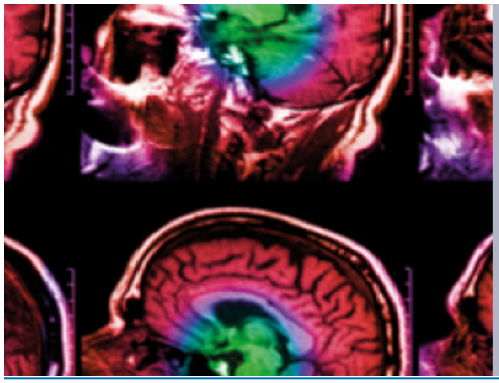
PAPER • OPEN ACCESS

A dilated inception CNN-LSTM network for fetal heart rate estimation



IOP Publishing Physiol. Meas. 42 (2021) 045007 https://doi.org/10.1088/1361-6579/abf7db

Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

PAPER

A dilated inception CNN-LSTM network for fetal heart rate estimation

E Fotiadou[1], R J G van Sloun[1], J O E H van Laar[2] and R Vullings[1]

[1] Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, 5612 AP, The Netherlands
[2] Department of Obstetrics and Gynaecology, Máxima Medical Center, Veldhoven, 5504 DB, The Netherlands

E-mail: E.Fotiadou@tue.nl

Keywords: convolutional neural networks, dilated convolution, fetal electrocardiogram, fetal heart rate, long short-term memory networks, noninvasive fetal ECG

OPEN ACCESS

RECEIVED
12 October 2020

REVISED
6 April 2021

ACCEPTED FOR PUBLICATION
14 April 2021

PUBLISHED
13 May 2021

Abstract

Objective. Fetal heart rate (HR) monitoring is routinely used during pregnancy and labor to assess fetal well-being. The noninvasive fetal electrocardiogram (ECG), obtained by electrodes on the maternal abdomen, is a promising alternative to standard fetal monitoring. Subtraction of the maternal ECG from the abdominal measurements results in fetal ECG signals, in which the fetal HR can be determined typically through R-peak detection. However, the low signal-to-noise ratio and the nonstationary nature of the fetal ECG make R-peak detection a challenging task. *Approach*. We propose an alternative approach that instead of performing R-peak detection employs deep learning to directly determine the fetal HR from the extracted fetal ECG signals. We introduce a combination of dilated inception convolutional neural networks (CNN) with long short-term memory networks to capture both short-term and long-term temporal dynamics of the fetal HR. The robustness of the method is reinforced by a separate CNN-based classifier that estimates the reliability of the outcome. *Main results*. Our method achieved a positive percent agreement (within 10% of the actual fetal HR value) of 97.3% on a dataset recorded during labor and 99.6% on set-A of the 2013 Physionet/Computing in Cardiology Challenge exceeding top-performing state-of-the-art algorithms from the literature. *Significance*. The proposed method can potentially improve the accuracy and robustness of fetal HR extraction in clinical practice.

1. Introduction

Hypoxia that occurs when the brain does not receive adequate oxygen poses a significant risk for fetuses. The failure of oxygen delivery to the fetus can cause permanent brain damage, developmental delays, or even death in severe cases. Although oxygen deficiency can happen at any stage during pregnancy most injuries typically occur during labor. The heart rate (HR) pattern of the fetus changes as a response to reduced oxygenation (Sundström *et al* 2000, Shaw *et al* 2016). Therefore, the fetal HR must be monitored during labor but also pregnancy, especially in high-risk pregnancies.

Many hospitals routinely use continuous electronic fetal HR monitoring during pregnancy and labor. Electronic fetal HR monitoring measures fetal HR in response to the contractions of the uterus and can be performed internally or externally. External fetal HR monitoring is the most common method and can be performed by placing an ultrasound transducer on the maternal abdomen. However, this method often provides inaccurate results, as it is affected by the movement of the mother and the fetus and suffers from signal loss in case of obese patients (Kahankova *et al* 2020). Internal monitoring is carried out by placing an electrode on the fetal scalp (Sundström *et al* 2000). It provides a more accurate and consistent HR because factors such as movement have a smaller impact on the quality of the measurement. However, scalp fetal HR measurements can only take place during labor after the membranes have ruptured and there is sufficient dilation of the cervix. Additionally, the scalp electrode may cause injury to the fetus.

An alternative way of measuring the fetal HR externally is by measuring the noninvasive fetal electrocardiogram (ECG) by placing electrodes on the maternal abdomen. This method has the potential to

provide accurate measurements and can be performed both during pregnancy and labor. As opposed to ultrasound measurements, the fetal ECG recordings enable beat-to-beat HR extraction, necessary for reliable analysis of the HR variability (Jezewski *et al* 2006). To extract the fetal HR from the fetal ECG recordings, one needs to detect the fetal R-peaks since the HR is related to the distance between two successive peaks. However, the noninvasiveness of the fetal ECG comes at the cost of a reduction in signal-to-noise ratio (SNR) (Sameni and Clifford 2010). The noninvasive abdominal recordings are severely contaminated by electrical interferences such as the maternal ECG, powerline interference, muscle noise, equipment noise, etc, thus complicating the detection of the fetal R-peaks.

Several methods have been proposed in the literature in the area of fetal ECG and fetal HR extraction from noninvasive abdominal recordings. The 2013 Physionet/Computing in Cardiology Challenge aimed to encourage fetal HR estimation from abdominal recordings (Silva *et al* 2013). Along with the challenge, a database became publicly available to allow for the comparison of different algorithms. The methods presented in the challenge were unique but, as pointed out in the review of Clifford *et al* (2014), most followed a five-step approach. This includes preprocessing of the abdominal signals, estimation, and suppression of the maternal ECG signal, R-peak detection, and postprocessing of the fetal HRs. A variety of algorithms were proposed for maternal ECG subtraction, such as adaptive filtering (Adam and Shavit 1990, Widrow *et al* 1975, Sameni 2008), template subtraction (Ungureanu *et al* 2007, Vullings *et al* 2009), blind source separation (Kanjilal *et al* 1997, Martín-Clemente *et al* 2002, Ye *et al* 2009, Camargo-Olivares *et al* 2011) or a combination of different algorithms (Vigneron *et al* 2003, Wu *et al* 2013, Behar *et al* 2014). For an extensive review of fetal ECG extraction methods see Sameni and Clifford (2010) or Behar *et al* (2016). However, even after the maternal ECG is removed, the SNR of the extracted fetal ECG signals is usually low, and this can result in faulty R-peak detections.

The winner of the challenge, Varanini *et al* (2013), used two QRS detectors in forward and backward directions after enhancing the extracted multichannel fetal ECG signal with independent component analysis (ICA). Afterwards, the fetal ECG channel with the best R–R series was selected for HR estimation. However, since the orientation of the fetus can change, causing variations to the SNR of each channel, using multichannel information can be more robust. Warmerdam *et al* (2018) suggested an adaptive multichannel R-peak detection method that combines HR information and ECG waveform. However, a linear autoregressive model was used as an HR model that is not able to describe complex accelerations and decelerations that occur during labor. When tested in set-A of the 2013 Physionet/Computing in Cardiology Challenge (Silva *et al* 2013), Warmerdam *et al* achieved 99.6% accuracy in R-peak detection while Varanini *et al* 98.6%. Since the challenge, more research has been performed in the area of fetal HR extraction with promising results (Zhang *et al* 2017, Jamshidian-Tehrani and Sameni 2018).

Deep learning methods have achieved remarkable success in tasks such as image classification (Waseem and Zenghui 2017) and speech recognition (Nassif *et al* 2019) and the expectations on how this technology could help to improve health care are high (Esteva *et al* 2019). Several works have been reported in the area of ECG processing such as adult ECG denoising (Xiong *et al* 2016, Antczak 2018), adult arrhythmia detection (Isin and Ozdalili 2017), prediction of fetal acidemia (Zhao *et al* 2019), fetal ECG denoising (Fotiadou *et al* 2020, Fotiadou and Vullings 2020) and fetal ECG signal reconstruction (Muduli *et al* 2016). Present signal processing algorithms for fetal HR extraction have limited performance because there are no accurate models for the noise that remains after the maternal ECG is suppressed. However, complex deep learning models might be more efficient in HR estimation especially in low-SNR signals. Two works in the literature already attempted fetal QRS detection with deep learning methods. Zhong *et al* (2018) proposed a convolutional neural network (CNN) model for QRS complex detection from noninvasive abdominal recordings without canceling the maternal ECG. Subsequently, Lee *et al* (2018) advanced this approach by using multichannel signals, deeper architecture, and postprocessing, leading to an improved positive predictive value of 92.77%. However, both methods did not manage to reach the performance of conventional signal processing algorithms. Moreover, there is a limitation in the validation of both approaches to only seven subjects of the set-A of the 2013 Physionet/Computing in Cardiology Challenge dataset (Silva *et al* 2013).

Recently, the authors proposed a deep learning model that directly estimates the fetal HR from the extracted fetal ECG signals (Fotiadou *et al* 2020). The model, which combines CNNs with long short-term memory (LSTM) networks, achieved comparable performance with a top-performing state-of-the-art HR extraction algorithm. A limitation of the method was that it failed to correctly estimate the HR in cases of extremely low-quality signals obtained during the second stage of labor, especially when decelerations happened. In clinical practice, absence of information is unfavorable but wrong information is worse as it might lead to misdiagnosis.

In this paper, we further improve and extend our previous work aiming at more robust and efficient HR estimation. Inspired by Shi *et al* (2017), we propose to use a dilated inception CNN encoder in our network to achieve multiscale feature extraction and a variety of receptive fields. An LSTM decoder network is chained to

the encoder to learn long-term temporal feature relations and extract the fetal HR. Our main contributions are outlined as follows:

- A deep hybrid dilated inception CNN-LSTM (DICNN-LSTM) encoder-decoder network that extracts the fetal HR from noninvasive abdominal recordings. To the best of our knowledge, we are the first that employed deep learning to estimate the fetal heart HR without explicitly detecting the QRS complexes.
- The reliability of our method is reinforced by a classifier, which uses a CNN network to identify the periods of time that the extracted fetal HR is inaccurate, increasing the suitability of our method for clinical application.
- Experimental results demonstrate the advantage of our method over top-performing state-of-the-art algorithms.

The rest of the paper is organized as follows. Section 2 presents the proposed fetal HR extraction method, the HR reliability classifier and the data used. Experimental results are provided in section 3. Finally, the results are discussed in section 4 and conclusions are drawn in section 5.

2. Materials and methods

2.1. Data

Two different datasets were used in this study. The first one is a private dataset obtained in a collaboration between the Eindhoven University of Technology and the Máxima Medical Center, Veldhoven, The Netherlands. This dataset is a part of the study described in Lempersz *et al* (2020). It contains 28 abdominal recordings measured by four electrodes during labor at a sampling frequency of 500 Hz. The data were collected from 28 women with a gestational age between 36 and 42 weeks with a total duration of roughly 91 h. Each recording was partially obtained during the first and second stage of labor. Simultaneous scalp HR recordings were performed and stored at 4 Hz. In nearly all recordings, clock drift was present between the scalp fetal HR and the fetal ECG, leading to desynchronization after a certain amount of time. To limit this desynchronization, the fetal HR was resampled using a manually determined resample factor. The fetal ECG signals were extracted from the abdominal measurements by the methods described in Vullings *et al* (2009), Warmerdam *et al* (2016). According to Warmerdam *et al* (2016), a fixed-lag Kalman smoother with adaptive noise estimation was used to filter the powerline interference. After this step, the maternal ECG was suppressed by a template subtraction technique, known as weighted averaging of maternal ECG segments (WAMES) (Vullings *et al* 2009). WAMES method dynamically segments the maternal ECG complex in separate parts and generates an individual template for each part. Each template is determined by linearly combining time-shifted, offset-compensated, and scaled corresponding parts in previous complexes. The individual templates are then combined to yield a maternal ECG template, which is subsequently subtracted from the recorded, preprocessed data. At the moment of conducting the study, the authors were provided merely with the extracted fetal ECG data, while access to the raw data was not available. Figure 1 illustrates two examples of fetal ECG signals contained in this dataset. Figure 1(a) presents a signal recorded during the first stage of labor, while figure 1(b) a signal obtained during second stage of labor. Note that the vertical axis limits for the signals in figures 1(a) and (b) differ for better visualization. Notably the signal in figure 1(a) has significantly higher quality than the signal in figure 1(b). All the signals of this dataset obtained during second stage of labor have high amounts of noise, making it virtually impossible to distinguish the fetal R-peaks.

The second dataset is the set-A of the 2013 Physionet/Computing in Cardiology Challenge (Silva *et al* 2013). It consists of 75 1 min noninvasive abdominal signals sampled at 1000 Hz. The data were obtained from multiple sources, using a variety of instrumentation with different frequency response, resolution and configuration, although in all cases they are presented at 1000 samples per second. Reference annotations of the fetal QRS complexes were made available as well which enabled us to determine and evaluate the fetal HR. The reference annotations were produced, usually with reference to a direct scalp fetal ECG signal. Following the suggestion of Behar *et al* (2013), organizer of the challenge, seven recordings (a33, a38, a47, a52, a54, a71, and a74) were discarded due to inaccurate annotations. The algorithm of Varanini *et al* (2013) was used to extract the fetal ECG signals from the abdominal recordings. According to the method of Varanini *et al* first the baseline wander and the powerline interference were removed. Afterwards, the maternal ECG was estimated through ICA and singular value decomposition and subsequently subtracted from the signals. Finally, a second ICA was employed to enhance the fetal ECG signal. We need to note that since our fetal HR extraction framework was developed for signals sampled on 500 Hz, before applying our algorithm on the Physionet dataset, but after applying Varanini's method, the signals were resampled to 500 Hz.

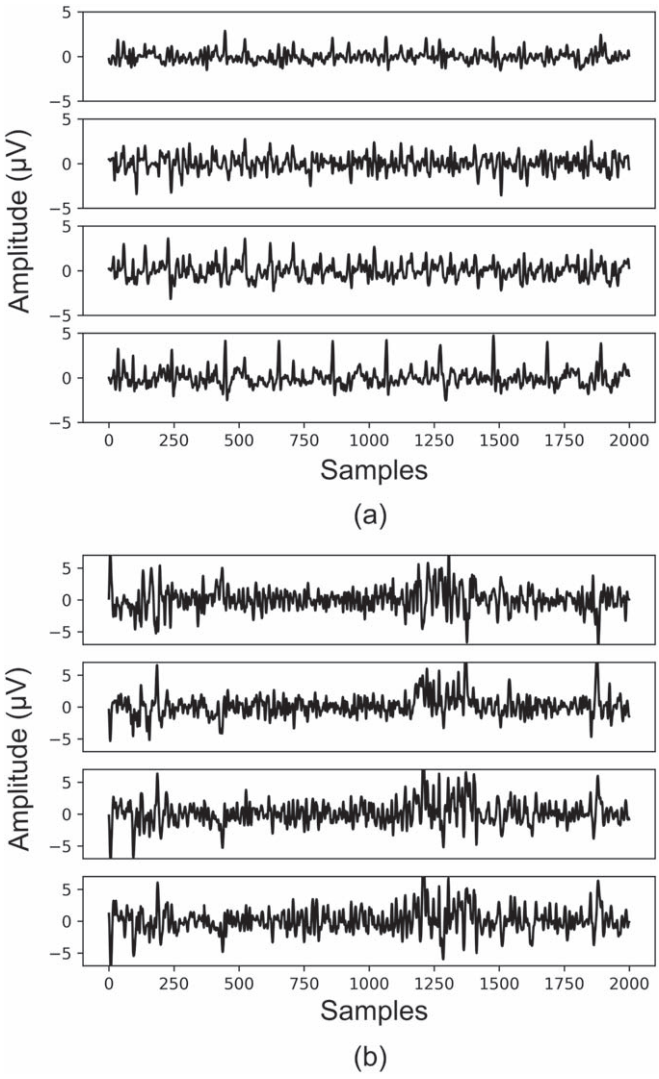


Figure 1. Two example 4-channel fetal ECG signals from our private dataset (Lempersz *et al* 2020), where the signal presented in (a) was obtained during first stage of labor and in (b) during second stage of labor.

Sixteen (54 h) out of the 28 recordings of our private dataset were randomly selected and used to train the fetal HR extraction network. Six recordings of 22.5 h were kept as validation set to tune the parameters of the network. The remaining six (14.5 h) recordings of our private set, together with the set-A of the Physionet database (68 min) (Silva *et al* 2013), were kept as a test set to evaluate the performance of the network. The scalp fetal HR was used as the desired output of the network (labels).

2.2. Fetal HR extraction

In this section, we present our DICNN-LSTM model for fetal HR extraction from noninvasive fetal ECG signals. The model, which is illustrated in figure 2, is comprised of two main blocks. The DICNN network block (encoder) consists of six stacked dilated convolution inception modules, which are depicted in detail in figure 3. The DICNN network is used as a feature extractor and the extracted features are fed to the LSTM network block (decoder) that is responsible for estimating the fetal HR.

2.2.1. Network description

2.2.1.1. Dilated convolution

Dilated convolutions were originally developed for wavelet transforms (Holschneider *et al* 1990) and later proposed for multiscale context aggregation in Yu and Koltun (2016).

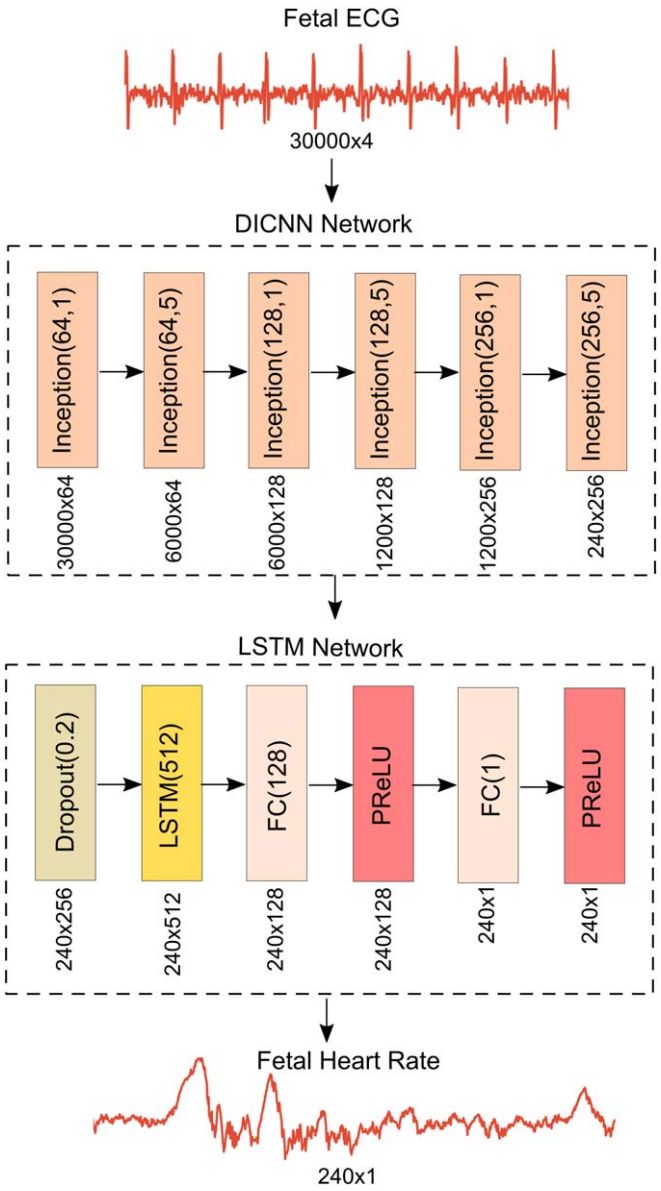


Figure 2. Overview of the proposed DICNN-LSTM network. A dilated inception CNN network extracts features from the input 4-channel fetal ECG signal (1 min duration), which are then fed to an LSTM network for the estimation of the fetal heart rate (four values/s). Next to each layer the output layer size is provided. The inception (*i*, *j*) block is given in more detail in figure 3, while *i* indicates the number of filters of the convolutional layers and *j* the size of the max pooling window. FC stands for a fully connected layer.

Given a 1D signal x and a kernel w, the output y of a dilated convolution is defined as

$$y[n] = \sum_{k=0}^{K-1} w[k]x[n - dk], \tag{1}$$

where K and d are the kernel size and the dilation factor respectively. Notice than when d is 1, then the dilated convolution is the same as the conventional convolution. Figure 4 illustrates the 1D dilated convolution operation with dilation rates 1, 2, and 4 and kernel size 3. As shown in the figure, d-1 signal samples are skipped during the dilated convolution.

Dilated convolutions are used to increase the receptive field of the network, i.e. the region in the input space that a CNN feature is affected by. When using conventional convolutions, the receptive field is linearly related to the depth of the layer. Dilated convolutions can achieve a receptive field that is exponentially related to the layer depth when exponentially increasing dilation rates ($d = 1, 2, 4, \ldots$) are used.

2.2.1.2. Encoder network

Szegedy *et al* (2015) proposed an inception model for image classification, intending to capture multiscale information in the input images. Salient signal parts can have large variations in size and thus choosing the right

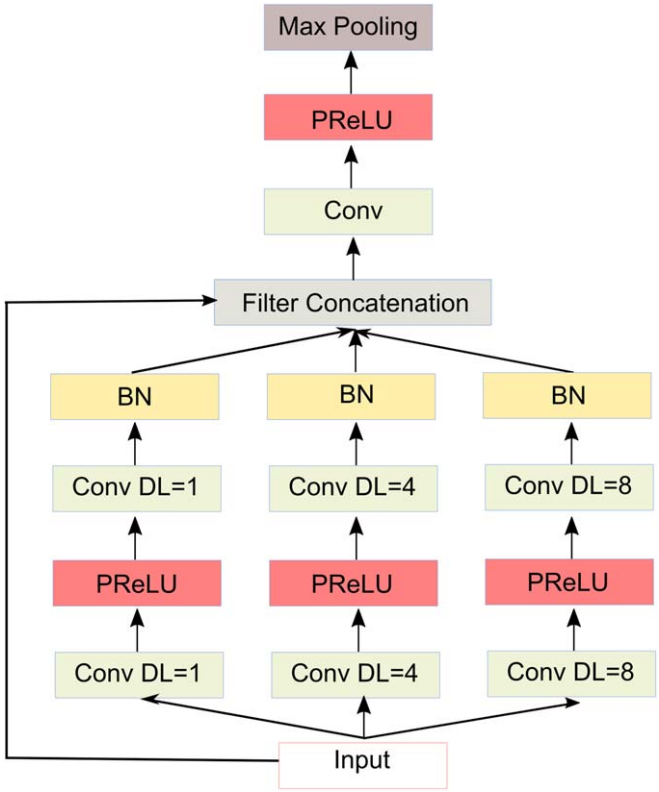


Figure 3. The dilated inception block used in the DICNN-LSTM network of figure 2. Conv stands for a convolutional layer, DL for dilation rate, and BN for a batch normalization layer.

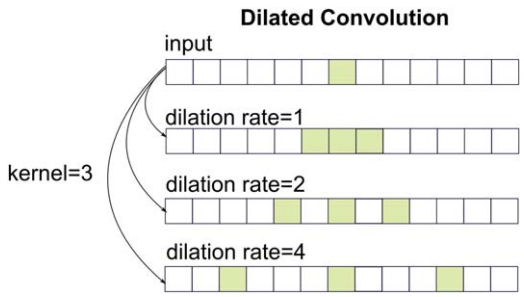


Figure 4. 1D dilated convolution with a kernel size of 3 and dilation rates 1, 2, and 4. The green block in the input signal (first row) indicates the unit of interest. In the output signals (second, third, and fourth row) the green blocks show the receptive field for each different dilation rate.

kernel size is not easy. Larger kernel sizes are preferred for more globally spread information, while smaller sizes are preferred for local information. Moreover, very deep networks are susceptible to overfitting and are computationally expensive. To address these issues, the original inception module was developed that uses a wider instead of deeper network by allowing filters of different kernel sizes to operate on the same level. More recently, Shi *et al* (2017) presented a dilated convolution inception model for single image super-resolution. In their model, instead of using convolutions with different kernel sizes, they employed convolutions of different dilation rates to learn multiscale information.

Inspired by both works we propose a stacked dilated inception convolutional encoder network to extract useful features for fetal HR extraction from noninvasive fetal ECG signals. We found that for our problem varying the dilation rate rather than the kernel size leads to better performance (in the validation dataset). As shown in figure 3, each inception module of the encoder consists of three parallel branches of two convolutional layers of different dilation rates (1, 4, and 8). Since our data are temporal, we use 1D convolutions. Batch normalization layers are used after the convolutions to speed up the training process by normalizing the intermediate outputs. The outputs of these layers are subsequently concatenated together with the input of the

module through a residual connection. The role of the residual connection is twofold. First, it allows the input features to be reused, leading to better performance. Second, it allows the gradients to propagate easier across our deep network preventing the vanishing gradient problem. A convolutional layer is applied after the concatenation of the features to select the most meaningful features and reduce the dimension of the feature vector. At the end of the module, a max pooling layer is used to reduce the temporal dimension of the signal.

The parametric rectified linear unit (PReLU) (He *et al* 2015) is used as an activation function between the layers. In contrast to the rectified linear unit (ReLU), in which the negative part is completely dropped, PReLU assigns a nonzero slope to it. The output of PReLU, $f(z)$, for an arbitrary input z, is defined as

$$f(z) = \begin{cases} z, & \text{for } z > 0 \\ az, & \text{for } z \leqslant 0 \end{cases}, \tag{2}$$

where the slopes α are optimized during the training of the network. The motivation to use PReLU is that it solves the 'dying ReLU' problem when ReLU neurons become inactive and output zero for any input. Moreover, when tested in the validation set, we noticed a slight increase in our HR estimation accuracy when using PReLU instead of ReLU.

Six of these dilated convolution-based inception modules are stacked together to form the encoder of our model. We tried also different numbers of inception modules but six gave the best performance based on tests on our validation dataset. The encoder is both wide and deep and thus relatively complex, which is necessary to extract relevant information from the typically very noisy fetal ECG signals. A 240 $\times$ 256 feature vector is extracted from the fetal ECG signals at the output of the encoder.

2.2.1.3. Decoder network

The extracted feature vector is sent to the decoder network that is responsible for estimating the fetal HR. The first layer of the decoder is a dropout layer with a dropout rate of 0.2. This layer randomly ignores 20% of the neurons in the corresponding layer during training. This means that the weights for these neurons will not be updated on the backward pass. As a result, the network becomes less sensitive to the specific weights of the neurons and consequently can generalize better and is less likely to overfit.

Then, an LSTM layer with 512 nodes is used to model the temporal dynamics of the extracted features. LSTMs (Sherstinsky 2020) were developed as an improved form of recurrent neural networks (RNNs) to handle the problem of vanishing and exploding gradient. Instead of simple RNN neurons, LSTMs are characterized by more complex memory blocks that consist of several gates to control the information flow in the internal memory cells. They were explicitly designed for learning long-term dependencies in sequential data, meaning that they can remember and associate past with present information. Thus, an LSTM layer is stacked after the feature extraction step and dropout to learn dependencies in the feature sequence. The LSTM output is fed to two fully connected layers with 128 and 1 neurons respectively that output the fetal HR. PReLU activation functions are applied also to the decoder of our network. The parameters of both the encoder and decoder of our network were determined such that the fetal HR estimation performance was maximized on our validation dataset. Due to the large number of parameters cross validation to tune them was prohibitive.

2.2.1.4. Input and output signals

The input signal of the network is a four-channel ECG signal of one-minute duration. Since the average normal fetal HR is 100–160 bpm, the input signal typically contains 100–160 heartbeats. Because our reference HR from the scalp electrode is 4 HR values/s (section 2.1), the output of the network is set to 240 fetal HR values (60 s $\times$ 4 values/s).

To speed up the training process, the input signal x is normalized over the four channels separately by subtracting the mean and dividing by the standard deviation of the signals through

$$x_{\text{norm}} = \frac{x - E[x]}{\sqrt{\text{Var}[x] + \epsilon}}, \tag{3}$$

where $E[x]$ and $\text{Var}[x]$ are the mean and the variance of the input signal respectively. A small constant ϵ, chosen as 0.001, is added to the denominator to prevent division by zero. The output signals (labels) are not normalized.

2.2.1.5. Loss function

Mean square error (MSE) is the most commonly used loss function in regression problems. As the name suggests, MSE measures the square difference between the HR estimations and the actual observations. Due to the squaring, estimations that are far away from their actual values are penalized more heavily. However, when estimating the fetal HR less importance should be given to clear outliers because these will be ignored by the clinicians in their visual inspection of the fetal HR traces. For this reason, the mean absolute error (MAE) was selected as a loss function for our problem. MAE is defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\mathrm{FHR}_{\mathrm{target}}i - \mathrm{FHR}_{\mathrm{predicted}}i|, \tag{4}$$

where N is the length of the output sequence ($N = 240$), $\mathrm{FHR}_{\mathrm{predicted}}$ is the estimated fetal HR and $\mathrm{FHR}_{\mathrm{target}}$ is the fetal HR measured by the scalp electrode.

2.2.1.6. Training

The Adam algorithm (Kingma and Ba 2015) was selected as an optimization algorithm while the learning rate was set to 0.0001. The batch size was set to 8. At each training iteration, eight random one-minute segments were chosen from the 54 h training sequences. The scalp fetal HR was used as the desired output of the network. The network was trained for 1000 epochs and the model that minimized the loss on the validation set was finally chosen.

2.2.2. Performance evaluation

2.2.2.1. Evaluation metrics

The accuracy of the proposed method was assessed by measuring the MAE (equation (4)) and the mean squared error defined by

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{FHR}_{\mathrm{target}}i - \mathrm{FHR}_{\mathrm{predicted}}i)^2. \tag{5}$$

We determined the reliability of our method in terms of positive percent agreement (PPA), which is the percentage of fetal HR outputs that were within 10% of the actual HR value (Cohen *et al* 2012). We also present results where we used a smaller tolerance of 5% for the calculation of PPA (we will call this PPA_5). In addition, we used a second reliability metric, the coverage, that corresponds to the time that the method outputted a nonzero fetal HR value.

2.2.2.2. Reference methods

The performance of our method was compared to the performance of the algorithm of Warmerdam *et al* (2018), that outperformed the algorithms of Varanini *et al* (2013) and Behar *et al* (2014), when tested in set-A of the 2013 Physionet /Computing in Cardiology Challenge dataset (Silva *et al* 2013). Warmerdam proposed a multichannel hierarchical probabilistic framework for detecting the fetal R-peaks. Two models are incorporated in this framework, a gaussian QRS model and an autoregressive HR model. Their framework consists of three inference levels for inferring each next R-peak location: state estimation, QRS and HR model estimation, and noise estimation. Initially, the QRS model is used to determine the next R-peak location. A sanity check is performed to evaluate if this is indeed an R-peak location and if not, the location is extrapolated based on the HR model. In addition to R-peak detection, the algorithm of Warmerdam *et al* identifies the periods of time that the extracted fetal HR is unreliable and does not output an HR value for these. The code of the algorithm was provided by the authors for our experiments.

We performed an additional comparison of our method with the algorithms of Varanini *et al* (2013) and Behar *et al* (2014) in the Physionet dataset. Both Varanini *et al* (2013) and Behar *et al* (2014) are fetal ECG extraction approaches that also calculate the fetal HR. We did not compare on our private dataset because we have access only to the extracted fetal ECG signals and not the raw data. For both algorithms we used online implementations provided by the authors in the Physionet website.

After the fetal ECG extraction process (section 2.1), the method of Varanini performs R-peak detection on individual channels using a two-step approach. First, a derivative filter is used for R-peak detection. Second, a forward and backward autoregressive model is trained from the obtained RR-series. The autoregressive model is subsequently used in combination with the derivative signal to detect the fetal R-peaks. In the end, the channel with the best RR-series is chosen based on some statistical features of the RR-series.

Behar uses a combination of various source separation techniques to cancel the maternal ECG such as template subtraction and blind source separation methods. Fetal R-peak extraction is then performed using a Pan and Tompkins R-peak detector on each separate fetal ECG channel. Finally, the channel with the smoothest RR-series is selected.

2.3. Fetal HR reliability estimation

When the fetal ECG signals have extremely low quality it is difficult to accurately estimate the fetal HR. To increase the robustness of our network, we propose to use a simple classification framework that determines whether the extracted fetal HR is reliable or not. In fact, the classifier determines if a fetal ECG segment could yield a reliable fetal HR depending upon its signal quality. Figure 5 illustrates the proposed classifier for fetal HR

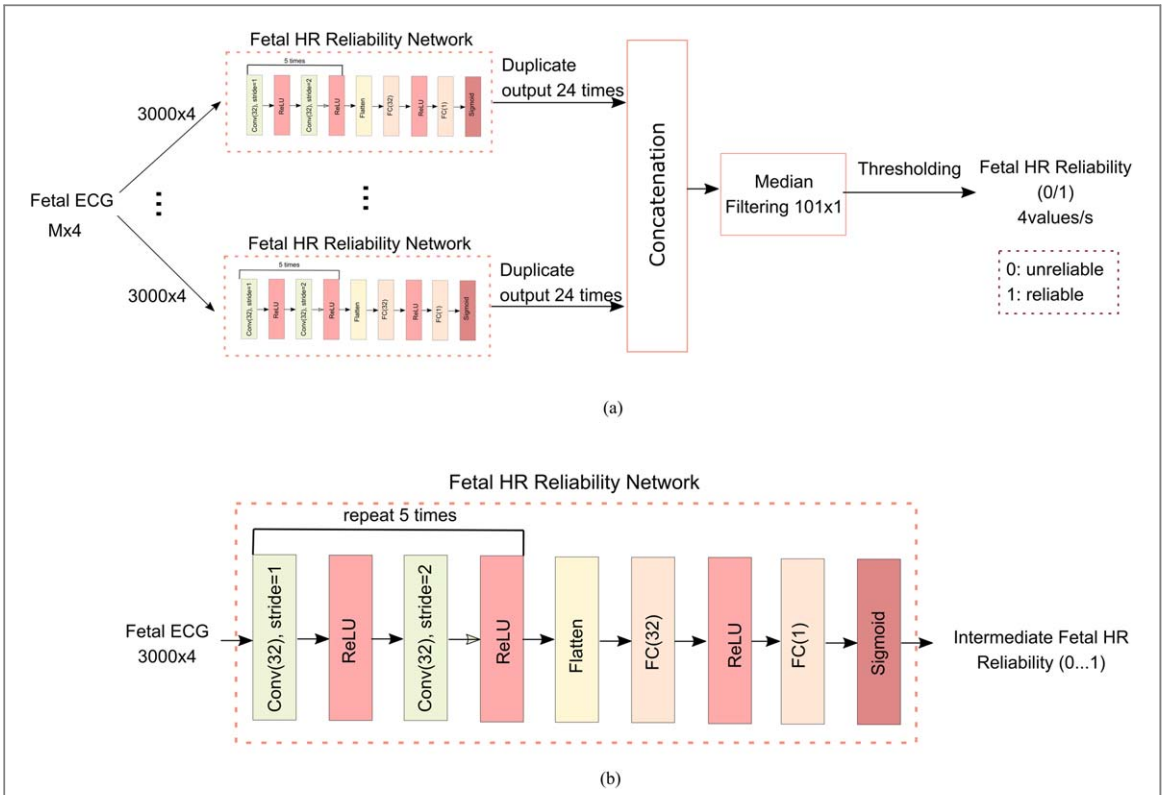


Figure 5. (a) Illustration of the fetal heart rate reliability classifier. A CNN network is used to classify each 3000-sample-long segment of the 4-channel fetal ECG (M samples) and the outputs are subsequently concatenated and filtered to obtain the final result. (b) Description of the CNN network used to classify each fetal ECG segment. Conv stands for convolutional layer and FC for fully connected.

reliability estimation. The input 4-channel fetal ECG signal is initially segmented in parts of 3000 samples that correspond to 6 s. Each segment passes through a CNN network that produces a value between 0 (unreliable) and 1 (reliable) indicating the reliability of the fetal HR. For each 6 s segment, one output value is obtained. However, since the DICNN-LSTM network outputs four fetal HR values/s, we copy the same value 24 times to obtain consistent outputs. Afterwards, we concatenate all outputs obtained from the 6 s signal segments. A median filter of size 101 (26 s) is subsequently applied to smoothen the reliability result for the whole input signal. Filtering removed some outliers and was found to slightly improve the classification performance on the validation dataset. The reasoning about our choice to filter the output as well as about the relatively large filter size is that it prevents the output from jumping between reliable and unreliable. Towards a clinical application we want to avoid such jumping as we want to prevent frequent raising of alarms. It is preferable to have more sustained alarms that could encourage the hospital staff to take measures to improve signal quality by e.g. asking the patient to move less. After the filtering operation, we set a cut-off threshold to determine if the final outcome will be 0 (unreliable) or 1 (reliable) for each fetal HR value.

2.3.1. Fetal HR reliability network

The fetal HR reliability network, shown in figure 5(b), consists of 10 1D convolutional layers followed by two fully connected layers. After every two convolutional layers subsampling by two is applied to the input (convolutional layers with stride 2). We empirically determined that employing 32 filters in each convolutional layer achieves a satisfactory result and avoids overfitting. The kernel size was set to 15 for all the convolutional layers. The two fully connected layers have 32 and 1 units respectively. All layers except for the last are followed by a ReLU activation function, while the last is followed by a sigmoid function. The latter produces a score between 0 and 1 indicating if the input will produce an unreliable (value close to 0) or reliable result (value close to 1).

2.3.1.1. Training

The same training data that were used to train the DICNN-LSTM model were employed for the training of the fetal HR reliability network, i.e. 16 recordings of our private dataset. However, the labels of the data were different because in this case we are not estimating HRs but reliabilities. To obtain the training labels, the data were initially passed through the DICNN-LSTM network and the HRs were estimated. Afterwards, the HR

estimations were compared with the actual HR values obtained by the scalp electrode. When the estimations were close to the target values ($|FHR_{target} - FHR_{prediction}| < 0.1 * FHR_{target}$), then the data were labeled as reliable with label 1, otherwise unreliable with label 0.

We trained the network by minimizing the binary cross-entropy between the HR estimations and the labels using the Adam optimizer with learning rate 0.0001. The batch size was 32, comprising 16 positive (reliable) and 16 negative (unreliable) stochastically selected examples. After training the network for 1000 epochs, the model that minimized the validation loss was selected.

2.3.2. Performance evaluation

The performance of the HR reliability classification network is assessed in terms of area under the receiver operator characteristic curve (AUC), true negative rate (TNR) or specificity, and true positive rate (TPR) or sensitivity. AUC metric ranges from 0 to 1, with 0 meaning that all estimations are wrong and 1 that all estimations are correct.

3. Results

In this section, we apply our proposed model for fetal HR extraction and report the experimental results for both our private and the Physionet datasets. The effectiveness of our method is validated both quantitatively as well as qualitatively. The results are presented in three parts. In the first part experiments regarding the DICNN-LSTM are provided. In the second part the fetal HR reliability classifier is evaluated and in the third part the results of the combined system are presented.

3.1. DICNN-LSTM network experiments

3.1.1. Kernel size

By changing the size of the kernels different features could be learned and the complexity of the network changes. We are interested in investigating if the detection accuracy changes when the dimensions of the convolutional window are altered. Thus, we performed experiments where we varied the kernel size of all the convolutional layers in the encoder of our DICNN-LSTM network but kept all the other parameters fixed. The experiments were performed on our validation dataset. The different kernels sizes that we tried are the following: 5, 11, 15, 19, 25, 35, and 50. For each one of them we repeated training the network three different times and then we averaged the performances of these three networks. The reason for this is that there is some randomness in the training process caused by differences in the initialization of the weights and random selection of training samples. Therefore, even when two networks are trained with the same architectural- and hyper-parameters, these trained networks will not be identical.

Figure 6 illustrates the performance of our network in terms of MSE, MAE and PPA when varying the kernel size. We notice that for smaller kernel sizes (up to 19) the performance of the network is relatively stable. This means that the kernel size does not seem to strongly affect the accuracy of the network and using any kernel size up to 19 is a good choice. However, for kernel sizes larger than 19 the performance starts to drop and becomes worse as the kernel size increases. This means that probably with large kernel sizes we miss details of some smaller features that are relevant for HR detection. Contrary to that, when using smaller kernel sizes, we detect not only smaller features but also larger ones since we have a relatively deep network with many convolutional layers stacked on top of each other. In addition, with the dilated inception scheme of our network we achieve a variety in the receptive fields of the network.

3.1.2. Model ensemble

Deep neural networks (DNNs) learn via a stochastic training algorithm that makes them sensitive to the training data and may learn a different set of weights each time they are trained that consequently produce different estimations. A way to reduce this high variance of the DNNs is to train multiple models and combine their outputs, i.e. to use ensemble learning. Ensemble learning typically results in more stable and improved estimations when compared to a single model. There are many ensemble learning techniques like varying the training data or the model architecture and the combination of the estimations. We decided to choose the combination of the three best models, trained with different kernel sizes, and average their outputs.

Table 1 shows the performance of the three best performing models together with the performance of the model ensemble. The best performances are marked in bold. The convolutional layers of these three models have kernel sizes of 5, 15, and 19. The performance of the three models is comparable while the combination of the separate models leads to more accurate results.

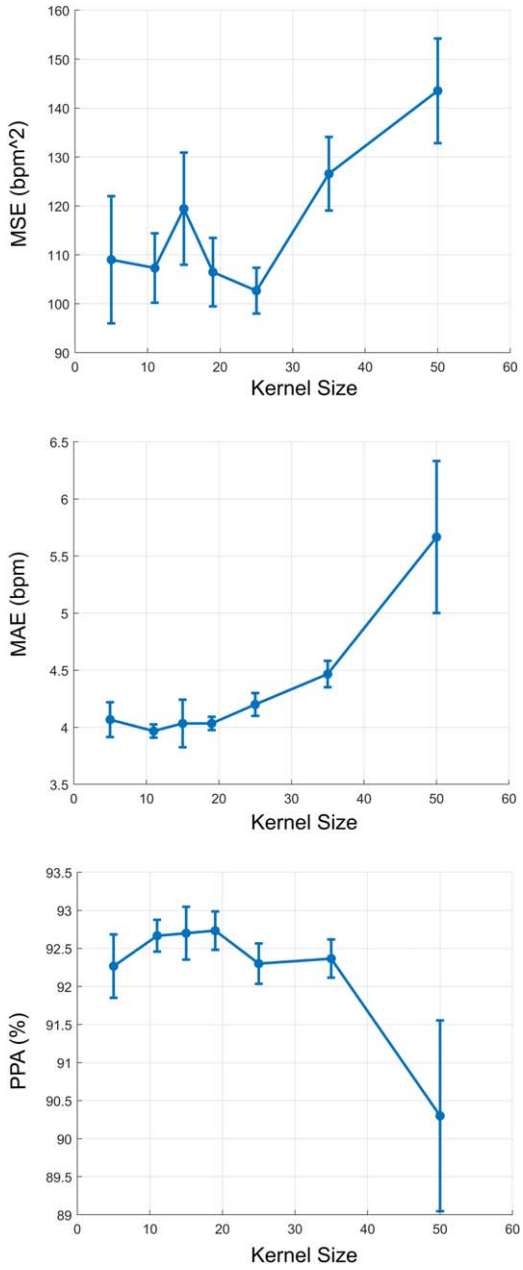


Figure 6. Performance of the DICNN-LSTM network when varying the kernel size of all the convolutional layers on the validation set. At each data point the standard deviation error bars show variations in performance due to differences in the initialization of the weights and random selection of training samples.

Table 1. Performance of the best perfoming DICNN-LSTM models and the ensemble of them on the validation set.

Model	MSE (bpm^2)	MAE (bpm)	PPA (%)
Model1_kernel5	101.5	3.9	92.6
Model2_kernel15	107	3.8	93.1
Model3_kernel19	111.6	4	93
Model Ensemble	**98.7**	**3.7**	**93.4**

3.2. Fetal HR reliability classifier performance

Figure 7 illustrates the performance of the fetal HR reliability estimations for the validation dataset. The classifier achieved AUC of 0.91. We additionally calculated the 95% AUC confidence interval but since our sample size is

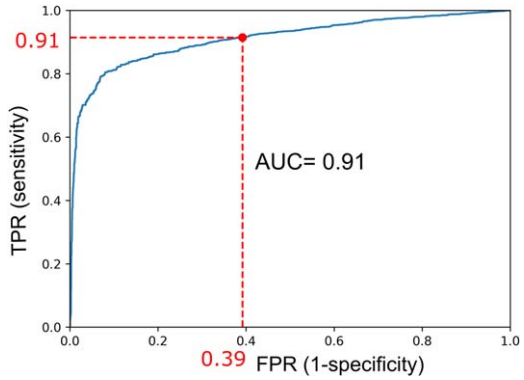


Figure 7. Receiver operator characteristic curve (true positive rate (TPR) plotted against false positive rate (FPR)) of the fetal HR reliability classifier for the validation set. The area under the curve (AUC) is specified in the plot. The 95% confidence margin of error was less than ±0.01. The red dot on the plot indicates the operating point when the classifier threshold is set to 0.55. In that case sensitivity and specificity have values 0.91 and 0.61 respectively.

Table 2. Performance of the fetal HR reliability classifier (for determining FPR and TPR classification threshold was set to 0.55).

Private test dataset			Physionet dataset		
AUC	TPR	TNR	AUC	TPR	TNR
0.92	0.92	0.66	0.83	0.82	0.75

very large (approximately 300 000) the upper and lower limits differ less than 0.01 from the AUC. By changing the threshold for our classifier, we can achieve the desired sensitivity and specificity for our problem. By choosing a high threshold we achieve higher sensitivity, meaning that most of the reliable data are indeed classified as reliable. Contrary to that, with a lower threshold higher specificity is obtained, leading to correctly identifying most of the unreliable data. We selected a threshold of 0.55 for our classifier and provided the TPR (sensitivity) and false positive rate (FPR) values as well in figure 7. The specificity can be calculated as 1-FPR. With the selected threshold the classifier can identify 91% of the reliable HRs and 61% of the unreliable ones.

We give slightly more importance to having a higher sensitivity than specificity for two reasons. First, we do not want to miss a lot of correctly estimated information to achieve a high coverage in the detection of the fetal HR values. Second, in clinical practice, even if some unreliable values are displayed on the screens of the clinicians, most of them look more like outliers, unconnected to the correct HR trace, and as such not taken into considerations in the clinical decision-making process.

Table 2 provides the classification performance for the two test datasets. The classifier achieved AUC of 0.92 for our private test dataset and 0.83 for the Physionet dataset. This means that the classifier can differentiate relatively well between the reliable and unreliable fetal HRs. The AUC of our private test dataset is similar to the one of the validation set but the AUC of the Physionet dataset is lower. The reason for this is that the Physionet dataset is highly unbalanced, containing 98.6% positive and only 1.4% negative examples. However, an AUC of 0.83 is still quite high meaning that our classifier generalized well to this dataset. The 95% AUC confidence interval for our private test set is very narrow ($<+-0.01$ difference from AUC) because of the large sample size. For the Physionet dataset, that is relatively small in size, the confidence interval for the AUC is [0.78, 0.86]. The TPR and TNR when the classifier threshold is set to 0.55 are also provided in table 2. According to the evaluation on the test sets the classifier can identify 82%–92% of the reliable HRs and 66%–75% of the unreliable ones.

3.3. Fetal HR extraction performance

Table 3 shows the performance of the fetal HR extraction network in comparison with the algorithm of Warmerdam *et al* (2018) for our private test dataset. In the table, we provided results both with and without incorporating the fetal HR reliability classifier. Our model achieved a PPA of 93.9% (91% for PPA_5), comparable with the method of Warmerdam with a PPA of 93.6% (91.5% for PPA_5). However, the algorithm of Warmerdam identifies the periods of time that the extracted HR is unreliable and does not output an HR value for these. Thus, for this dataset it has coverage of 94.5%. To have a fair comparison we evaluated the performance of our method also for the case where we excluded exactly the same periods. In this case the PPA achieved by our network is 96%, clearly outperforming the algorithm of Warmerdam. In addition, the proposed

Table 3. Fetal heart rate extraction performance on our private test dataset.

Algorithm	MSE (bpm^2)	MAE (bpm)	PPA (%)	PPA_5 (%)	Coverage (%)
DICNN-LSTM	104.5/59.6[a]	3.3/2.4[a]	93.9/96[a]	91/93.6[a]	100
DICNN-LSTM + Fetal HR reliability classifier	49.4	2	97.3	95.4	87.9
Warmerdam *et al* (2018)	129.7	3.4	93.6	91.5	94.5

[a] Calculated only in the periods that Warmerdam *et al* (2018) outputs a heart rate value.

method achieved significantly lower MAE and MSE, (2.4 versus 3.4 bpm for MAE and 59.6 versus 129.7 bpm^2 for MSE). When we discarded the unreliable fetal HR values, estimated by our reliability classifier, our PPA increased to 97.3% and the MAE and MSE decreased to 2 bpm and 49.4 bpm^2. However, our coverage fell to 87.9%, lower than Warmerdam's method. We additionally calculated the percentage of the time that the two methods in comparison agree on estimating the reliability of a fetal HR segment and we found that it is 90.1% on our test set.

We should note here that our private dataset is very challenging because it is partially obtained during the second stage of labor. At this stage, the uterine contractions are stronger and more frequent, and the woman is actively pushing leading to additional interferences from the abdominal muscles. As a result, the extracted fetal ECG signals have very low quality as they are strongly contaminated by noises. The significantly low SNR of the signals complicates the HR extraction, but by identifying these periods as unreliable, we are more confident that correct information will be presented to the clinicians. As an example, our reliability classifier labeled the fetal HR extracted from the signals of figure 1(a) as reliable and from the lower quality signals of figure 1(b) as unreliable.

Figure 8 depicts the fetal HR extraction by our network in comparison to the ground truth HR measured by the scalp electrode and the algorithm of Warmerdam. Four cases (a)–(d) of 7.5 min fetal HR segments are presented in the figure.

Results are demonstrated both with and without employing the HR reliability classifier. Note that the vertical axis limits for each case (a)–(d) are not the same for better visualization. As can be seen in figure 8(a) our method follows very precisely the reference HR obtained by the scalp electrode and so does the method of Warmerdam. The reliability classifier determined successfully that the estimations for this segment were reliable. Figure 8(b) is another example of almost perfect estimation. Here, the reliability classifier wrongly classified merely small signal portions as unreliable. In the case of figure 8(c) both methods failed to correctly determine the fetal HR deceleration parts. However, they successfully identified that the estimation was wrong. Figure 8(d) depicts a case that was partially wrongly estimated by our network with the reliability classifier being also partially successful in that case.

Table 4 demonstrates the performance of the fetal HR extraction on set-A of the Physionet/Computing in cardiology dataset for our network in comparison to the methods of Warmerdam, Varanini and Behar. Our network slightly surpasses the methods of Warmerdam and Varanini for this dataset, achieving a PPA of 98.6% as opposed to 97.9% and 98.4% respectively. The algorithm of Behar achieved lower performance with PPA of 91%. However, when we lowered the tolerance for the calculation of PPA the algorithm of Warmerdam was the one to achieve the best performance.

When we incorporated the reliability classifier our PPA increased to 99.6% but our coverage fell to 82% even though this dataset does not contain particularly low-quality fetal ECG signals. Moreover, our method achieved significantly lower MSE than the other methods, indicating that it provides fewer outliers. In terms of MAE the performances of all the methods are similar, apart from the one of Behar that is less accurate.

4. Discussion

Fetal R-peak detection in noninvasive fetal ECG recordings is demanding due to the low quality and the non-stationarity of the fetal ECG signals. In this work, we employed deep learning for directly determining the fetal HR from fetal ECG signals, without the need of R-peak detection. We proposed a deep hybrid dilated inception CNN-LSTM network that captures both short-term as well as long-term temporal HR patterns. To increase the reliability of our method, a classifier based on a CNN network was developed that estimates the accuracy of the detected fetal HR.

Initially, we performed experiments where we varied the kernel size of the convolutional layers of our DICNN-LSTM network. According to the results, we concluded that smaller kernel sizes are preferred, probably because features on smaller timescales are relevant for fetal HR extraction. Moreover, even when using small kernels, larger features are still exploited by our network due to the large receptive field achieved by using

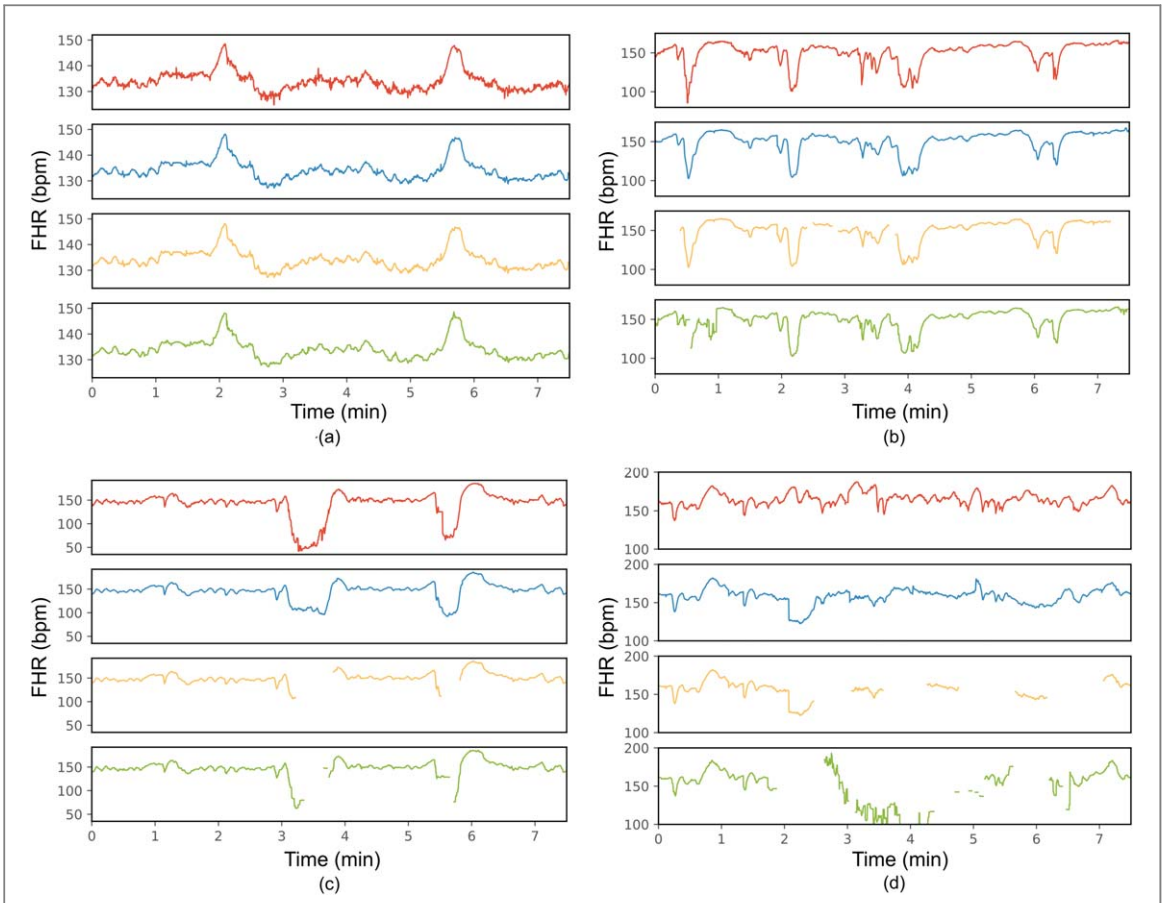


Figure 8. Four fetal heart rate (FHR) segments from our private test set of 7.5 min duration each. For each case a–d: the first subplot (red) depicts the scalp FHR, the second (blue) the estimated FHR from the DICNN-LSTM network, the third (yellow) the estimated FHR from the DICNN-LSTM network in combination with the reliability classifier (only reliable FHR values are displayed) and the fourth (green) the FHR obtained from the algorithm of Warmerdam *et al* (2018) (only when an FHR value is outputted).

Table 4. Fetal heart rate extraction performance on set-a of 2013 Physionet/Computing in Cardiology Challenge.

Algorithm	MSE (bpm^2)	MAE (bpm)	PPA (%)	PPA_5 (%)	Coverage (%)
DICNN-LSTM	14.2	1.6	98.6	95.6	100
DICNN-LSTM + Fetal HR reliability classifier	6.9	1.1	99.6	98.7	82
Warmerdam *et al* (2018)	30.8	1.5	97.9	96.5	100
Varanini *et al* (2013)	23.8	1	98.4	95.4	100
Behar *et al* (2014)	172	5.7	91	78.8	100

dilations in combination with a relatively deep network. In addition, we found that a model ensemble of several trained models with different kernel sizes leads to better performance than a single model. We decided to use the average of the three best performing models that we trained in our experiments as our fetal HR estimator.

The network achieved accurate estimations of the fetal HR for the most cases tested. However, mainly during the second stage of labor, there were cases that our method failed in the fetal HR extraction. Presenting false information to the clinicians can be very dangerous as this can lead to wrong decision making in critical moments during labor. Therefore, we developed a simple classifier that decides if the output of the DICNN-LSTM network is reliable so that in case of clinical application of the method only reliable information is presented to the clinicians. The developed classifier that is based on a simple deep CNN network managed to differentiate relatively well between the reliable and unreliable HRs (AUC of 0.92 and 0.83 for our private and the Physionet test dataset respectively). However, since the hyper-parameters of the classifier were not fully optimized, we are confident that the classification accuracy can be improved, and an even more robust system can be achieved.

Our DICNN-LSTM network achieved a PPA of 93.9% on a dataset obtained during labor similar to a top-performing state-of-the-art algorithm (Warmerdam *et al* 2018) (PPA of 93.6%) in the area of fetal HR extraction. However, since (Warmerdam *et al* 2018) excludes some unreliable fetal HR signal parts, when we also

excluded the same parts our PPA increased to 96%. It is remarkable that Doppler ultrasound, which is widely adopted in clinical practice, according to the literature, achieves much lower PPA. In a study conducted to a population of 75 women in labor (Cohen *et al* 2012), Doppler ultrasound obtained a PPA of 73%, while abdominal fetal electrocardiography 81.7%. In a second study (Euliano *et al* 2017), which recruited 71 women in labor, PPA for ultrasound was measured as 63%, while for noninvasive fetal ECG it was 84.4%.

When we combined the DICNN-LSTM network with the HR reliability classifier the PPA was raised to 97.3%, while the coverage of our method dropped to 87.9%. We should note here that we selected a threshold of 0.55, however, this might not be the optimal threshold according to a clinician. For clinical applications, the threshold could be adapted so as the desired importance should be given to the sensitivity and specificity of the classifier. Moreover, as an alternative to the binary output of the classifier (i.e. suppress fetal HR output or not), the classifier could also be used to provide reliability of the fetal HR and have clinicians appraise this information themselves to trust or distrust the provided fetal HR. We could also use a different approach to assess the reliability of the fetal HR. The variance of the estimations of the different models used in the ensemble might give some information about the reliability of the outcome. Possibly for that we need a bigger ensemble than the one used in this work.

Next to our private dataset, we tested our network on a public set of Physionet achieving a PPA of 98.6% outperforming the algorithms of Warmerdam, Varanini and Behar that scored 97.9%, 98.4% and 91% respectively. However, the algorithm of Warmerdam achieved the best performance when computing the PPA metric with lower tolerance. As reported in Warmerdam *et al* (2018), in the same dataset of Physionet, Warmerdam scored 99.6%, Varanini 98.6% and Behar 92.9% in terms of accuracy in R-peak detection. This implies that even though R-peaks can be detected accurately, the resulting fetal HR that is calculated from the R-peaks can be inaccurate. The main reason for this is the relatively wide range around an annotated R-peak during which a detected R-peak is still considered accurate. In combination with the HR reliability classifier the PPA achieved by our network increased to 99.6% for the Physionet database. However, the coverage of our method fell to 82% because many correctly estimated HRs were wrongly classified as unreliable.

Comparing the results obtained on the two different datasets we notice that the accuracy of the detection on the Physionet dataset is better. We believe that this is mainly due to the fact that our private dataset was partially obtained during the second stage of labor resulting in fetal ECG signals of lower quality as compared to the Physionet dataset. In addition, the Physionet dataset is significantly smaller in size than our dataset (68 min versus 14.5 h). Lastly, the algorithms used to extract the fetal ECG were different between the two datasets. We do not believe that this is the cause of the difference but can possibly account for a small part of it. Moreover, we notice that our approach clearly outperforms the one of Warmerdam on our private set (when compared in the same fetal HR signal parts) but in the Physionet dataset the performances are comparable. We believe that this is due to the limited ability of the model of Warmerdam to describe complicated accelerations and decelerations during labor. The DICNN-LSTM method, being capable of modeling more complex dynamics, provided consistent results in both datasets. However, we believe that even in cases when the performances are comparable, there is still value in our work because we demonstrate that a neural network can achieve similar performance to complex signal processing methods.

Finally, we have built a DNN to estimate fetal HR directly from the extracted fetal ECG signals. An alternative strategy would be to first extract features of the fetal HR manually and then apply a network on them for estimating the fetal HR. Although we believe that the encoder of our network can extract informative HR features, it would be interesting to investigate also this option.

4.1. Limitations of the study

This study has several limitations. First, the proposed fetal HR extraction network has many parameters such as the number of layers, dilation rates, number of nodes for each layer, type of layers, size of input and output signals etc. To select these parameters, first we divided our private dataset into training, validation and test sets. Then, we chose the parameter values that led to the best performance on the validation set. An alternative way to choose the parameters is leave-on-out cross validation that is appropriate in cases of small datasets like ours. The reason we did not chose this approach is that it is computationally very expensive. Moreover, due to the large number of parameters we did not perform sensitivity analysis by varying their values (apart from the kernel size), although it could potentially reveal erratic behavior resulting from overfitting to the training set. Our training set is relatively small for training such a complex model. The training recordings are long, leading to many available data (54 h), but they come from 16 subjects and a single acquisition system. However, our method was successfully tested not only on a test set from the same database but also to a completely unseen set from Physionet. The Physionet data constitute heterogeneous data obtained from multiple sources with different acquisition systems. It is promising that, even though our network was trained on data obtained from a single

system, it was able to generalize to the Physionet dataset obtained by different systems. Nevertheless, in order to confirm our findings, validation on more diverse datasets is needed.

According to our results, as already mentioned, the DICNN-LSTM network sometimes failed to correctly estimate the fetal HR during the second stage of labor. The second stage of labor is the phase of labor that begins after full cervical dilatation and ends with the delivery of the baby. This phase lasts approximately 20–60 min. During this stage uterine contractions are stronger and more frequent, and the maternal abdominal muscles are particularly active due to intense pushing. The second stage is very critical for the fetus because exactly at this period it is often subjected to reduced oxygenation due to the increase in the intensity of the contractions. During this stage fetal HR decelerations often happen that are synchronous to the contractions. Our network frequently failed to precisely estimate these decelerations. Fortunately, the HR reliability classifier mostly classified these results as unreliable. However, it has been demonstrated that prolongation of the second stage of labor and the fetal HR decelerations is correlated with perinatal mortality and morbidity (Tranquilli 2012). Therefore, the study of the fetal HR during this period is extremely important. Merely excluding these parts as unreliable should be avoided as much as possible. Possibly the capacity of our network is not sufficient to handle these cases of extreme noise and a separate network should be trained particularly for this stage. It might also be that we need more training data obtained during this stage, so our network learns better how to handle those segments. Finally, the error could lie much earlier in our signal processing chain, in the fetal ECG extraction step from the abdominal recordings. Considering the high amount of muscle noise, we could argue that the maternal-fetal ECG separation might not have been correctly performed. This has a strong impact on subsequent processing.

In general, regarding our fetal HR reliability classifier, we need to stress that it would benefit from further improvement. Our main purpose in this work was to show that it is possible to classify unreliable fetal HRs with a relatively simple network. According to our experiments in the heterogenous Physionet dataset we found that the classifier rejected too much data of good quality. Unlike our classifier, the algorithm of Warmerdam successfully estimated the reliability of the result in this dataset. Thus, our results suggest that more investigation is required for robust fetal HR reliability estimation. We might do need two separate classifiers, one trained for second stage of labor and one for pregnancy and first stage of labor due to the distinct characteristics of the different stages. On the other hand, one could argue that we do not need an HR reliability classifier outside the second stage of labor since the network estimations are already accurate during the first stage of labor. These hypotheses need to be confirmed with additional experiments and more data.

Our method extracts the fetal HR sampled at 4 Hz. The noninvasive fetal ECG can potentially provide beat-to-beat fetal HR information, while beat-to-beat HR variability analysis has been reported to provide important information about fetal distress (van Laar *et al* 2010). However, according to a study (Goncalves *et al* 2013), differences in variability indices between beat-to-beat and 4 Hz sampled HR signals were found to not affect physiological changes observed during labor progression, while 4 Hz sampling provided better results in entropy indices. In addition, most central monitoring systems require the fetal HR values to be communicated at a frequency of 4 Hz and thus choosing for this frequency allows our method to be easier implemented in clinical practice.

Moreover, we provided the performance of our method in terms of PPA as in Cohen *et al* (2012), Euliano *et al* (2017) and PPA_5 for assessing the fetal HR extraction reliability with lower error tolerance. For an actual fetal HR of 140 bpm, the PPA metric considers acceptable HR values in the range [126,154], while PPA_5 in the range [133,147]. We need to stress that this error of 7 bpm (accepted by PPA_5) might still be relatively high for reliable calculation of fetal HR variability.

Another important issue to consider when developing a method intended to run online as well as when comparing different algorithms is computational complexity. Our model ensemble is relatively complex having 137M trainable parameters and our reliability classifier has 237k. The advantage of our method is that it is a neural network and as such its deployment can exploit massive parallelization, which can be exploited via GPU computation. The other algorithms that we compared to might be less complex but do not have this benefit and consequently might be even slower in practice.

5. Conclusion

In this study, we presented a deep dilated inception CNN-LSTM network for fetal HR extraction from noninvasively obtained fetal ECG signals. A quality assessment method, based on a CNN network, was additionally developed to exclude signal parts that will yield an unreliable fetal HR. The proposed method achieves accurate HR detection outperforming top-performing methods proposed in the literature. Our method may be used to achieve more reliable HR monitoring and contribute to the spread of noninvasive

electrocardiography in clinical practice. Our results indicate that more complex algorithms and more data are needed to also make accurate fetal HR estimations during the second stage of labor.

Conflict of interest

RV has shares in Nemo Healthcare BV, The Netherlands.

ORCID iDs

E Fotiadou https://orcid.org/0000-0003-3877-8961

References

Adam D and Shavit D 1990 Complete foetal ECG morphology recording by synchronized adaptive filtration *Med. Biol. Eng. Comput.* **28** 287–92

Antczak K 2018 Deep recurrent neural networks for ECG signal denoising (arXiv:1807.11551)

Behar J *et al* 2014 Combining and benchmarking methods of foetal ECG extraction *Physiol. Meas.* **35** 1569–89

Behar J, Andreotti F, Zaunseder S, Oster J and Clifford G D 2016 A practical guide to non-invasive foetal electrocardiogram extraction and analysis *Physiol. Meas.* **37** R1–35

Behar J, Oster J and Clifford G D 2013 Non-invasive FECG extraction from a set of abdominal sensors *Computing in Cardiology 2013 (Zaragosa)*

Camargo-Olivares J L, Martín-Clemente R, Hornillo-Mellado S, Elena M M and Roman I 2011 The maternal abdominal ECG as input to MICA in the fetal ECG extraction problem *IEEE Signal Process. Lett.* **18** 161–4

Clifford G D, Silva I, Behar J and Moody G B 2014 Non-invasive fetal ECG analysis *Physiol. Meas.* **35** 1521–36

Cohen W R *et al* 2012 Accuracy and reliability of fetal heart rate monitoring using maternal abdominal surface electrodes: maternal surface electrode fetal monitoring *Acta Obstet. Gynecol. Scand.* **91** 1306–13

Esteva A *et al* 2019 A guide to deep learning in healthcare *Nat. Med.* **25** 24–9

Euliano T Y, Darmanjian S, Nguyen M T, Busowski J D, Euliano N and Gregg A R 2017 Monitoring fetal heart rate during labor: a comparison of three methods *J. Pregnancy* **2017** 8529816

Fotiadou E *et al* 2020 Deep convolutional long short-term memory network for fetal heart rate extraction *2020 42nd Annual Int. Conf. of the IEEE Engineering in Medicine & Biology Society (EMBC) (Montreal, QC, Canada)* pp 1–4

Fotiadou E, Konopczyński T, Hesser J and Vullings R 2020 End-to-end trained encoder–decoder convolutional neural network for fetal electrocardiogram signal denoising *Physiol. Meas.* **41** 015005

Fotiadou E and Vullings R 2020 Multi-channel fetal ECG denoising with deep convolutional neural networks *Front. Pediatr.* **8** 508

Goncalves H, Costa A, Ayres-de Campos D, Costa-Santos C, Rocha A P and Bernardes J 2013 Comparison of real beat-to-beat signals with commercially available 4 Hz sampling on the evaluation of foetal heart rate variability *Med. Biol. Eng. Comput.* **51** 665–76

He K *et al* 2015 Delving deep into rectifiers: surpassing human-level performance on ImageNet classification *IEEE Int. Conf. on Computer Vision (ICCV) (Santiago)* (https://doi.org/10.1109/ICCV.2015.123)

Holschneider M *et al* 1990 A real-time algorithm for signal analysis with the help of the wavelet transform *Wavelets* (Heidelberg: Springer) pp 286–97

Isin A and Ozdalili S 2017 Cardiac arrhythmia detection using deep learning *Proc. Comput. Sci.* **120** 268–75

Jamshidian-Tehrani F and Sameni R 2018 Fetal ECG extraction from time-varying and low-rank noninvasive maternal abdominal recordings *Phys. Meas.* **39** 125008

Jezewski J, Wrobel J and Horoba K 2006 Comparison of Doppler ultrasound and direct electrocardiography acquisition techniques for quantification of fetal heart rate variability *IEEE Trans. Biomed. Eng.* **53** 855–64

Kahankova R *et al* 2020 A review of signal processing techniques for non-invasive fetal electrocardiography *IEEE Rev. Biomed. Eng.* **13** 51–73

Kanjilal P P, Palit S and Saha G 1997 Fetal ECG extraction from single-channel maternal ECG using singular value decomposition *IEEE Trans. Biomed. Eng.* **44** 51–9

Kingma D and Ba J 2015 Adam: a method for stochastic optimization *Int. Conf. for Learning Representations (San Diego)*

Lee J S, Seo M, Kim S W and Choi M 2018 Fetal QRS detection based on convolutional neural networks in noninvasive fetal electrocardiogram *4th Int. Conf.on Frontiers of Signal Processing, Poitiers* (https://doi.org/10.1109/ICFSP.2018.8552074)

Lempersz C *et al* 2020 Intrapartum non-invasive electrophysiological monitoring: a prospective observational study *Acta Obstet. Gynecol. Scand.* **99** 1387–95

Martín-Clemente R, Camargo-Olivares J L, Hornillo-Mellado S, Elena M and Roman I 2002 Fast technique for noninvasive fetal ECG extraction *IEEE Trans. Biomed. Eng.* **58** 227–30

Muduli P R, Gunukula R R and Mukherjee A 2016 A deep learning approach to fetal-ECG signal reconstruction *2016 22nd National Conf. on Communication (NCC) (Guwahati)* (https://doi.org/10.1109/NCC.2016.7561206)

Nassif A B *et al* 2019 Speech recognition using deep neural networks: a systematic review *IEEE Access* **7** 19143–65

Sameni R 2008 Extraction of fetal cardiac signals from an array of maternal abdominal recordings *PhD Thesis* Sharif University of Technology—Institut National Polytechnique de Grenoble

Sameni R and Clifford G D 2010 A review of fetal ECG signal processing; issues and promising directions *Open Pacing Electrophysiol. Ther. J.* **3** 4–20

Shaw C J, Lees C C and Giussani D A 2016 Variations on fetal heart rate variability *J. Physiol.* **594** 1279–80

Sherstinsky A 2020 Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network *Physica* D **404** 132306

Shi W *et al* 2017 Single image super-resolution with dilated convolution based multi-scale information learning inception module *IEEE Int. Conf. on Image Processing (ICIP) (Beijing, China)* (https://doi.org/10.1109/ICIP.2017.8296427)

Silva I, Behar J, Sameni R, Zhu T, Oster J, Clifford G D and Moody G B 2013 Noninvasive fetal ECG: the PhysioNet/Computing in Cardiology Challenge *Comput. Cardiol.* **40** 149–52

Sundström A K *et al* 2000 *Fetal Surveillance* (Gothenburg, Sweden: Neoventa Medical AB)

Szegedy C *et al* 2015 *Going deeper with convolutions IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* pp 1–9
Tranquilli A L 2012 Fetal heart rate in the second stage of labor: recording, reading, interpreting and acting *J. Matern. Fetal Neonatal. Med.* **25** 2551–4
Ungureanu M, Bergmans J W M, Oei S G and Strungaru R 2007 Fetal ECG extraction during labor using an adaptive maternal beat subtraction technique *Biomed. Tech.* **52** 56–60
van Laar J O E H, Peters C H L, Vullings R, Houterman S, Bergmans J W M and Oei S G 2010 Fetal autonomic response to severe acidaemia during labor *BJOG* **117** 429–37
Varanini M, Tartarisco G, Billeci L, Macerata A, Pioggia G and Balocchi R 2013 A multi-step approach for non-invasive fetal ECG analysis *Computing in Cardiology 2013 (Zaragosa)*
Vigneron V *et al* 2003 Fetal electrocardiogram extraction based on non-stationary ICA and wavelet denoising *Seventh Int. Symp. on Signal Processing and Its Applications, 2003. Proc. (Paris, France)* (https://doi.org/10.1109/ISSPA.2003.1224817)
Vullings R, Peters C H L, Sluijter R J, Mischi M, Oei S G and Bergmans J W M 2009 Dynamic segmentation and linear prediction for maternal ECG removal in antenatal abdominal recordings *Physiol. Meas.* **30** 291–307
Warmerdam G J J, Vullings R, Schmitt L, Van Laar J O E H and Bergmans J W M 2018 Hierarchical probabilistic framework for fetal R-peak detection, using ECG waveform and heart rate information *IEEE Trans. Signal Process.* **66** 4388–97
Warmerdam G J J, Vullings R, Schmitt L, Laar J O E H V and Bergmans J W M 2016 A fixed-lag Kalman smoother to filter power line interference in electrocardiogram recordings *IEEE Trans. Biomed. Eng.* **64** 1852–61
Waseem R and Zenghui W 2017 Deep convolutional neural networks for image classification: a comprehensive review *Neural Comput.* **29** 1–98
Widrow B, Glover J R, McCool J M, Kaunitz J, Williams C S, Hearn R H, Zeidler J R, Dong E and Goodlin R C 1975 Adaptive noise cancelling: principles and applications *Proc. IEEE* **63** 1692–716
Wu S *et al* 2013 Research of fetal ECG extraction using wavelet analysis and adaptive filtering *Comput. Biol. Med.* **43** 1622–7
Xiong P, Wang H, Liu M, Zhou S, Hou Z and Liu X 2016 ECG signal enhancement based on improved denoising auto-encoder *Eng. Appl. Artif. Intell.* **52** 194–202
Ye Y, Sheu P C Y, Zeng J, Wang G and Lu K 2009 An efficient semi-blind source extraction algorithm and its applications to biomedical signal extraction *Sci. China Ser.* F **52** 1863–74
Yu F and Koltun V 2016 Multi-scale context aggregation by dilated convolutions *ICLR*
Zhang N, Zhang J, Li H, Mumini O, Samuel O, Ivanov K and Wang L 2017 A novel technique for fetal ECG extraction using single-channel abdominal recording *Sensors* **17** 457
Zhao Z *et al* 2019 DeepFHR: intelligent prediction of fetal acidemia using fetal heart rate signals based on convolutional neural network *BMC Med. Inform. Decis. Mak.* **19** 286
Zhong W, Liao L, Guo X and Wang G 2018 A deep learning approach for fetal QRS complex detection *Physiol. Meas.* **39** 045004