA toolbox [19]. We chose not to use the MATLAB® Deep Learning toolbox to reduce the user requirements. On CPU (Intel® i7-11800H), the computation time for FSDop for 1 h of recording was 0.9 s. When several recordings were computed in multiple threads, the computation time for 1 h of recording fell to 0.2 s.

3. Evaluation method

Performance was evaluated on data unseen by any model. Test datasets were created after the final model was established, so that we were not tempted to improve the models once we had seen the results with the test data (i.e. risk of over-evaluation). Two evaluation systems and datasets were developed:

(i) **Test dataset for routine clinical practice**: An evaluation of the three models on 30 recordings per model, selected at random and fully annotated by the experts. This evaluation was intended to provide an idea of the model's performance in data of routine clinical practice not biased by selection criteria.

(ii) **A dual signal test dataset**: FSDop was evaluated on 34 recordings with the simultaneous scalp ECG sensor and Doppler sensor. The scalp ECG was used by two experts to set the ground truth. Two other experts analyzed the same periods but

doi:10.20944/preprints202207.0131.v1

were blinded to the scalp ECG signal. This dataset was used to compare FSDop's performance with that of the experts.

For each dataset and for each model, we measured the accuracy (setting a threshold of P=0.5 for each classifier), the contingency table, the area under the receiver operating characteristic curve (AUC), and the CE. These metrics were also calculated for the training and validation datasets and for each stage (antepartum, first stage of delivery, and second stage of delivery). Given that a Doppler recording is not always accompanied by a simultaneous MHR recording in routine clinical practice (due to obsolete CTG monitor, ill-trained staff or difficulties to position the sensor), we also assessed the performance without considering the MHR channel.

The following subsection provides details of how the two test datasets were built.

3.1. The test dataset for routine clinical practice

In the training and validation datasets, the proportion of FSs is higher than the routine clinical practice ; since the periods were selected because they contained FSs or at least ambiguities. Thus, this test dataset was intended to assess the model's performance in routine clinical practice and was not biased by selection criteria.

Thirty Doppler sensor recordings (during the first stage of delivery and, in some cases, the second stage) were selected at random from among those recorded in 2019 This dataset was used to evaluate FSMHR and FSDop.

Thirty scalp electrode recordings (during the first stage of delivery and, in some cases, the second stage) were selected at random from among those recorded between January 1st, 2019, and May 31st, 2021. This dataset was used to evaluate FSScalp.

The 60 recordings were analyzed by three experts (two obstetricians and a midwife, all of whom were involved in research and giving lectures on FHR analysis). Each expert analyzed a third of the recordings. On the Doppler/MHR dataset, both the MHR and the FHR were fully annotated. Each sample of an FHR or an MHR from the start of the recording through to delivery was annotated as a TS, an FS, or an uncertain signal. On the scalp electrode dataset, the scalp ECG FHR channel was fully annotated as a TS, an FS, or an uncertain signal. The scalp FHR recording was generally shorter than the entire recording because the scalp electrode was applied in second-line, after the Doppler sensor.

3.2. The dual signal test dataset

The evaluation on test dataset for routine clinical practice is limited by potential errors made by the expert; one cannot say whether the model is better or worse than the expert. Moreover, not all the recordings contained periods that the model might fail to analyze.

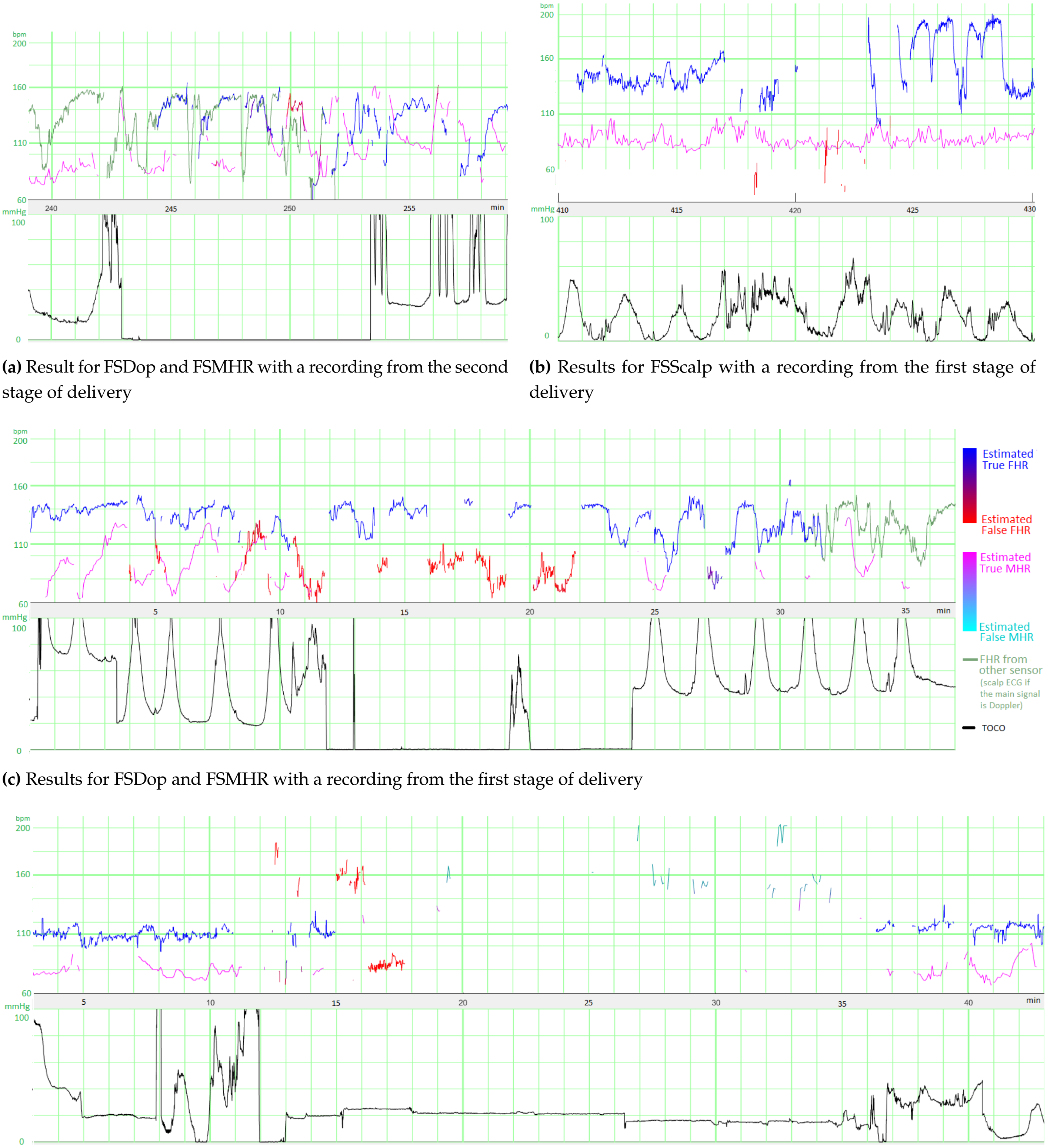
We therefore built a second test dataset by selecting all periods of a least 10 min between January 1st, 2019, and May 31st, 2021 (this inclusion period is after the one of the training dataset A), on which both Doppler and scalp sensor signals were simultaneously recorded. This yielded 45 periods in 37 recordings.

Next, to form the solid ground truth, two experts (a senior obstetrician and an engineer) consensually annotated the Doppler channel with the help of the scalp ECG signal. These experts annotated all samples in a period as a TS, FS or uncertain signal, even when scalp ECG is temporary MS. Next, two other experts (an obstetrician and a midwife) analyzed these periods but were blinded to the scalp ECG signals. These experts had to annotate each sample in the period as a TS or an FS; annotating a signal as "uncertain" was not allowed. It was then possible to compare the method's performance with that of the two experts.

4. Result and discussion

4.1. Illustrative results

Figure 5a shows an illustrative result for FSDop (and for FSMHR, even though there are no MHR FSs) and a period at the start of the second stage of delivery. Although the MHR was measured discontinuously, we can see that MHR has accelerations synchronized



(a) Result for FSDop and FSMHR with a recording from the second stage of delivery

(b) Results for FSScalp with a recording from the first stage of delivery

(c) Results for FSDop and FSMHR with a recording from the first stage of delivery

(d) Result of FSDop and FSMHR with a recording from the first stage of delivery, during the administration of an epidural

Figure 5. Examples of results for the three models (FSMHR, FSDop and FSScalp) with recordings from the first and second stages of delivery. The likelihood of an FS estimated by each model is represented as a color gradient. On examples (a), (c) and (d), the blue/red signal corresponds to the Doppler channel and the green signal corresponds to the scalp ECG channel. On example (b), the blue/red signal is the scalp ECG signal. On all recordings, the time lag in the MHR is corrected as described in section 2.4.

with contractions (if recorded). In contrast, the FHR decelerates and is difficult to analyze because it crosses the MHR. The start of the recording has a scalp ECG signal (in green), which shows the true FHR; the model did not see this scalp ECG signal. The probability of an FS estimated by the model is shown on a color scale. One can see that the method identified the period of FSs at minute 250 because it matches the MHR exactly. During a period in which the Doppler signal is superposed on the scalp ECG signal, the model predicted correctly that the signals were TSs. Another short period of FS appears to have been identified correctly at minute 256. Looking very closely, one can see that FSs at minutes 246 and 249 were not identified by the model; although this constitutes a minor error. This example would be very difficult to interpret in the absence of an MHR signal.

The figure 5b shows an illustrative result for FSScalp and the first stage of delivery. The few probable FSs appears to have been detected correctly.

The figure 5c shows another illustrative result for FSDop (and FSMHR, even though there are no MHR FSs) and a period corresponding to a first stage of delivery. The MHR showed accelerations and the FHR showed decelerations, as confirmed by the scalp electrode. The MHR and FHR values were sometimes identical. There was a long period of MS on the MHR channel but FSDop correctly identified the corresponding FHR FS - even during periods with MHR MS. The model might have failed to identify a possible short FSs at minute 27.

The figure 5d shows a second illustrative result for FSDop and FSMHR and a period corresponding during which an epidural was given. There is a long period of MS on the Doppler channel. During this time, the MHR channel has FSs – most of which were detected. One can see that these MHR FSs do not prevent FSs on the Doppler FHR channel from being detected correctly.

4.2. Statistical results

All the results for the training and test datasets are summarized in Table 2. The left and right parts of the table correspond to the datasets and models, respectively. The most intuitive metric is accuracy, which was usually greater than 99%. However, since most of the signal samples are TSs, trivial classification of all samples as TSs would also produce a relatively high accuracy. This trivial model would have an accuracy corresponding to the "Percentage of "true" among annotated" column. An accuracy that is lower than the percentage of TSs means that the model removes more TS than FSs. Hence, if we assume arbitrarily that a false negative has the same importance as a false positive (as in section 2.7), the model would be of no use. This does not mean, however, that the method is worse than chance – as shown by other metrics.

The contingency table contains the sensitivity (Se), specificity (Sp), positive predictive value (PPV) and negative predictive value (NPV). $Se + PPV > 1$ is equivalent to $Acc > Percentage\ of\ TS$, so this condition should be met for useful model. However, the strongly imbalanced data meant that this is not a trivial problem, and so some models did not meet this condition. To perform better than chance, $Se + Sp$ should be greater than 1; this was the case for all models.

The AUC is a guide to the classifier's performance, independently of the threshold for the output probability. If the AUC >0.5, we know that the model performs better than chance.

The CE is the most precise measure of performance (since it measures both accuracy and the model's ability to recognize uncertainty) but is less intuitive for humans. A trivial random model in which all samples have a probability of $P = Percentage\ of\ ''true''$ has $CE = -P * ln(P) - (1 - P)ln(1 - P)$. CE is the optimized criterion during the training (section 2.7) but we did not weight the samples for the evaluation.

4.3. Results with the test dataset for routine clinical practice

The accuracy values for this dataset showed that all three models are highly effective (FSDop: 99.66%, FSMHR: 99.92% and FSScalp: 99.93%). The percentage of TSs on the

doi:10.20944/preprints202207.0131.v1

				Statistics on the dataset						Statistics on model results						
Reference labels set by:	Dataset	Stage	Model or expert	Number of recordings	Number of analysed window	Avergage window Length (min)	Percentage of missing signal	Percentage of annotated	Percentage of "true signal" among annotated (=trivial model accuracy)	Accuracy	Sensibility (=rate of FS detected)	Specificity	Positive Predictive Value (=rate of true FS)	Negative Predictive Value	AUC	Cross entropy
Experts: 0 (Engineer) & 3 (senior obstetrician)	Train	All	**FSDop**	341	546	39	19.2%	64.3%	82.9%	**99.37%**	97.7%	99.7%	98.6%	99.5%	0.9996	0.0169
			FSMHR	74	98	32	39.6%	57.4%	94.4%	**99.10%**	87.6%	99.8%	96.1%	99.3%	0.9957	0.0295
			FSScalp	109	188	27	16.0%	97.7%	99.5%	**99.83%**	78.6%	99.93%	85.2%	99.90%	0.9948	0.0056
	Validation	All	**FSDop**	82	129	41	16.3%	68.8%	89.5%	**99.30%**	97.0%	99.6%	96.3%	99.7%	0.9992	0.0213
			FSDop WO MHR	82	129	41	16.3%	68.8%	89.5%	**97.75%**	85.9%	99.1%	92.0%	98.4%	0.9928	0.0643
			FSMHR	22	30	37	39.4%	39.4%	89.3%	**98.68%**	90.9%	99.6%	96.6%	98.9%	0.9952	0.0395
			FSScalp	25	39	31	19.2%	97.2%	99.4%	**99.90%**	85.3%	99.99%	98.3%	99.91%	0.9949	0.0056
		Antepartum	**FSDop**	7	8	34	18.9%	96.6%	88.2%	**99.37%**	99.2%	99.4%	95.6%	99.89%	0.9997	0.0205
			FSDop WO MHR	7	8	34	18.9%	96.6%	88.2%	**99.18%**	97.8%	99.4%	95.3%	99.7%	0.9991	0.0316
		1st stage	**FSDop**	71	114	39	14.9%	67.4%	89.4%	**99.49%**	97.6%	99.7%	97.6%	99.7%	0.9995	0.0158
			FSDop WO MHR	71	114	39	14.9%	67.4%	89.4%	**98.22%**	86.2%	99.6%	96.7%	98.4%	0.9971	0.0470
			FSMHR	17	23	37	37.3%	39.5%	89.8%	**98.64%**	90.1%	99.6%	96.4%	98.9%	0.9973	0.0368
			FSScalp	22	34	27	20.3%	97.2%	99.3%	**99.89%**	85.6%	99.99%	98.4%	99.90%	0.9953	0.0060
		2nd stage	**FSDop**	26	26	20	26.5%	66.8%	92.2%	**97.30%**	86.3%	98.2%	80.4%	98.8%	0.9860	0.0766
			FSDop WO MHR	26	26	20	26.5%	66.8%	92.2%	**91.90%**	67.3%	94.0%	48.5%	97.1%	0.8663	0.2619
			FSMHR	12	12	22	46.1%	39.0%	87.4%	**98.83%**	93.4%	99.6%	97.2%	99.1%	0.9918	0.0501
			FSScalp	15	17	16	15.6%	97.1%	99.7%	**99.94%**	83.7%	99.99%	97.5%	99.94%	0.9962	0.0043
Experts: 1 (obstetrician), 2 (midwife) & 3 (senior obstetrician) ; Each a third	Test dataset for routine clinical practice: Doppler and MHR	All	**FSDop**	30	30	331	7.8%	98.9%	97.2%	**99.66%**	91.3%	99.91%	96.7%	99.7%	0.9992	0.0112
			FSDop WO MHR	30	30	331	7.8%	98.9%	97.2%	**98.88%**	69.9%	99.7%	88.1%	99.1%	0.9951	0.0322
			FSMHR	30	30	331	15.4%	99.75%	99.86%	**99.92%**	63.2%	99.98%	79.6%	99.95%	0.9706	0.0044
		1st stage	**FSDop**	30	30	312	7.1%	99.5%	97.2%	**99.75%**	93.1%	99.94%	97.7%	99.8%	0.9996	0.0086
			FSDop WO MHR	30	30	312	7.1%	99.5%	97.2%	**99.08%**	72.3%	99.84%	92.7%	99.2%	0.9979	0.0235
			FSMHR	30	30	312	13.9%	99.86%	99.91%	**99.95%**	75.8%	99.98%	75.8%	99.98%	0.9893	0.0032
		2nd stage	**FSDop**	20	20	29	18.2%	87.2%	95.7%	**97.88%**	66.3%	99.3%	80.8%	98.5%	0.9780	0.0667
			FSDop WO MHR	20	20	29	18.2%	87.2%	95.7%	**94.72%**	37.0%	97.3%	38.0%	97.2%	0.8322	0.2140
			FSMHR	20	20	29	38.9%	97.3%	98.7%	**99.21%**	42.3%	99.96%	93.5%	99.2%	0.9589	0.0317
	Test dataset for routine clinical practice: Scalp ECG	All	**FSScalp**	30	30	177	3.2%	99.85%	99.90%	**99.93%**	49.1%	99.98%	75.2%	99.95%	0.9792	0.0032
		1st stage	**FSScalp**	30	30	165	2.8%	99.89%	99.90%	**99.94%**	47.3%	99.99%	78.6%	99.95%	0.9798	0.0030
		2nd stage	**FSScalp**	17	17	21	8.9%	99.3%	99.82%	**99.86%**	63.4%	99.92%	60.0%	99.93%	0.9877	0.0054
Experts: 0 & 3 (double reading) with help of scalp ECG	Dual signals test dataset	All	**Dop by expert 1**	37	45	20	22.0%	98.0%	96.6%	**96.91%**	15.6%	99.8%	74.5%	97.1%	-	0.3352
			Dop by expert 2	37	45	20	22.0%	98.0%	96.6%	**98.24%**	60.7%	99.6%	83.7%	98.6%	-	0.2020
			FSDop	37	45	20	22.0%	98.0%	96.6%	**97.29%**	53.3%	98.9%	62.5%	98.3%	0.9645	0.0902

Table 2: Statistical results for the study's three models (FSDop, FSMHR an FSScalp) and datasets.

corresponding dataset were respectively 97.2%, 99.86% and 99.90%. The fact that the accuracy was greater than the percentage of TSs means that the models reject more FSs than TSs. The good performance is also confirmed by the AUC of respectively 0.9992, 0.971 and 0.9792 as well as the CE 0.0112 (vs 0.1286 for the trivial model), 0.0044 (vs 0.0108) and 0.0032 (vs 0.0074).

FSDop was the most useful model for routine clinical practice because the latter recordings contain many more FSs; the AUC of 0.9992 is impressive. The FSMHR and FSScalp models were less useful for this dataset, given to the very small number of FSs. However, the AUCs of FSMHR and FSScalp were re respectively 0.971 and 0.978, which correspond to good, significantly better-than-chance performance. Even though FSMHR and FSScalp have little effect (since there are very few FSs on those channels), it would make sense to implement them on the central monitor or as preprocessing steps.

FSDop's performance fell to 97.88% for the second stage of delivery (Table 2). This was expected because the second stage is often more complicated, with more MSs, more FSs, more FHR decelerations (which can be mistaken for the MHR) , and more MHR accelerations (which can easily be confused with the FHR). The percentage of FSs in the second stage was 4.3%, although 12.8% of the second-stage samples were annotated as "uncertain" by the experts. This confirms that the second stage is more complicated to analyze, and the true proportion of FSs was probably around 10% - much more than the 2.8% in the first stage of delivery.

When we removed the MHR from the model's input, the performance fell from 99.75% to 99.08% for the first stage of delivery, and from 97.88% to 94.72% for the second stage.

Thus, the model is still efficient for the first stage of delivery. However, for the second stage (and even though the AUC was 0.83), the model removed more TSs than FSs and so might not be relevant. For example, if these models were applied to the CTG-UHB public dataset [24], the absence of MHR data will limit their value. The difficulty of interpreting the second stage of delivery in the absence of an MHR signal was confirmed by the experts during their annotation of the study's datasets; ; hence, this was not a method-specific problem. We encourage practitioners to check that the MHR is recorded well during the second stage of delivery, since poor obstetric decisions prompted by FHR/MHR confusion is probably more common than thought. We also suggest that the MHR could be recorded with a smartwatch, which might be more comfortable for the mother and possibly more reliable and we hope CTG monitor manufacturers will study this possibility.

In this dataset, the experts were able to annotate features as "uncertain", and we so did not expect any incorrect (false) annotations. However, some of the models' remaining errors might still be expert errors. The second test dataset was designed to overcome this problem but could be applied to FSDop only.

4.4. Results on double signals test dataset

The dual signal test dataset was more difficult to interpret than the routine clinical practice dataset. Nevertheless, the FSDop model's accuracy was 97.29%, which was still higher than the proportion of true samples in the dataset (96.6%). The AUC was 0.965, which corresponds to good classification performance. The accuracy rates for the two experts (who analyzed the recordings under the same conditions as the models, i.e. blinded to the scalp ECG signal) were 96.91% and 97.29%; hence, the model was better than one expert and worse than the other. The part of randomness in those statistics is difficult to assess although the statistics confirmed our impression that the method was as accurate as a competent practitioner but a well-trained expert who understands the FS mechanism described in Appendix A could probably do slightly better than the models. Thus, we encourage other researchers to try to improve the models' performance levels and this a reason why we shared resources of this study; a few ideas are given in section 4.6.

4.5. Results on validation data

As expected, performance on the validation dataset was slightly worse than the performance on the training data. One cannot compare the levels of performance between validation dataset and test datasets because they do not have the same selection criteria. However, the validation dataset provides greater precision (but not accuracy), since there are more FS data. However, there should be a probable bias in evaluation due to the settings of hyper-parameters in this same dataset. The models' accuracies were very satisfactory (99.30% for FSDop, 98.68% for FSMHR, and 99.90% for FSScalp) and much higher than the percentage of TSs (89.5%, 89.3%, and 99.4%, respectively); hence, the number of errors was divided by a factor of 15.0 for FSDop, 8.1 for FSMHR, and 6.0 for FSScalp. The FS rate was much lower in the FSScalp dataset (0.6%) than in the FSDop dataset (10.5%), and so FSDop appeared to have more interest. The FS rate for the MHR was high (10.7%) but this was mainly due to selection bias during annotation.

We judged unnecessary to realize a test dataset for antepartum recordings since globally, they do not exhibit difficulties not already presents in first stage of delivery. This is confirmed by the accuracy on the validation dataset (99.37 %) which is approximately equivalent to the accuracy on first stage of delivery (99.49 %), for a percentage of TS equivalent (88.2 % vs 89.4%). The only details which did not work with first models were in a recording with FHR baseline above 210 bpm. The problem was solved with the data augmentation of random multiplication described in sec. 2.6.

Poor performance in the second stage (with a censored MHR channel) was apparent in the validation data. Indeed, the CE was 0.2619, which was not significantly better than the CE of a trivial random model (0.273). We tried to train another model independently for this particular task (the second stage of delivery without an MHR sensor) and achieved a CE of around 0.23; however, we considered that this performance was too poor to justify adding another model. The performance improves drastically when an MHR sensor is present, and it would be better to change current medical practice and ensure that the MHR is always recorded.

4.6. Perspectives

The models' levels of performance were satisfactory but could probably be improved:

- Our models do not use the information from the tocography signal. There are some complicated cases in which the synchronization with contraction can help to determine whether the FHR is a deceleration or a FS. Adding this information to the model would require the automatic detection of the start and end of the UCs unlikely to be an easy task and would probably not emerge accurately from optimization of the FS detection task in deep learning.
- Access to the raw Doppler signal from which the auto-correlation is computed (for estimation of the FHR) would have provided additional information for recognition of the FS. However, obtaining the raw Doppler signal is not possible with most commercial devices.
- Some of our preprocessing steps (FHR/MHR delay compensation, and removing "holds") were specific to Philips® monitors and should be adapted for other manufacturers' monitors. Once this delay compensated, we expect our models to work on other monitors but we have not yet evaluated this aspect.
- We have not yet fully evaluated whether FS detection can improve the performance of automatic FHR analysis for acidosis detection. Our initial results suggest that some of the computed features were improved by this preprocessing, but not all (e.g. deceleration surface), and this is difficult to explain. Our future research will focus on this topic.
- Once trained, the models have a very low computation time (<1s per hour of recording). However, the training is lengthy, and the differences in performance between two training sessions obliged us to train the models several times. To improve the models, we recommend to reducing the training time by (for example) mixing Convolutive and recurrent neural networks or using transformers. It would also be better to create a model with less variability between training sessions.
- Data augmentation enabled us to greatly improved the models. Our efforts to generate realistic FSs from TSs could probably be continued. It may also be possible to synthesize realistic Doppler FHR signals from the scalp ECG signal but we have not tried this yet.
- It might be possible to increase the number of recordings with a solid ground truth (i.e. better than the analysis by experts, who would only use the same information as the models do) by putting two Doppler sensors instead of one on the tocography belt. Although this would be less precise than a scalp ECG, adding a scalp ECG sensor with no medical justification is ethically problematic. Thousands of recordings could be recorded and annotated automatically by considering that two similar signals are necessarily either both TSs or (less likely) both FSs. If two signals differ, one is likely to be false. Originally, we tried this method for annotating training dataset A (with dual Doppler/Scalp ECG sensors). Although this method worked, only 90 recordings were concerned and it was simpler and more informative to annotate by experts using Scalp ECG as indicator.

5. Conclusion

We developed intelligent models for detecting FSs in FHR and MHR recordings. The detection of FSs is particularly useful for Doppler recordings, in which FSs are frequent particularly due to the high probability of capturing MHR instead of FHR. The models performed at expert level, although a well-trained expert could probably do better in some cases. We hope that these models will be able to improve clinical care by providing alarm to practitioner of possible FSs in the FHR recording. Moreover, the models can be used to preprocess FHR recordings for a subsequent automatic analysis. We showed that FS detection is a complicated problem and is often neglected in the literature; however, tackling this problem might have a major impact on both clinical care and automatic analysis.

Our results suggest that the MHR sensor is very important (and often essential) for recognizing the FS - particularly during the second stage of delivery. The MHR is also often missing during the second stage of delivery. One simple, comfortable solution would be to add a smartwatch with a wireless connection. We also strongly encourage CTG monitor manufacturers to compensate for the delay between the FHR and the MHR (5 s or 12.5 s on a Philips system); this would make it easier to distinguish the MHR on the Doppler FHR channel and (as display software) could be patched relatively easily into centrals of monitoring.

The proposed approach used a GRU and required several developments with a great degree of optimization (e.g. data augmentation, kernel/recurrent sparsity, and a symmetric GRU). We also developed a technique for concatenating recordings inside a GRU unit (Appendix B); this avoids zero padding and thus useless operation, and might be useful for other problems. The present study also generated a large, annotated dataset.

To encourage other researchers to work on this problem, we have made all the data and our source code available via the open-access FHRMA project [19]. Annotation on test datasets are not provided, and so researchers who want to evaluate their models will have to send us their results for evaluation; hence, a competition has been opened. We also encourage researchers working on FHR signal processing to use our FS detection methods as preprocessing steps and to use our open-source WMFB method (also in the FHRMA toolbox) [13] for FHR baseline estimation.

Our future research will evaluate the impact of FS detection on computation of FHR features for the detection of fetal acidosis.

Author Contributions: Conceptualization, S.B., A.H., R.D. and D.H.; methodology, S.B.; software, S.B.; validation, S.B., A.H., R.D. and D.H.; formal analysis, S.B., L.P.; investigation, S.B, A.H., R.D., D.H.; resources, A.H., R.D., D.H.; data curation, S.B., A.H., R.D., D.H.; writing—original draft preparation, S.B.; writing—review and editing, all authors; visualization, all authors; supervision, S.B., D.H.; project administration, S.B., A.H., D.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research did not receive any external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and was approved by the institutional review board at the Lille Catholic Hospitals (GHICL, IRB 00013355) on August 4th, 2016 (reference: 2016-06-08).

Informed Consent Statement: All study participants provided their informed consent.

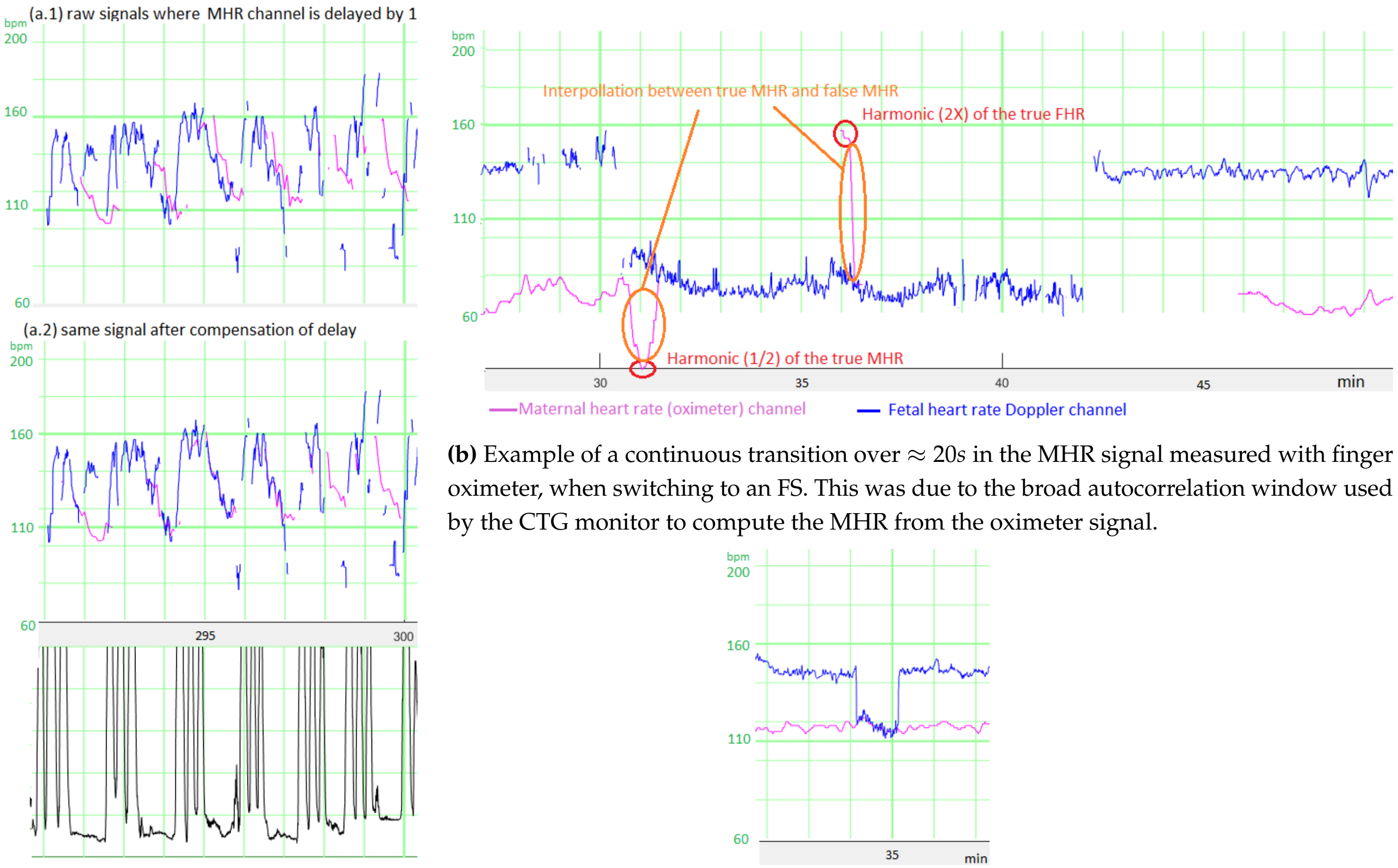
Data Availability Statement: All data and source code are available in the Fetal Heart Rate Morphological Analysis Project at https://github.com/utsb-fmm/FHRMA.

Acknowledgments: We thank Google TPU Research Cloud (TRC) program (https://sites.research.google/trc) for providing free access to cloud TPUs to support and accelerate our study.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. How to know if a signal is true or not on Doppler channel

Saed et al. [14] presented details of how to differentiate between the FHR and the MHR. We have added further details for researchers who would like to understand how an artificial intelligence "thinks" and for clinicians who might not be aware of certain technical details. This appendix highlights the difficulty of the problem and why the use of logic rules is complicated.



(b) Example of a continuous transition over ≈ 20*s* in the MHR signal measured with finger oximeter, when switching to an FS. This was due to the broad autocorrelation window used by the CTG monitor to compute the MHR from the oximeter signal.

(a) Example of an FHR FS in a second stage recording, with and without correction for the lag in the MHR channel.

(c) Example of a continuous transition over ≈ 1*s* in the FHR signal

Figure A1. Illustration of the effects produced by CTG monitors, complicating the recognition of FSs. These effects are due to the autocorrelation algorithm used to estimate heart rate.

- **Normal frequencies:** the FHR normal baseline is between 110 and 160 bpm, whereas the MHR during delivery is around 80 bpm. However, the possible occurrence of maternal tachycardia, MHR acceleration, FHR acceleration or FHR deceleration means that these signals can have the same value at times, and the FHR can fall below the MHR. More details are given in [14].
- **FS/TS switch:** A continuous period is most likely a single class (TS or FS. In most cases, there is MS between a change of class. If not, there is a large difference (> ≈25 bpm) between two consecutive signals or the transition may not exceed 1s (four samples) for a Philips monitor with a Doppler sensor (e.g. Fig. A1c). For the MHR sensor (on Philips monitors), the monitors can interpolate the transition between TS and FS over longer periods (≈ 20s, as in Fig. A1b). There are also some very rare cases in which the FHR crosses the MHR and the FHR sensor switches without a discontinuity.
- **FHR/MHR superposition:** when the MHR is recorded with its own sensor, the MHR on the Doppler FHR channel is approximately superposed. Nonetheless, on the Philips device, the MHR sensors (both the sensor on the tocometry belt and the finger oximeter) provide a smoother signal than the Doppler FHR sensor. Thus, a variation of less than 2s approximately coincides, although faster variations do not. On the Philips monitor (and probably other brands), the lag between the FHR channel and the MHR channel is 12.5 s (for the finger oximeter) or 5 s (for the tocometry belt), as shown in Fig. A1a. This occurs on both paper recordings and on the central server.

These lags make it harder to see the coincidence between the FHR channel and the MHR channel.

- **Possible FS values:** The Doppler FHR sensor have several possible FS values. It can be either the MHR (e.g. Figs. 5a,5c and 5d), a harmonic of the FHR (double, half or (more rarely) triple or a third) or a harmonic of the MHR (e.g. 5d). The MHR signal is less subject to FS but it can switch to its own harmonic (double, half, or (more rarely) triple) (e.g. 5d). Over short periods, it is also possible for these sensors to take on other values, although the variance is high and the situation does not last for more than a few seconds. The FSs on the scalp ECG sensor are also short and have high variance (often a vertical line) (e.g. fig. 5.b). In our dataset, we did not observe MHR value nor harmonics on the scalp ECG sensor. However, cases of this have been reported in the literature [4].
- **Possible MHR values during MSs on the MHR sensor:** During the first stage of delivery and during antepartum recordings, the MHR rarely changes. Hence, even though there are MSs on the MHR sensor, the MHR's range can thus be estimated. There are a very few MHR decelerations, and values below the MHR baseline are probably the true FHR or, more rarely, a low harmonic FS. The MHR can accelerate during uterine contractions (UCs), particularly during the second stage of delivery. MHR accelerations are most often higher (>≈30bpm) and longer (>1 min) than FHR accelerations (ex. figs 5a, 5c). The MHR accelerations are normally synchronized with UCs, although the latter are not always well measured. In contrast, the FHR tends to have decelerations during or just after UCs, and thus the FHR and the MHR might overlap. The characteristic MHR acceleration has a round peak (Fig. 5c). During the second stage, three bumps (of variable amplitude) can be often observed on the peak; these correspond to the three expulsive pushes during UCs (e.g. Fig. 5a at min 250 and Fig. A1a).
- **Sudden FHR change:** MHR variability (the max-min difference over 1 minute, outside decelerations/accelerations) often differs from FHR variability. Hence, a change in variability might indicate that the FHR has switched to the MHR. Likewise, an FHR that shows several decelerations and that suddenly stop might also indicate a switch to the MHR.
- **A hyperkinesia may confirm FHR TS:** A prolonged deceleration is often associated with hyperkinesia (increase of rhythm and amplitude of CU); if so, a prolonged FHR deceleration might be more likely than an MHR switch.
- **Epidural may confirm FHR FS:** : a long period of MHR captured on the Doppler FHR channel ($\approx$ 20 min) is common when an epidural is being given (Figs. 1, 5d, A1b).

Appendix B. Multiple recordings on the same unit in the GRU

In our problem, there are many short recordings and few very long recordings. Using the traditional zero padding approach could produce a lot of useless calculations. We concatenated the recordings in the same unit and add a constraint to the GRU matrices, to ensure that the GRU state was reset when switching between recordings. This could also be done with standard RNN and LSTM approaches. To this end, we added reset samples between recordings in the input matrix, together with a reset channel (Table 1).

The GRU equations are shown in equation A1, where x_t is the input vector (size n_x), h_t is the output vector (also corresponding to the GRU state, size n_h); $\hat{h}_t$ s the candidate activation vector (size n_h), z_t is the update gate vector (size n_h), r_t is the reset gate vector (size n_h). W_* are the kernel matrices, U_* are the recurrent matrices, and b_* are the bias. The operator $\odot$ denotes the Hadamard product. σ_g is a sigmoid function and ϕ_h is the hyperbolic tangent. The reset channel $x_t^{n_x}$ is 1 when t is a reset sample and 0 on other samples. On reset samples, $x_t^i = 0, \forall i < n_x$.

$$\begin{aligned} z_t =&\sigma_g(W_z x_t + U_z h_{t-1} + b_z) \\ r_t =&\sigma_g(W_r x_t + U_r h_{t-1} + b_r) \\ \hat{h}_t =&\phi_h(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \\ h_t =&(1 - z_t) \odot \hat{h}_t + z_t \odot h_{t-1} \end{aligned} \tag{A1}$$

The n_x-th lines of matrices W_* (noted $W_*^{n_x}$) had to be constrained in order to make $h_t = 0$ on all reset samples, whatever the values of h_{t-1} and whatever the other values on W_*, H_* and b_*. This can be achieved by forcing $z_t = 0$ and $\hat{h}_t = 0$. Next, $W_z x_t + U_z h_{t-1} + b_z$ should be the $-\infty$ vector and thus, the values of $W_z^{n_x}$ can be set to $-\infty$ (on practice, we set to -1.e30 to avoid computing problems).

$W_h x_t + U_h(r_t \odot h_{t-1}) + b_h$ should be 0 vector. Thus, we can force r_t to be 0 vector (so that $U_h(r_t \odot h_{t-1}) = 0$), and $W_h x_t + b_h = 0$. This last part can be achieved by setting $W_h^{n_x} = -b_h^T$ (the ^{T} means "transpose").

$r_t = 0$ is equivalent to $W_r x_t + U_r h_{t-1} + b_r = -\infty$. Hence, $W_r^{n_x}$ should be set to $-\infty$. In conclusion, to ensure that the state is reset at each reset sample, we added the following constraint to the kernel matrices:

$$\begin{aligned} W_z^{n_x} =&(-\infty) \\ W_r^{n_x} =&(-\infty) \\ W_h^{n_x} =&-b_h^T \end{aligned} \tag{A2}$$

References

1. Ayres-de-Campos, D.; Spong, C.Y.; Chandraharan, E. FIGO Consensus Guidelines on Intrapartum Fetal Monitoring: Cardiotocography. *International Journal of Gynecology & Obstetrics* **2015**, *131*, 13–24. doi:10.1016/j.ijgo.2015.06.020.
2. External and Internal Heart Rate Monitoring of the Fetus. In *Health Encyclopedia*.
3. Maternia, A.; Kupka, T.; Horoba, K.; Jezewski, J.; Martinek, R.; Wrobel, J.; Kahankova, R.; Czabanski, R.; Graczyk, S. New Possibilities for Fetal Monitoring Using Unobtrusive Abdominal Electrocardiography. 2019 MIXDES - 26th International Conference "Mixed Design of Integrated Circuits and Systems"; IEEE: Rzeszów, Poland, 2019; pp. 413–418. doi:10.23919/MIXDES.2019.8787051.
4. Odendaal, H.J. False Interpretation of Fetal Heart Role Monitoring in Cases of Intra-Uterine Death. *S Afr Med J* **1976**, *50*, 1963–1965.
5. Reinhard, J.; Hayes-Gill, B.R.; Schiermeier, S.; Hatzmann, H.; Heinrich, T.M.; Louwen, F. Intrapartum Heart Rate Ambiguity: A Comparison of Cardiotocogram and Abdominal Fetal Electrocardiogram with Maternal Electrocardiogram. *Gynecol Obstet Invest* **2013**, *75*, 101–108. doi:10.1159/000345059.
6. Murray, M.L. Maternal or Fetal Heart Rate? Avoiding Intrapartum Misidentification. *Journal of Obstetric, Gynecologic & Neonatal Nursing* **2004**, *33*, 93–104. doi:10.1177/0884217503261161.
7. Kiely, D.J.; Oppenheimer, L.W.; Dornan, J.C. Unrecognized Maternal Heart Rate Artefact in Cases of Perinatal Mortality Reported to the United States Food and Drug Administration from 2009 to 2019: A Critical Patient Safety Issue. *BMC Pregnancy Childbirth* **2019**, *19*, 501. doi:10.1186/s12884-019-2660-5.
8. Melchior, J.; Cavagna, J.; Bernard, N. Le Rythme Cardiaque Foetal Pendant l'expulsion de l'accouchement Normal. Médecine Prénatale, 6e Journées Nationales, Paris, 1977, pp. 225–32.
9. Riethmuller, D. How Long Is Too Long ? A Dilatation Complète, Peut-on Attendre Jusqu'à 4 Heures ?, 2018.
10. Nurani, R.; Chandraharan, E.; Lowe, V.; Ugwumadu, A.; Arulkumaran, S. Misidentification of Maternal Heart Rate as Fetal on Cardiotocography during the Second Stage of Labor: The Role of the Fetal Electrocardiograph: Erroneous Recording of Maternal Heart Rate. *Acta Obstetricia et Gynecologica Scandinavica* **2012**, *91*, 1428–1432. doi:10.1111/j.1600-0412.2012.01511.x.
11. Houzé de l'Aulnoit, A.; Boudet, S.; Demailly, R.; Delgranche, A.; Génin, M.; Peyrodie, L.; Beuscart, R.; Houzé de l'Aulnoit, D. Automated Fetal Heart Rate Analysis for Baseline Determination and Acceleration/Deceleration Detection: A Comparison of 11 Methods versus Expert Consensus. *Biomedical Signal Processing and Control* **2019**, *49*, 113–123. doi:10.1016/j.bspc.2018.10.002.
12. Pinto, P.; Costa-Santos, C.; Gonçalves, H.; Ayres-De-Campos, D.; Bernardes, J. Improvements in Fetal Heart Rate Analysis by the Removal of Maternal-Fetal Heart Rate Ambiguities. *BMC Pregnancy Childbirth* **2015**, *15*, 301. doi:10.1186/s12884-015-0739-1.
13. Boudet, S.; Houzé de l'Aulnoit, A.; Demailly, R.; Peyrodie, L.; Beuscart, R.; Houzé de l'Aulnoit, D. Fetal Heart Rate Baseline Computation with a Weighted Median Filter. *Computers in Biology and Medicine* **2019**, *114*, 103468. doi:10.1016/j.compbiomed.2019.103468.
14. Saeed, F.; Abeysuriya, S.; Chandraharan, E. Erroneous Recording of Maternal Heart Rate as Fetal Heart Rate During Second Stage of Labour: Isn't It Time to Stop This? *J Biomed Res Environ Sci* **2021**, *2*, 315–319. doi:10.37871/jbres1233.

15. Bhogal, K.; Reinhard, J. Maternal and Fetal Heart Rate Confusion during Labour. *British Journal of Midwifery* **2010**, *18*, 424–428. doi:10.12968/bjom.2010.18.7.48781.
16. Petrozziello, A.; Jordanov, I.; Aris Papageorghiou, T.; Christopher Redman, W.; Georgieva, A. Deep Learning for Continuous Electronic Fetal Monitoring in Labor. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); IEEE: Honolulu, HI, 2018; pp. 5866–5869. doi:10.1109/EMBC.2018.8513625.
17. Houzé de l'Aulnoit, A.; Parent, A.; Boudet, S.; Rogoz, B.; Demailly, R.; Beuscart, R.; Houzé de l'Aulnoit, D. Development of a Comprehensive Database for Research on Foetal Acidosis. *European Journal of Obstetrics & Gynecology and Reproductive Biology* **2022**, *274*, 40–47. doi:10.1016/j.ejogrb.2022.04.004.
18. Houzé de l'Aulnoit, A.; Boudet, S.; Génin, M.; Gautier, P.F.; Schiro, J.; Houzé de l'Aulnoit, D.; Beuscart, R. Development of a Smart Mobile Data Module for Fetal Monitoring in E-Healthcare. *J Med Syst* **2018**, *42*, 83. doi:10.1007/s10916-018-0938-1.
19. Boudet, S.; Houzé l'Aulnoit, A.; Demailly, R.; Delgranche, A.; Peyrodie, L.; Beuscart, R.; Houzé de l'Aulnoit, D. A Fetal Heart Rate Morphological Analysis Toolbox for MATLAB. *SoftwareX* **2020**, *11*, 100428. doi:10.1016/j.softx.2020.100428.
20. Karpathy, A. The Unreasonable Effectiveness of Recurrent Neural Networks, 2015.
21. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **1997**, *9*, 1735–1780. doi:10.1162/neco.1997.9.8.1735.
22. Cho, K.; van Merrienboer, B.; Bahdanau, D.; Bengio, Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. arXiv, 2014, pp. 103–111. doi:10.48550/ARXIV.1409.1259.
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Las Vegas, NV, USA, 2016; pp. 770–778. doi:10.1109/CVPR.2016.90.
24. Chudáček, V.; Spilka, J.; Burša, M.; Janků, P.; Hruban, L.; Huptych, M.; Lhotská, L. Open Access Intrapartum CTG Database. *BMC Pregnancy Childbirth* **2014**, *14*, 16. doi:10.1186/1471-2393-14-16.