End-to-end learning with interpretation on electrohysterography data to predict preterm birth

Computers in Biology and Medicine 158 (2023) 106846

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

End-to-end learning with interpretation on electrohysterography data to predict preterm birth

A.M. Fischer [a,b,*], A.L. Rietveld [b,c], P.W. Teunissen [d,e], P.C.A.M. Bakker [b,c], M. Hoogendoorn [a]

[a] Department of Computer Science, Vrije Universiteit, De Boelelaan 1105, Amsterdam, 1081 HV, The Netherlands
[b] Department of Obstetrics and Gynecology, Amsterdam UMC Location AMC, Meibergdreef 9, Amsterdam, 1105 AZ, The Netherlands
[c] Amsterdam Reproduction and Development, Amsterdam, The Netherlands
[d] Department of Gynaecology&Obstetrics, Maastricht UMC, P. Debyelaan 25, Maastricht, 6229 HX, The Netherlands
[e] School of Health Professions Education, Faculty of Health Medicine and Life Sciences, Maastricht University, Universiteitssingel 60, Maastricht, 6229 ER, The Netherlands

ARTICLE INFO

Keywords:
Preterm birth prediction
Electrohysterography
Machine learning
Deep learning
Explainable AI
Interpretability framework

ABSTRACT

Prediction of preterm birth is a difficult task for clinicians. By examining an electrohysterogram, electrical activity of the uterus that can lead to preterm birth can be detected. Since signals associated with uterine activity are difficult to interpret for clinicians without a background in signal processing, machine learning may be a viable solution. We are the first to employ Deep Learning models, a long-short term memory and temporal convolutional network model, on electrohysterography data using the Term–Preterm Electrohysterogram database. We show that end-to-end learning achieves an AUC score of 0.58, which is comparable to machine learning models that use handcrafted features. Moreover, we evaluate the effect of adding clinical data to the model and conclude that adding the available clinical data to electrohysterography data does not result in a gain in performance. Also, we propose an interpretability framework for time series classification that is well-suited to use in case of limited data, as opposed to existing methods that require large amounts of data. Clinicians with extensive work experience as gynaecologist used our framework to provide insights on how to link our results to clinical practice and stress that in order to decrease the number of false positives, a dataset with patients at high risk of preterm birth should be collected. All code is made publicly available.

1. Introduction

The diagnosis of premature labor, i.e., labor before the 37th week of pregnancy, and its effective prevention are challenges faced by obstetricians on a daily basis. In 2014 the World Health Organization (WHO) estimated the rate of preterm births worldwide around 10%. As premature birth is the leading cause of mortality and neonatal morbidity [1], clinicians try to minimize the negative effects of preterm birth. When a pregnant woman presents herself while having premature contractions, clinicians attempt to accurately assess the risk of her going into actual preterm labor. If assessed to be at risk, hospital admission often combined with pharmacological treatment follows to prepare the baby for a preterm birth and offer it the best possible start given the circumstances. Nowadays, clinicians try to estimate the risk of preterm labor using a set of clinical tools such as ultrasound, laboratory tests and the tocogram, a tool to quantify the number of contractions per time unit. This assessment is flawed and results in many false positives, i.e., more than 50% of patients hospitalized for imminent preterm labor deliver at term [2]. It is becoming increasingly evident that admission to the hospital for potential preterm labor accounts for a huge burden on society, not only because of the costs of hospital admission but also because of the emotional and physical impact on the patient's wellbeing. To decrease the frequency and impact of false positive preterm labor estimations, and to improve ability to detect those who are going to deliver preterm, we need to reach for new techniques to increase accuracy of the current estimation of imminent preterm birth.

Uterine contractions are always present during pregnancy and during different phases of pregnancy they differ in frequency, strength, amplitude and propagation [3]. Therefore, profound knowledge of the process of uterine contractions and the ability to diagnose abnormalities that may lead to preterm birth are important. Two clinical tools that monitor the uterine contractions are the external tocometer and the intrauterine pressure catheter (IPC). Although the IPC is the only instrument that can accurately measure the strength of uterine contractions, the tocometer is the de facto instrument used, as the IPC requires rupture of the membranes for placement and carries

* Corresponding author at: Department of Computer Science, Vrije Universiteit, De Boelelaan 1105, Amsterdam, 1081 HV, The Netherlands.
E-mail address: a.m.fischer@amsterdamumc.nl (A.M. Fischer).

https://doi.org/10.1016/j.compbiomed.2023.106846
Received 27 December 2022; Received in revised form 3 March 2023; Accepted 30 March 2023
Available online 31 March 2023

potential risks for mother and child [4]. The tocometer also measures frequency of contractions, however, it has its restrictions due to the inter- and intra-variability in interpretation between clinicians [5] and the poor quality of the signal for obese women [6,7]. Furthermore, the tocometer is incapable of measuring the electrical activity of the uterus. Consequently, the tocometer cannot provide any information about the underlying physiological process in the uterus, such as whether the entire uterus contracts or only specific parts of the uterus.

A promising technology, called the electrohysterogram (EHG), is available that allows for more detailed analysis of uterine contractions. The EHG represents the signal associated with action potentials propagating from the uterus to the abdomen via smooth muscle cells, which can be thought of as the electrical activity of the uterus [8].

Finding physiological markers in EHG that are discriminative for predicting preterm and term birth has been widely studied in the past years. This includes handcrafted features such as velocity, directionality and synchronization [9–22], and these parameters have proved to be useful predictors in combination with machine learning models [20,22–25]. While this has led to valuable information, designing handcrafted features from EHG data does lead to loss of information. In recent years, there has been a shift from feature engineering to end-to-end learning; namely Deep Learning (DL) [26], which has been applied in many medically related research areas, including clinical imaging [27]. Nevertheless, to the best of our knowledge, these models have not been used to predict preterm birth using EHG data as time series. To address this gap, we employ two deep learning models in this study, namely Long-Short Term Memory (LSTM) networks [28] and Temporal Convolutional Neural (TCN) networks [29], to explore the potential benefit of end-to-end learning on EHG time series data.

However, to make these models applicable in clinical settings we need to innovate, because a major drawback of these DL models is the limited understanding of what factors contribute to a prediction. Recent work on interpretability for DL time series models has focused on attention mechanisms [30–34], which requires an additional layer in the network to obtain an attention vector that contains importance weights of different parts of the time series. Attention mechanisms, however, add additional hyperparameters to tune for the network, require large amounts of data to be trained on and have proven to fail to accurately identify feature importance over time in multivariate time series data [35,36]. Similarly, the Transformer Model [37], which removes all convolutional and recurrent layers and uses only attention mechanisms to simplify the network, adds an additional level of complexity by incorporating position encodings to preserve some information about the order of the time series.

We bypass the need for using attention mechanisms to highlight significant parts of time series, by introducing an interpretability framework that segments time series into subsequences and produces a prediction over each subsequence. In effect, we highlight parts of the time series a clinician should pay further attention to and we have clinicians evaluate our findings.

Another contribution of this research is combining high sampled frequency EHG data and clinical data into a single DL model. While many relevant clinical data is collected in electronic health records, the potential benefit of using both EHG time series data and clinical data for preterm birth prediction has not been studied using these models. As there are ample risk factors associated with preterm birth, such as diabetes, hypertension, placental abnormalities, multiple gestation, previous preterm delivery, age, weight and cervical length [38–43], this could potentially lead to better predictive power. In summary, the main contributions of this research are:

- Combine temporal EHG data and static clinical data for Deep Learning models
- Provide an interpretability framework for Deep Learning models for time series classification on small datasets
- Evaluate these DL models on EHG data for the first time by clinicians

2. Related work

This sections covers two main aspects. First, related work on predicting preterm birth using EHG data is presented. Second, literature related to machine learning explainability in the context of time series classification is outlined.

2.1. Preterm birth prediction using EHG data

Many studies have been published using the public available Term–Preterm Electrohysterogram database (TPEHG) and show almost perfect scores when differentiating preterm from term patients. However, recent work has shown that these results are based on a methodological flaw, namely applying oversampling on the dataset *before* partitioning the data into a separate train and test set [44], which leads to data leakage. When oversampling was carried out correctly, the results where often not better than random guessing [44].

The TPEHG database consists of 300 EHG records (belonging to 300 unique patients), and is highly imbalanced as there are 38 preterm cases and 262 term cases. We will focus on studies that performed data oversampling on this dataset and these studies did not explicitly mention partitioning the dataset in a mutually exclusive train and test set before oversampling. All studies used handcrafted features and machine learning and the results reported by these studies on the oversampled data had AUC scores even up to 0.99. Many of these studies have been reproduced by van de Wiele et al. [44]. In Table 1 we show the results from both the original paper and the reproduced results from van de Wiele et al. [44].

When looking at Table 1, the gap between reported AUC of the original study and correctly oversampled AUC of the reproduced study is compelling. The highest AUC, when correctly oversampled, was 0.65 and this study used median frequency as handcrafted feature [10]. Of studies [23,45,46], also AUC on the original TPEHG dataset was reported, which yielded a highest AUC of 0.62.

Also more recent work used the TPEHG database and calculated centroid frequency as feature from the EHG signal. Degbedzui et al. [47] report an accuracy of 99%, however, also they applied oversampling on the dataset and therefore we believe these results should be treated with caution.

Few studies did not apply oversampling on the TPEHG dataset [48–50] and used machine learning to classify preterm/term patients. Ryu et al. [48] filtered EHG signals using Multivariate Empirical Mode Decomposition (MEMD) instead of Fourier transform and hereafter they extracted sample entropy as feature. They bypassed the imbalanced data problem by subsampling a 100 times a balanced dataset of 38 term and 38 preterm records from the original dataset. A maximum AUC of 0.60 was achieved using their set-up.

Two studies segmented the original EHG signals into sub segments, the work of Khalil et al. [51] did not focus on classifying preterm/term birth but used these segments to identify four types of events, namely contractions, fetus motions, Alvarez waves and long-duration low-frequency waves and had them validated by an expert. The work of Shardad et al. [52] uses another approach to separate EHG signals and uses Linear Predictive Coding to extract features from these segments. Hereafter they cluster these events and classify each event independently into term and preterm birth, but do not consider patterns over each entire EHG recording. To bypass the problem of unbalanced data, they perform undersampling, resulting in a scenario that does not resemble reality for this application.

The work of Janjarasjitt et al. [49] uses single wavelet-based features to predict preterm birth and they evaluate their feature in a leave-one-out cross-validation scheme. First they split the dataset into two groups, of which one group had their EHG recording early during their pregnancy (around 22nd week of gestation) and the other group had their EHG recording later during pregnancy (around 32nd week of gestation). The classifier on the early group achieves a sensitivity and

Table 1
Overview of studies and their results on the TPEHG database.

	Original study AUC[a]		Reproduced study AUC[b]	
	original data	Oversampled	Correctly oversampled	Incorrectly oversampled
[10]	–	0.99	0.65	0.99
[20]	–	–	0.59	–
[22]	–	0.99	0.57	0.99
[23]	0.61	0.95	0.61	0.97
[53]	–	0.99	0.58	0.96
[54]	–	–	0.54	–
[45]	0.62	0.94	0.52	0.96
[46]	0.58	0.94	0.52	0.96
[55]	–	–	0.56	–
[56]	–	–	0.57	–
[57]	–	0.88	0.55	0.92

[a]We report the highest AUC scores (rounded to two decimals) from the original study. A — indicates that the AUC score was not provided (either missing or only accuracy was reported) in the original paper.
[b]We report the highest AUC scores (rounded to two decimals) from [44].

specificity of 0.6842 and 0.7133 respectively. Side note is that using a leave-one-out scheme may lead to too optimistic results. Lastly, Sadi-Ahmed et al. [50], only reported an accuracy of 0.89 while no other metrics were reported, which impedes assessing the true value of the model, since always predicting term birth already yields an accuracy of 0.86. They used features such as total number of contractions and average duration of contractions.

Our work will deliberately not apply oversampling on the imbalanced TPEHG dataset and is the first method to apply Deep Learning on this highly imbalanced dataset. Since Deep Learning approaches can also have problems in dealing with unbalanced data, we propose cost-sensitive learning by means of adding a class weight to the loss function. In order to make a fair comparison to results of existing studies, we will only compare to correctly oversampled datasets (third column in Table 1) and studies that did not apply oversampling.

2.2. DL explainability and time series

While explainability for deep learning models on static data is widely discussed in literature, explainability on time series domain has received much less attention. Broadly there are two classes to explain model behavior in time series. The first is instance-level feature importance from supervised learning on static data that is alternated in such way to make it suitable for time series data. Either the explanation is backpropagation based, in which gradients throughout the network (or in specific layers) are computed with respect to the input [31,58–60] or perturbation methods where parts of the input are masked or alternated to calculate the effect on the outcome [35,58,61]. However, gradient based methods do pick up on the important features, but fail to identify the important time steps [62]. Also perturbation methods are ambiguous to utilize, as observations need to be replaced with new samples and we do not know the data distribution of EHG data a priori.

An extension to instance-level feature importance on multivariate time series data is proposed by Tonekaboni et al. [62], in which they propose a framework that quantifies the importance of observations over time. They use generative models to learn the underlying distribution of time series and then approximate the counterfactual effect of subset of observations over time. The counterfactual is contrasted against the predictive distribution to quantify the contribution of observations of a time series. However, they only estimate the contribution of a *single* time point, whereas for high-frequency data such as EHG, we expect individual time points to contribute little, since individual time points capture only a fraction of a second, but their contribution can only be captured over time within a segment.

The second class consists of attention models, which is mostly used in the domain of natural language processing [37,63], but also have been used within the healthcare domain [30,32–34,64]. However, attention models are only well-suited in cases of large amounts of data, and fail to reliably identify important parts of time series if data consists of multi-variate time series [35].

We create an interpretability framework that is well-suited in case of limited data and as domain knowledge about EHG signals is limited, we will make a first attempt to let clinicians interact with the outcome of a ML model on EHG data. For this purpose we segment EHG time series into subsequences and let the ML model make a prediction over each subsequence. As a result, important parts of the time series are highlighted and we have clinicians assess the output.

3. Material and methods

3.1. Dataset

The data used to train the LSTM and TCN model comes from the Term–Preterm EHG Database (TPEHG DB) [14], which is publicly available on PhysioNet [65]. This dataset contains 300 EHG recordings from 300 patients, meaning that one recording per patient was made. The recordings were made either around the 22nd week of gestation or around the 32nd week of gestation during regular check-ups and last for 30 min. Only records from pregnancies where the onset of labor was spontaneous are included and cesarean sections are excluded.

We followed the definition of the WHO of 37 weeks to distinguish between preterm and term birth and also made a distinction between recordings made before 30 weeks gestation and recordings made after 30 weeks gestation. Since uterine activity (UA) increases significantly after 30 weeks of gestation [66], more UA is expected to be visible on the electrohysterogram and might result in different model behavior. In Fig. 2(a), the distribution of patients over the different categories is shown. The preterm cases are the minority class, with 38 patients having a preterm delivery against 262 patients with a term delivery.

Besides time of recording, the interval between time of recording and date of delivery is also important, as the closer to date of delivery, the more UA can be expected [66]. Since the recordings were made around the 22nd or 32nd week of pregnancy, we also see two clear distributions of time (in weeks) between time of EHG recording and birth, as shown in Fig. 2(b). The average time-to-delivery (TTD) from time of recording was 12.2 weeks, with the group that had a recording around the 32nd week of gestation having a TTD of 8–9 weeks. On the other hand, there is the group who had a recording around the 22nd week of pregnancy, and for the majority, their TTD is between 17 and 18 weeks.

Electrode set-up

Data was collected using four electrodes and placement of the electrodes was identical for all recordings and. The set-up is shown in Fig. 1. Three channels were constructed from each recording by calculating the differences in electrical potentials of the electrodes:

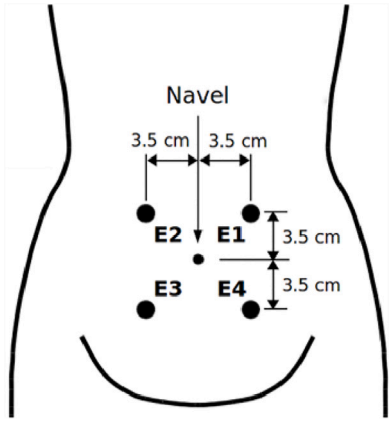


Fig. 1. Placement of the electrodes of the TPEHG DB. *Source:* [10].

Table 2
Clinical data variables present in TPEHG DB.

Variable	Variable
Abortions	Hypertension
Age	Parity
Bleeding first semester	Placental position
Bleeding second semester	Smoker
Diabetes	Weight (at time of recording)
Funneling[a]	

[a]Funneling represents the dilatation of the internal part of the cervical canal and is clinically observed through ultrasound.

- Channel 1 = E2 - E1
- Channel 2 = E2 - E3
- Channel 3 = E4 - E3

Each record is a single continuous acquisition of around 30 min in duration and the signals have been digitized at 20 samples per second per channel with 16-bit resolution over a range of ±2.5 millivolts. As a result, each record consists of about 20*60*30 = 36.000 data points and hence we are dealing with long time sequences.

An example of an unfiltered EHG recording is presented in Fig. 3. This patient gave birth at 38.5 weeks (therefore term birth) and the EHG recording was made at 31 weeks gestation. There are some noticeable differences between the channels, for example, the amplitudes of channel 3 are smaller compared to channel 1 and 2. Also there is a clear spike around minute 17 visible in channel 1 and 2, which is not visible in channel 3. This could be induced by inference of a physiological component (such as maternal heart rate or respiration rate). As some components' influence are higher in certain EHG signals than other, e.g., electrodes E1 and E2 are positioned closer to the maternal heart meaning that the heart rate will have a higher impact in those signals compared to electrodes E3 and E4 which are positioned close to the cervix [10]. How EHG data will be filtered will be discussed in Section 3.5.

Clinical data

Next to EHG records some clinical data was collected. In Table 2 an overview of all variables is shown.

Not all values are present for each patient and in Fig. 4(a) the percentage of missing values for each categorical variable is shown. The percentage of missing values ranges from 9% to 37%, with the least missing values for variables *funneling, bleeding first trimester* and *bleeding second trimester* and the highest percentage (37%) of missing values are of variables *abortions, smoker, hypertension* and *diabetes*. How missing values are handled during modeling process is discussed in Section 3.5. As for the numerical variables *age* and *weight*, the difference between the preterm and term group is negligible, as is shown in Fig. 4(b). The overall mean and standard deviation of age is 29.4 and 4.7 years respectively and the overall mean and standard deviation of weight is 69.8 and 9.8 kilograms.

3.2. LSTM architecture

In this research, a LSTM network was used as a baseline model as it is among the most popular models for time series modeling. LSTMs were first proposed in 1997 by Hochreiter and Schmidhuber [28] as a solution for the vanishing and exploding gradient problem. The main building blocks of LSTM are the input, output and forget gate, each of which have associated weights that are learned during training. Taking into account that each record consists of about 36,000 data points, we consider a specific configuration of the LSTM model that allows us to apply data reduction to the original sequence and at the same time preserve memory during processing the reduced sequence.

In our configuration, we use a so-called *stateful* LSTM, in which we first apply data reduction on the original sequence and hereafter we split up the reduced sequence into n subsequences. This amounts to processing the first subsequence of m samples (patients) in a batch and make a prediction over each first subsequence. The hidden states and cells are then retained and passed to the next batch containing the second subsequence of the same m samples and then again making a prediction, until the entire reduced sequence for m samples has been processed. Then the hidden states and cells are reset and the process is repeated for the remaining samples until all samples have been processed. In effect, all subsequences are provided as input and for each batch of subsequences the forward pass is executed, but backpropagation is only applied on the last subsequence. The advantage of a stateful LSTM is that it allows us to process shorter bits of the entire sequence at once, thus reducing the chance of forgetting long-term dependencies, while at the same time passing on long-term information throughout the sequence.

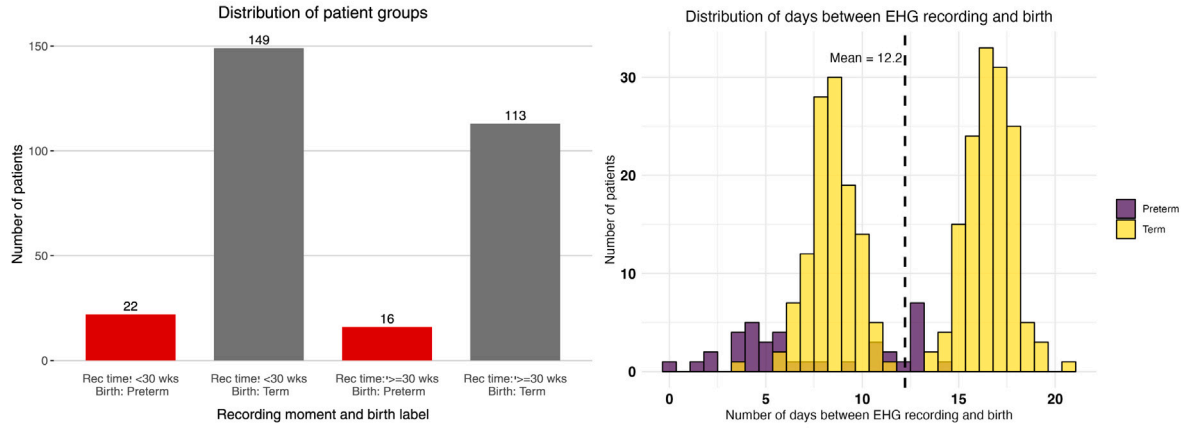
Furthermore, we initialize the value of the forget-gate bias to 1 at the beginning of training, to enhance learning long-term dependencies [67]. The process of how data reduction is realized will be explained in Section 3.5. There are many possible compositions of a LSTM model, including hidden dimension size, layer dimension size (in effect creating a stacked LSTM model), or a bidirectional LSTM. Different compositions will be tried during hyperparameter optimization, as explained in Section 3.5, and the number of trainable parameters for each network will be specified in the results section.

3.3. TCN architecture

As LSTMs can greatly suffer from vanishing or exploding gradients and thus be incapable of effectively modeling long time series tasks, Bai et al. [29] show that convolutional neural networks (CNNs) can sidestep these problems and achieve better model performance. Main advantages of the TCN are flexible receptive field size, parallelism (because convolutions over time steps can be done in parallel, unlike in LSTMs where time steps have to be processed in series) and stable gradients. The proposed architecture of a TCN adopts a 1D fully-convolutional network (FCN) [68] and uses causal convolutions, meaning that there is "no information leakage from future to past" [29].

In our research we have a 3-dimensional multivariate time series, where $X = [x_0, x_1, x_2, \ldots, x_t]$ consists of 3 different univariate time series with $X_i \in \mathbb{R}^T$. The constraint the TCN composes is that given an input sequence $x_0, x_1, \ldots, x_t$, where one wishes to predict the output y_t for some time t after the entire input sequence has been processed, one must only use the inputs that have been previously observed $(x_0, x_1, \ldots, x_t)$. This makes the structure causal.

An example of a time series processed by a TCN model is depicted in Fig. 5, where an input sequence $X = [x_0, x_1, x_2, \ldots, x_{10}]$ is given and



(a) Distribution of patient EHG recordings.

(b) Distribution of time to delivery.

Fig. 2. Data characteristics.

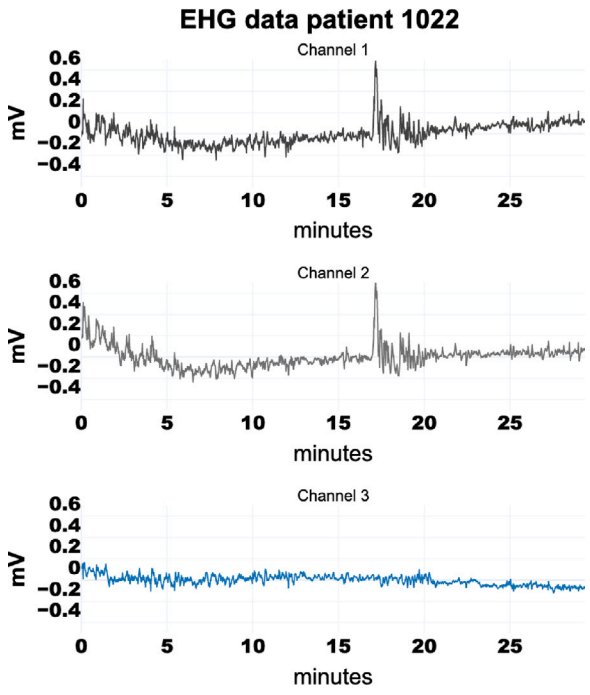


Fig. 3. Example of EHG recording of one patient.

the network flow for output y_8 is visualized. To enable a long effective history, dilated convolutions and residual blocks [69] are used. Each of these elements will be briefly explained in this section.

1D fully-convolutional network

CNNs have been widely used for image classification tasks, where an image is processed by a series of 2D convolution layers. As we are using time series data, the TCN uses 1D convolution where time series input is convolved with a filter size of $1 \times k$. The TCN can take a sequence of any length as input and produces an output sequence of the same length. This is accomplished by employing a 1D FCN where each hidden layer is kept the same length as the input layer and zero-padding is used to keep consecutive layers the same length as previous ones. For further details on FCN we refer to the work of Long et al. [68].

Dilated convolutions

In dilated convolutions one employs a convolution filter to a larger receptive field by skipping input values with step size d^i for the i^th layer. Due to the exponential dilation factor, dilated convolutions enable an exponentially large receptive field which can be beneficial for sequence tasks with a long history. If we take a 1D sequence input $X \in \mathbb{R}^T$ and filter $f : \{0, \ldots, k-1\} \rightarrow \mathbb{R}$, we can define the dilated convolution operation F on element s of the sequence as follows:

$$F(s) = (x *_d f)(s) = \sum_{i=0}^{k-1} f(i) * x_{s-d*i}$$

with d the dilation factor,

k the filter size,

and s - d . i accounts for the direction of the past.

The dilation factor d increases exponentially with the depth of the network, i.e., $d = 2^i$ for the i^th layer. As is shown in Fig. 5, this network uses a dilation factor of 1, 2 and 4 and requires only two hidden layers. If non-dilated causal convolutions were used the receptive field would have been the same, but the network would have required six hidden layers instead of two. This would lead to extra parameters to be learned and requires both more data and extensive computational resources. In short, the receptive field of the TCN can be increased by either choosing a larger filter size k and increasing the dilation factor d. The covered input of one such layer is $(k-1) * d$. The downside is that a higher dilation factor d leads to larger skipping steps in the time series input. In effect, the dependency between adjacent time steps might not be extracted at higher layers.

Residual blocks

Another architectural aspect of the TCN is the residual block instead of a convolutional layer. The residual block is used between each layer in the TCN, and contains a series of transformations (F) where the outputs are added to the input x of the block:

$$o = Activation(x + F(x)) \quad (1)$$

He et al. [69] has shown that very deep networks benefit from such architecture. Instead of learning modifications to the entire transformation, modifications to the identity mapping are learned. As explained, the TCN allows for a large receptive field to be modeled. However, this requires the network to become deeper and larger, making it more challenging to develop a stable TCN network. For example, if a time series consists of 2^{15} (32.768) time steps, a network of up to 15 layers may be required and in addition each layer consists of multiple filters to extract features. As a result, a multitude of parameters need to be learned, making it harder to develop a stable TCN. Bai et al. [29] alleviated this issue by employing a residual block, consisting of the following series of transformations:

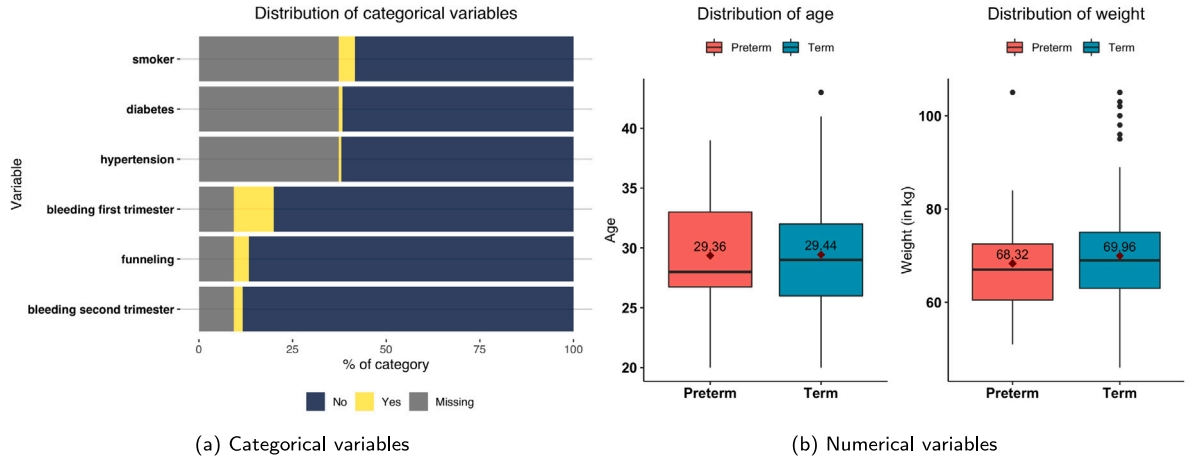


Fig. 4. Distribution of categorical and numerical variables.

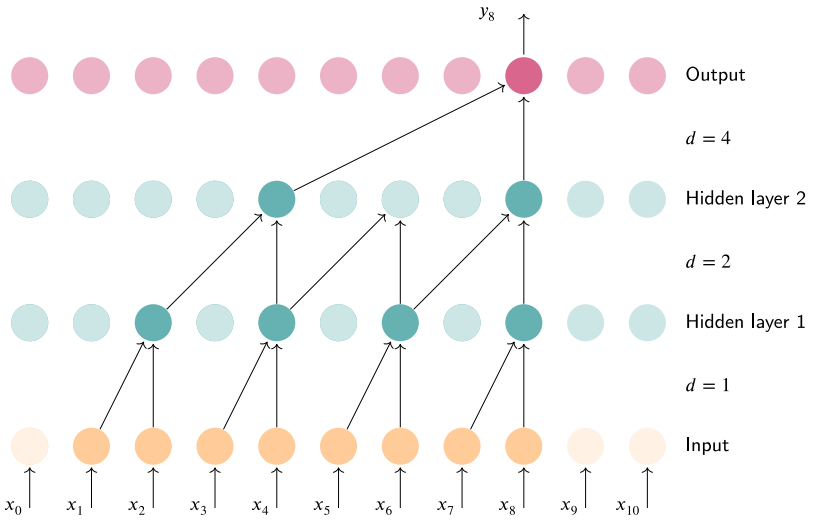


Fig. 5. Stacked dilated convolutions with dilation factors $d = 1, 2, 4$ and filter size $k = 2$.

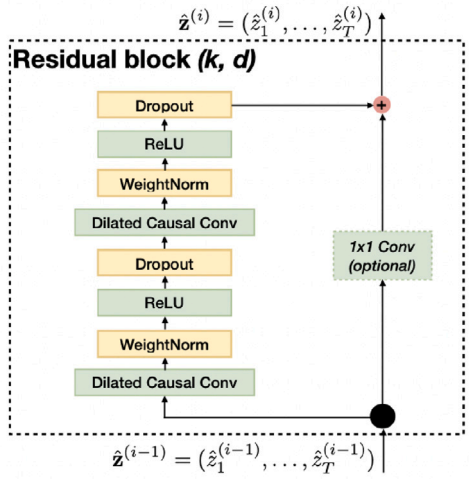


Fig. 6. Residual block (k, d) consisting of a series of transformations [29].

The series of transformations consists of two layers of dilated causal convolutions, and each layer is followed by weight normalization, non-linearity is added by applying rectified linear unit (ReLU) and at last a spatial dropout is added for regularization. The 1×1 convolution is added to ensure that the input and output have the same widths (see Fig. 6). As the residual block includes two causal layers instead of one causal layer, twice as much receptive field is added, therefore drastically reducing the minimum number of required layers necessary to achieve full coverage of the time series. Several architectures will be explored during hyperparameter optimization and the number of trainable parameters for each network will be specified in the results section.

3.4. Adding static data to LSTM and TCN model

The static data is added outside the LSTM/TCN model by means of additional fully connected layers. The static data is concatenated in a hidden layer together with the time series data that has been processed by the LSTM/TCN. Hereafter the data is processed through a linear layer, some optional layers, consisting of activation layers and/or dropout layers and finally pushed through a final linear layer. The final linear layer will output logits after which the logits will be processed by a sigmoid (output) layer to obtain probability predictions between 0 and 1. The optimal configuration of the optional layers will be determined during hyperoptimization. Finally, the data will go through the output layer. The data flow is depicted in Fig. 7.

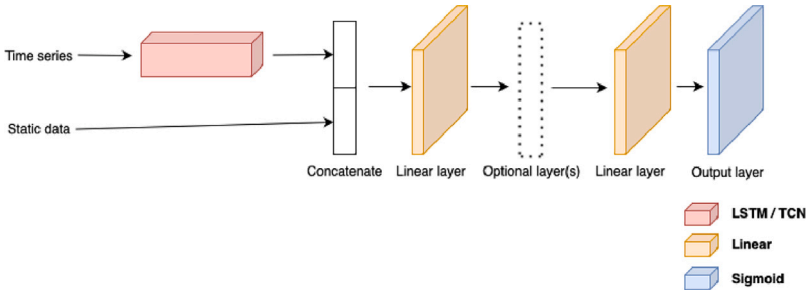


Fig. 7. Data flow of time series and static data into one model.

3.5. Classification of EHG signals and clinical data

The EHG signals will be used to predict preterm birth. This is essentially a time series classification task, where the model takes three EHG channels as input and outputs a vector indicating the whether the input signals are classified as preterm or term birth. To put it more formally, we have a 3-dimensional multivariate time series, where $X = [x_1, x_2, x_3, \ldots, x_m]$ consists of 3 different univariate time series with $X_i \in \mathbb{R}^T$.

Data preprocessing

Different noises can be present in the EHG signals, induced by either physiological interference (e.g., maternal respiration and heart rate) or non-physiological interference (motion artefacts) [70]. This has led to various different frequency bandwidths reported in literature for preterm birth prediction. Many studies have used a frequency bandwidth of 0.34–1.0 Hz, but several other studies have expanded the upper bandwidth of 1.0 Hz to 3.0 or 4.0 Hz, and even up until 16.0 Hz [10]. We decided to filter all signals with the 4th order Butterworth bandpass filter with a bandwidth of 0.34–1.0 Hz, as previous studies showed that this avoids interference from respiratory or heart rate activity [10,71,72]. After filtering the signals, the first and last 180 s of the signals are removed since these intervals contain transient effects of the filter.

Since it is computationally infeasible to process the original time series, we look for ways to reduce data while retaining most of the information present in the data and still apply the principle of end-to-end learning on EHG data. To this end, we consider features cited in literature as most predictive for preterm birth using EHG data. These are sample entropy (SE), peak frequency (PF) and median frequency (MF) [10,14,18,19,23–25]. To treat the data as a time series, retain as much information as possible, and keep the time series computationally feasible, we chose to divide each original time series into 50 non-overlapping consecutive time windows. Next, we calculate the SE, PF and MF over each time window, resulting in 50 values of SE, PF and MF. Afterwards, we bin together the values in groups of 10 values, leaving us with 5 adjacent subsequences. This process is depicted in Fig. 8 and the result of transforming the original EHG time series to a reduced time series is depicted in Fig. 9.

Sample entropy

Sample entropy is used for assessing the complexity of physiological time-series signals and is defined as follows:

$$H(x, m, r) = -\log \frac{C(m+1, r)}{C(m, r)} \tag{2}$$

Where m is the embedding dimension (= order), r is the radius of the neighborhood (0.2 * std(x)), $C(m+1, r)$ is the number of embedded vectors of length $m+1$ having a Chebyshev distance.

Peak frequency

The peak frequency represents the peak of the power distribution in the power spectral density (PSD) and is the frequency that occurs most often in the power of the signal. For each signal $x(t)$ the power spectrum P was calculated using fast discrete Fourier transform (FFT) and the PSD is defined as taking the square of the absolute value of FFT. The peak frequency is then:

$$F_{max} = \frac{f_s}{N} \max_{i=0}^{N-1} P(i) \tag{3}$$

Where f_s is the sampling frequency and N the number of samples (data points in time series).

Median frequency

The median frequency represents the midpoint of the power distribution in the PSD and is the frequency below and above which lies 50% of the total power in the signal:

$$F_{med} = \sum_{i=1}^{j} P(i) = \sum_{i=j}^{M} P(i) \tag{4}$$

Static clinical data

The dataset contains seven categorical variables ('hypertension', 'diabetes', 'placental position', 'bleeding first trimester', 'bleeding second trimester', 'funneling', and 'smoker'), each of which will be one-hot-encoded (including the missing values). The missing values of the numeric variables will be imputed with either median ('parity', 'abortions') or mean ('age', 'weight'). The last numeric variable, 'gestation at moment of recording', does not have missing values. After imputing missing values we scale all features to a mean of 0 and a standard deviation of 1. In total there are 26 static features after preprocessing, of which we remove one of the correlated feature pairs that have a correlation higher than 85%.

Experimental set-up

The TCN model will be compared to the LSTM model, predictive power of different data reduction methods will be evaluated and also the potential benefit of adding static clinical data to the model will be assessed. This amounts to 12 separate models to evaluate, which are shown in Table 3. All models are implemented using PyTorch and code is available online.[1]

Since the dataset contains only 300 records and we also want to assess generalizability, we apply nested cross validation in a stratified manner. The nested loops will be used for hyperparameter optimization and the outer loops will be used to assess model performance and generalizability. Meaning, we first divide the data in 5 (outer) folds and within each fold we create another 3 stratified folds which are used for hyperparameter optimization. Each outer fold will have their own optimal hyperparameters on which a final model will be trained and tested. Hyperparameter optimization will be performed using Bayesian Optimization [73,74]. The range of possible values for the hyperparameters for the LSTM and TCN model are shown in Tables 6 and 7

[1] https://github.com/AnneFischer/cocoon-project.

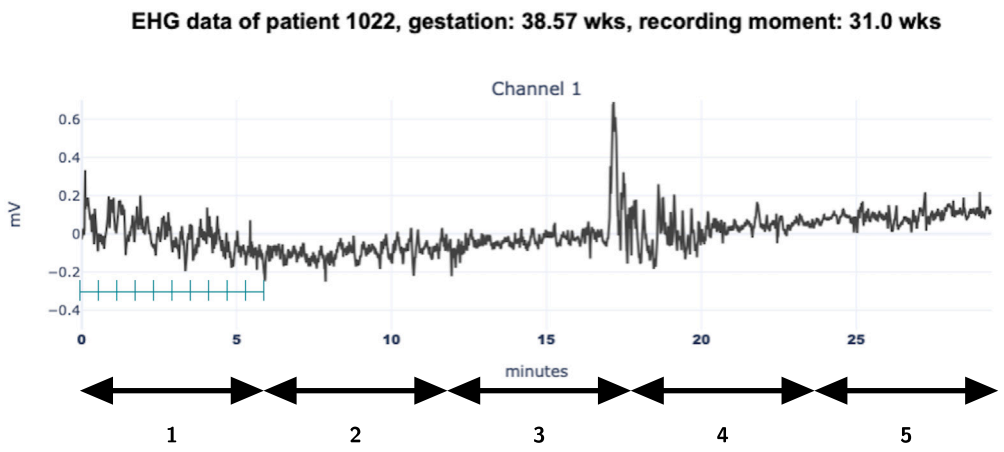


Fig. 8. Create 5 adjacent subsequences, each consisting of 10 values of sample entropy/peak frequency/median frequency.

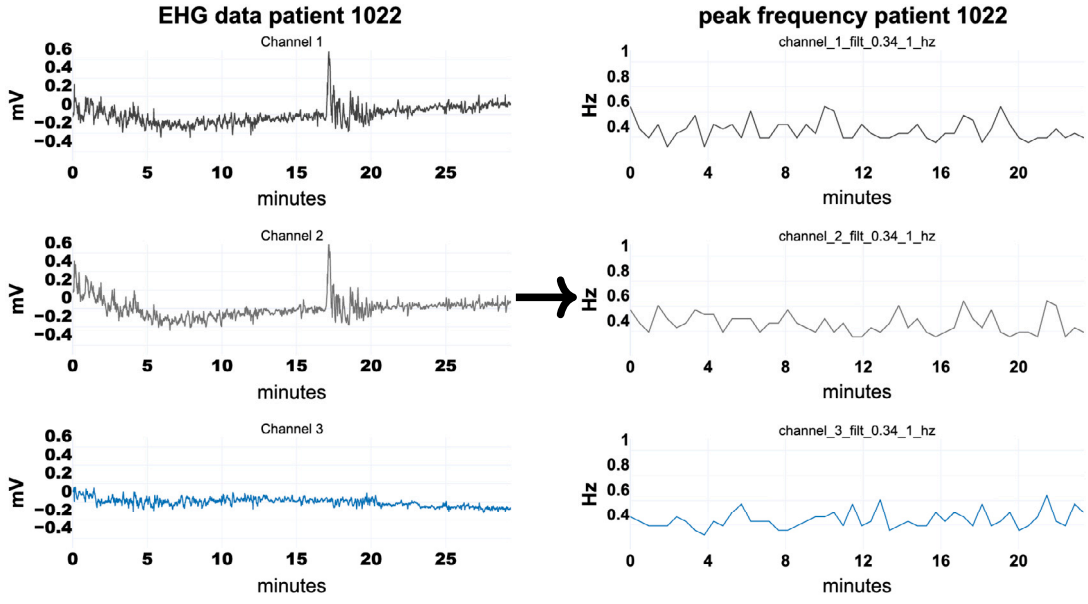


Fig. 9. Transformation of original EHG time series (left) to reduced sequence of Peak Frequency measurements (right).

Table 3
Overview of all models.

Model	Model
TCN SE time series	LSTM SE time series
TCN PF time series	LSTM PF time series
TCN MF time series	LSTM MF time series
TCN SE time series + clinical data	LSTM SE time series + clinical data
TCN PF time series + clinical data	LSTM PF time series + clinical data
TCN MF time series + clinical data	LSTM MF time series + clinical data

in Appendix. In total, 100 hyperparameter settings for each outer fold will be tested and the configuration resulting in highest AUC on nested validation set will be used to train and test the outer fold on. The number of preterm cases in each of the 5 outer test folds will be 8, 8, 8, 5 and 9 respectively.

We use binary cross-entropy with logits loss as our loss function with adding weights to positive examples to trade of precision and recall:

$$l_{n,c} = -w_{n,c}[p_c y_{n,c} * log\sigma(x_{n,c}) + 1 - y_{n,c} * log(1 - \sigma(x_{n,c}))] \quad (5)$$

Where c is the class number, n is the number of samples in the batch and p_c is the weight of the positive answer for the class c. We will use the ratio between the total number of samples in the training set and the number of preterm samples in the training set as value for p_c.

Two evaluation criteria are used to evaluate the models, namely the area under the ROC-curve (AUC) and the area under the precision–recall curve, also known as average precision (AP). As pointed out by [75,76], AP is better suited when dealing with imbalanced datasets. As we have 5 subsequences and thus 5 predictions per patient, we take both the mean probability prediction over all 5 subsequences and the highest probability prediction over all 5 subsequences for each patient. Reason for taking the highest probability is that we want to alert clinicians if a specific segment of an EHG recording triggers a higher prediction and not have segments average each other out. The mean and highest probability prediction will be used to calculate AUC and AP and this procedure will be carried out for each outer fold, resulting in 5 AUC and AP values. The final AUC and AP and their corresponding standard deviation will be calculated by taking the mean and standard deviation over all 5 folds.

3.6. Interpretability framework

The best performing model will be used to analyze subsequences from the test set to which the model assigned the top-10 highest or lowest prediction score. We will look for patterns between the two groups based on visual inspection. We chose to have clinicians analyze both the data reduced subsequences (which was what was processed by the model) and corresponding original EHG subsequences to try to translate the model's findings to clinical practice. In summary, we asked the clinicians to answer these questions:

- Do you observe any difference in patterns between the top 10s?
- The idea behind 5 predictions per EHG recording is to give more interpretability than if you would only give one prediction at the very end. Do you feel this would add anything from a clinical point of view? What factors do you feel are missing that would make the model more interpretable?

4. Results and discussion

4.1. Results

In this paragraph, we show the results based on the experiments we explained in the previous section. We will start with the results obtained using the best hyperparameters we found using Bayesian Optimization using only the EHG data. Followed by the results of using EHG data + static clinical data. Hereafter, we will discuss interpretability of the best model using our framework.

Performance LSTM and TCN model on EHG data

In Table 4 the results of each model using only EHG data are shown. The highest AUC score for LSTM model is 0.544 by using Peak Frequency as method of data reduction. Highest AP for LSTM, however, is obtained using Median Frequency as method of data reduction. None of the models outperformed other models with a significant difference. Noted is that the difference between taking mean or maximum probability over all 5 subsequences resulted in only minor differences in AUC or AP scores. We analyzed all predictions made on test set with the different LSTM models and found that all predictions were very close to the decision threshold of 0.5. In essence this means that LSTM models were not able to differentiate between different subintervals. As SE as method of data reduction resulted in AUC scores of even below 0.5, affiliated AP scores were above or around baseline (which is the incidence of preterm cases; $\frac{38}{300} = 0.13$), showing that no good comparison between AUC and AP scores is possible in case of an unbalanced dataset.

The best performing TCN model is the Peak Frequency data reduction model by taking the mean probability over all subsequences resulting in an AUC of 0.527 and AP of 0.214 if the maximum probability was taken. Also for the TCN models, the difference between taking the mean or maximum probability over all subsequences did not lead to significant differences in performance. After analyzing all predictions on the test set we found that the TCN model did differentiate between predictions on subsequences, as its predictions lay further away from the decision threshold of 0.5. The confidence interval (CI) of the best performing model of LSTM and best performing model of TCN have overlap. For TCN using SE as data reduction led to the worst performance across all models, which also holds for the LSTM models.

Performance LSTM and TCN model on EHG data + clinical data

In Table 5 the results of each model using both EHG and clinical data are shown. Also the results of a simple Logistic Regression (LR) model applied on only clinical data is shown in Table 5 to act as a baseline. The highest AUC and AP score for the LSTM model is 0.487 and 0.169 respectively, by using Peak Frequency as method of data reduction. Also for these results it holds that there is no significant difference between taking mean or maximum probability prediction over all 5 subsequences. Strikingly, adding clinical data resulted in overall worse performance for all data reduction types, except for SE, which was a similar score. Even though difference is not significant for most cases, none of the models showed an absolute gain in performance. In particular, adding clinical data to PF LSTM model, which was the best performing model with only EHG data, produced worse results.

After analyzing the results of EHG LSTM PF model against results of EHG+clinical LSTM PF model, we found that predictions made by the first model was on average a value of 0.50 while the latter had an average prediction of 0.56. Even though these averages do not differ substantially, the latter model had a standard deviation of 0.14 while EHG data PF LSTM model only had a deviation of 0.02 for its predictions. Thus, the EHG+clinical LSTM PF model on average predicts a higher value on subsequences and deviates more in its predictions. Meaning that this model appears to differentiate between different subsequences, but at the same time is unable to correctly classify subsequences to the preterm class.

As for the results for EHG+clinical TCN models, these do not change much compared to the EHG TCN models. In absolute terms, the TCN SE model with EHG+clinical data outperforms PF and MF as method for data reduction, but CIs all have overlap. In absolute terms, the TCN SE model with EHG+clinical data outperforms PF and MF as method for data reduction, but CIs all have overlap. Likewise for the EHG TCN models, the difference between taking the mean or maximum over all subsequences does not lead to significant differences in performance. When analyzing the differences in predictions between EHG TCN PF and EHG + clinical TCN SE model, we observe different behavior as for LSTM models. The mean prediction of the first model was 0.53 over all subsequences with a standard deviation of 0.14, while this was 0.51 with a deviation of 0.12 for the latter model. The average prediction is similar when clinical data is added to the model and deviations between predictions are also similar.

When we compare the results of the LR baseline model with the results of the combined models, we observe that adding EHG data to clinical data does not result in a performance gain. Compared to an AUC of 0.571, the best performing combined model (TCN SE+clinical) scores the same. For the LSTM combined models, these perform worse than the clinical baseline model.

4.2. Results interpretability framework

After consulting with our clinical experts they decided to use the EHG TCN PF model, because this model had the highest AP across all models, making it clinically most relevant if the model can correctly handle preterm cases. This model was used to evaluate the subsequences from the test sets of all folds to which the model assigned the top-10 highest or lowest prediction score. We chose to have three clinicians evaluate both the data reduced sub-sequences (which was what was processed by the model) and corresponding original EHG subsequences to try to translate the model's findings to clinical practice. Two clinicians have 10+ years of work experience as a gynaecologists in both general and academic hospital and one clinician has 8 years of work experience as resident in obstetrics and gynaecology.

Data reduced sub-sequences

In Fig. 10 (a–c) and 10 (d–f) respectively the highest and lowest predictions over the sub-sequences are shown.

Observed differences in patterns between the top 10s

All three clinicians state that, at first glance, major differences are hard to detect. All subsequences show a baseline peak frequency of around 0.4 Hz with some deviations, although deviations seem smaller for the lowest predictions. These observations are in line with expectations, since electrical activity of the uterus is of small potential,

Table 4
Overview of results of 5-fold cross-validation for LSTM/TCN models with only EHG data.

Model	AUC mean prediction	AP mean prediction	AUC max prediction	AP max prediction	# Trainable params**
LSTM SE	0.441 [0.033]	0.127 [0.017]	0.470 [0.075]	0.142 [0.039]	[446–5203]
LSTM PF	**0.544** [0.07]	0.176 [0.012]	0.507 [0.08]	0.181 [0.069]	[206–10427]
LSTM MF	0.538 [0.101]	0.160 [0.022]	0.539 [0.084]	**0.203** [0.084]	[206–607]
TCN SE	0.474 [0.109]	0.136 [0.015]	0.517 [0.079]	0.189 [0.025]	[624–2129]
TCN PF	**0.527** [0.08]	0.173 [0.058]	0.518 [0.096]	**0.210** [0.086]	[356–1161]
TCN MF	0.521 [0.100]	0.156 [0.044]	0.492 [0.095]	0.149 [0.040]	[356–1345]

Number between squared brackets is the standard deviation over 5-folds.
** This is the range of trainable parameters of the models of all 5-folds.

Table 5
Overview of results for LSTM/TCN models with EHG data + clinical data.

Model	AUC mean prediction	AP mean prediction	AUC max prediction	AP max prediction	# Trainable params[b]
LSTM SE + clinical	0.423 [0.11]	0.142 [0.03]	0.430 [0.12]	0.142 [0.03]	[1025–2411]
LSTM PF + clinical	**0.487** [0.05]	**0.169** [0.02]	0.481 [0.04]	0.167 [0.021]	[859–2463]
LSTM MF + clinical	0.480 [0.151]	0.154 [0.06]	0.483 [0.161]	0.164 [0.06]	[1758–9500]
TCN SE + clinical	**0.578** [0.089]	**0.296** [0.09]	0.562 [0.08]	0.234 [0.08]	[1961–2866]
TCN PF + clinical	0.464 [0.09]	0.186 [0.07]	0.473 [0.08]	0.152 [0.044]	[2714–3341]
TCN MF + clinical	0.528 [0.12]	0.206 [0.10]	0.549 [0.13]	0.211 [0.118]	[734–3031]
Clinical LR baseline[a]	0.571 (AUC) [0.10]		0.191 (AP) [0.02]		

[a]No hyperparameter optimization was done for this baseline. L2 regularization with C=1.0 was chosen and the same class weight as for the binary cross-entropy logits loss function was used.
[b]This is the range of trainable parameters of the models of all 5-folds.

about 50μV [10], so subtle changes in the frequency domain are to be expected and more difficult to observe with the human eye. When we calculated the standard deviation for the channels between the two top-10 groups, we observe somewhat similar variability between the two groups, except for channel 2 where the standard deviation is higher. In channels 1, 2, 3 there is a standard deviation of 0.10, 0.10, 0.09 Hz respectively for the highest predictions vs. 0.10, 0.07, 0.09 Hz for the lowest predictions. In general, the model tends to classify more false positives, as the mean prediction over all sub-sequences in test sets in the folds was 0.53.

Value of proposed interpretability framework from a clinical point of view

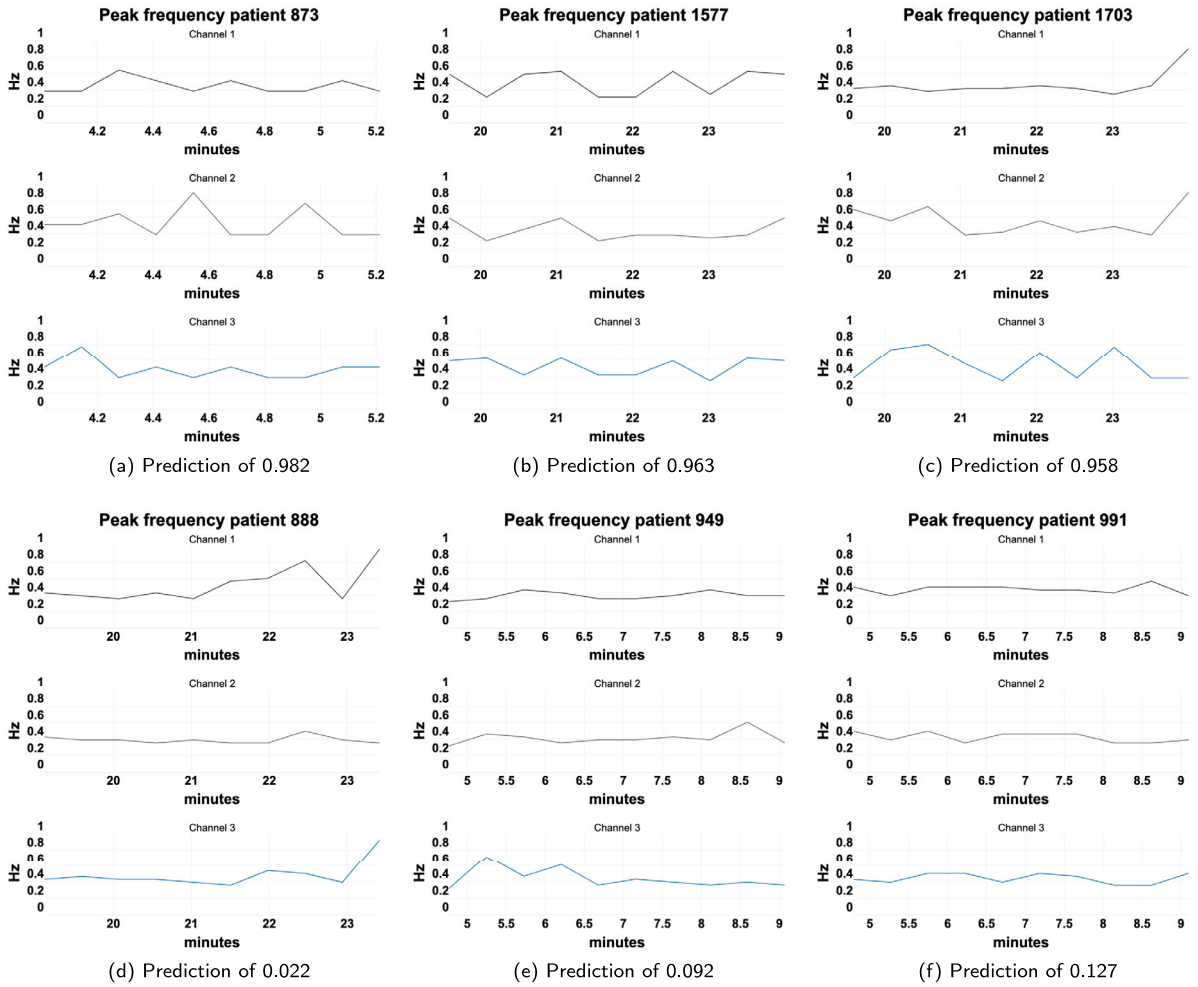
The conclusion of the clinicians on the interpretability framework can be summarized as follows:

- To make 5 predictions over an entire EHG recording can be clinically relevant if making one prediction over the entire recording takes much longer, whereas with 5 predictions over adjacent sub-sequences, you could have your first prediction sooner. However, in case the EHG recording takes only 30 min, one prediction at the end of the recording would be sufficient.
- To work with predictions over multiple intervals may be interesting in order to be able to estimate how consistently the model is close to 0 or 1 with the prediction. However, one interval now seems to equate to about 5 min. Clinically, that means that there could be 1 or 2 contractions of the uterus in that interval. If you would look at just one interval, it would not be possible to see a pattern in it that can reasonable be linked to clinical uterine activity. In other words, even if you create 5-minute intervals with a prediction per interval, to look for patterns we need to look at the entire recording (at least about 30 min) of one patient. To summarize, the value of having 5 predictions per recording, gives insight into what the model 'thinks' per interval about the chance of term or preterm delivery, but clinically speaking it is ultimately about one estimate that has to be made.

EHG sub-sequences with higher expected uterine activity

From literature it is known that between 30 to 44 weeks of gestation there is a significant increase in uterine activity (UA), and this progressive increase in UA has been reported before 36 weeks of gestation for patients destined to deliver preterm [66]. Since the average moment of recording was 26 weeks of gestation for patients in the TPEHG database, it is less likely that signals from the electrohysterogram related to UA are present. We attempt to connect the model's output to a scenario in which more UA is to be expected and clinician's expertise may be needed to assess the UA. Therefore we showcase examples of subsequences of original EHG data from test set belonging to patients who had a time-to-delivery of 6 weeks or less from the moment of EHG recording.

In Fig. 11 (a–c), the highest predictions are shown and in Figs. 11 d–f the lowest predictions for this subgroup of patients are shown. Again, differences are subtle and difficult to identify but the amplitudes of



(a) Prediction of 0.982 (b) Prediction of 0.963 (c) Prediction of 0.958

(d) Prediction of 0.022 (e) Prediction of 0.092 (f) Prediction of 0.127

Fig. 10. Peak frequency during intervals with the highest (a–c) and lowest prediction (d–f) by TCN PF model.

the EHG subsequence with the highest predictions appear to be slightly larger than those of the lowest predictions and UA also appears to be more irregular to some extent. This claim is hard to substantiate as only 6 examples are depicted, but when comparing more recordings after 30 weeks of gestation, more UA is visible on the EHG compared to patients in whom the recording was made before 30 weeks of gestation. Also in Fig. 11e (EHG recording was made at 26 weeks of gestation), it can be seen that UA is smaller compared to Fig. 11 a–d and f.

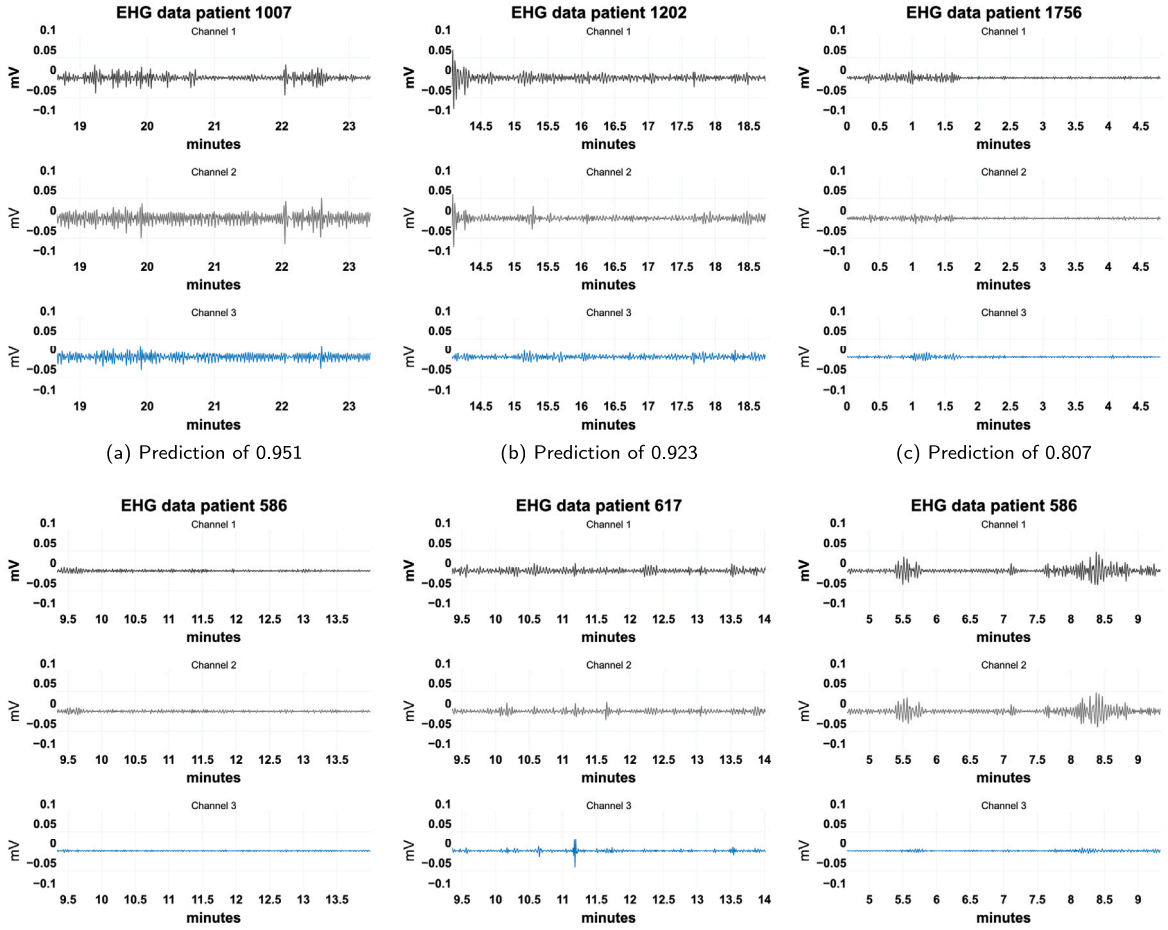
5. Discussion

The best performing model, EHG + clinical TCN with Sample Entropy for data reduction, with an AUC of 0.578 and AP of 0.296 has similar predictive power as other ML models developed on the original (not oversampled) TPEHG database. We deliberately did not apply oversampling on the TPEHG database, as recent research has shown that when oversampling was applied properly (i.e., oversampling after data partitioning), these models often did not perform better than random guessing [44]. The best performing study, which has been reproduced by van de Wiele et al. [44] and when correctly oversampled, had an AUC of 0.65.

Papers that did also report evaluation metrics on the original (not oversampled) TPEHG database achieved an AUC of 0.615 using a Random Forrest model [45], an AUC of 0.58 using a feed-forward Neural Network [46], an AUC of 0.61 using support vector machine [23] and an AUC of 0.60 [48] using a linear classifier. All these models used handcrafted engineered features as input. Although an AUC score of 0.578 when using a TCN model is not close to a perfect score, it is obtained in a realistic scenario where incidence is low and thus having a class imbalance. The use of DL models on this small dataset works as good as models that rely on specific feature engineering and have the advantage of not being dependent on scenarios where domain knowledge may be limited.

Clinicians who evaluated results of our interpretability framework stressed that differences between predictions are hard to detect with the human eye and as neither EHG nor frequency values are part of daily clinical practice, interpretation is not straightforward. In general, having a framework that provides predictions over subsequences gives them some insight on how the model 'thinks' per subsequence what the chance is of preterm or term birth. But as an interval of 5 min could reasonably only be linked to 1 or 2 uterine contractions, more intervals are needed to observe patterns. Ultimately clinicians have to translate multiple predictions into one risk estimate.

Combining the available static clinical data to EHG data and train a model led in most cases to similar performance when a model was trained on only EHG data. This implies that adding the available clinical data does not bring extra predictive power to the model or the number of samples in the dataset are too limited to learn extra trainable parameters. From our experiments, we can conclude that with the given data, the combination of EHG and clinical data does not



(a) Prediction of 0.951 (b) Prediction of 0.923 (c) Prediction of 0.807

(d) Prediction of 0.230 (e) Prediction of 0.270 (f) Prediction of 0.283

Fig. 11. EHG during sub-sequences from patients with a time-to-delivery ≤ 6 weeks with predictions by TCN PF model.

work well in one model and could be modeled separately. In clinical practice this might in fact be the favored scenario, as clinical data and related risk factors are well understood by clinicians and can therefore be assessed independently. While EHG monitoring is not yet part of standard clinical practice and therefore not many clinicians have acquired expertise in assessing an electrohysterogram, a ML model may be of more added value. In addition, a baseline LR model using only the available static clinical data resulted in an AUC score of 0.57, showing that the potential predictive power was limited a priori. Nevertheless, since there are clinical factors that have predictive power regarding spontaneous preterm birth, such as short cervical length and increased cervical-vaginal fetal fibronectin concentration [43], we believe that combining EHG and clinical data in a single ML model is worth considering if these clinical factors are available.

Drawback of the proposed TCN model is the relatively large number of false positives it infers, and as more than 50% of hospitalized patients for imminent premature labor deliver at term [2], overtreatment is likely to occur when a model trained on this patient population would be used in practice.

6. Conclusion and future work

In this research, we have employed end-to-end learning on EHG data to predict preterm birth for pregnant women, added static clinical data to assess potential increase in predictive power and provided an interpretability framework that can be used in scenario's where data is scarce and let clinicians evaluate findings of the model. In order to make high-sampled frequency data like EHG feasible for time series DL modeling, we applied data reduction by means of calculating feature values over adjacent subsequences. In effect we drastically reduced computation time while DL models still automatically learn representations of the EHG signals. We show that DL models achieve comparable performance to ML models with handcrafted features for preterm birth prediction.

Next, we assessed the potential benefit of combining EHG and clinical data into a single model and found that aggregation of the two data sources does not lead to a gain in performance. Also, we showed intervals that were given a high or low prediction by the model to clinicians, and conclude that although predictions at successive time points provide some insight into how the model 'judges' over time, clinically you would need only one definitive risk estimate of preterm birth.

Various research directions can be promising to develop a model for more accurate preterm birth prediction. First, in line with previous research [44], we underpin the importance to collect a dataset containing a more balanced fraction of high-risk patients who are more likely to deliver preterm to develop a clinically relevant model. Second, further research endeavours are needed to combine EHG and clinical data in a way that it adds value to a model. A method that can determine at the individual level which single or combination of clinical fixed variables are most important for a patient, and then add only these variables to a DL model, could be promising.

Table 6
Range or options for hyperparameters for LSTM model.

Variable	Range/options
Bidirectional	Yes/No
Hidden dimension for time series	[5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
Hidden dimension for static data[a]	[15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25]
Layer dimension	[1, 2, 3]
Learning rate	Loguniform distribution over [1e−5, 1e-3]
Number of epochs	3 (for EHG model), 6 (for EHG + clinical)
Drop out	Uniform distribution over [0.1, 0.5]
Batch size	[10, 20, 30, 40, 50, 60]
Optimizer	Adam
Optional model	Combination of [BatchNorm, activation layer, Linear layer or None]

[a]Hidden dimension of the linear layer (as depicted in Fig. 7) for combined time series and static data will be: hidden dim time series + hidden dim static.

Table 7
Range or options for hyperparameters for TCN model.

Variable	Range/options
Number of hidden units per layer for time series	[5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
Hidden dimension for static data	[15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25]
Kernel size	[3, 5, 7, 9]
Learning rate	Loguniform distribution over [1e−5, 1e-3]
Number of epochs	3 (for EHG model), 6 (for EHG + clinical)
Drop out	Uniform distribution over [0.1, 0.5]
Batch size	[10, 20, 30, 40, 50, 60]
Optimizer	Adam
Optional model	Combination of [BatchNorm, activation layer, Linear layer or None]

*Hidden dimension of the linear layer (as depicted in Fig. 7) for combined time series and static data will be: hidden dim time series + hidden dim static.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Anne Fischer, dr. A.L. Rietveld and dr. P.C.A.M. Bakker are funded by a grant of public–private partnerships from Amsterdam UMC, The Netherlands. This study has been performed in the context of the COCOON (Combining cord-free uterine electrohysterography and standard clinical measurements for refining the detection of premature birth) study, a cooperation of Stichting VUmc, Stichting VU and Bloom Technologies NV. All funding bodies played no role in the creation of this paper.

Appendix

See Tables 6 and 7.

References

[1] J.E. Lawn, M. Kinney, Preterm birth: now the leading cause of child death worldwide, Sci. Transl. Med. 6 (263) (2014) 263ed21.
[2] M.L. McPheeters, W.C. Miller, K.E. Hartmann, D.A. Savitz, J.S. Kaufman, J.M. Garrett, J.M. Thorp, The epidemiology of threatened preterm labor: a prospective cohort study, Am. J. Obstet. Gynecol. 192 (4) (2005) 1325–1329.
[3] A. McEvoy, S. Sabir, Physiology, pregnancy contractions, in: StatPearls [Internet], StatPearls Publishing, 2022.
[4] J.J. Bakker, P.F. Janssen, K. Van Halem, B.Y. Van der Goes, D.N. Papatsonis, J.A. van der Post, B.W.J. Mol, Internal versus external tocodynamometry during induced or augmented labour, Cochrane Database System. Rev. (12) (2012).
[5] E.I. Emin, E. Emin, A. Papalois, F. Willmott, S. Clarke, M. Sideris, Artificial intelligence in obstetrics and gynaecology: is this the way forward? In Vivo 33 (5) (2019) 1547–1551.
[6] M.W. Vlemminx, K.M. Thijssen, G.I. Bajlekov, J.P. Dieleman, M.B. Van Der Hout-Van, D. Jagt, S.G. Oei, Electrohysterography for uterine monitoring during term labour compared to external tocodynamometry and intra-uterine pressure catheter, Euro. J. Obstetrics Gynecol. Reproduct. Biol. 215 (2017) 197–205.
[7] P.C. Bakker, M. Zikkenheimer, H.P. van Geijn, The quality of intrapartum uterine activity monitoring, J. Perinat. Med. 36 (3) (2008) 197–201.
[8] R. Parameshwari, S.S. Devi, Acquisition and analysis of electrohysterogram signal, J. Med. Syst. 44 (3) (2020).
[9] C. Rabotti, M. Mischi, Propagation of electrical activity in uterine muscle during pregnancy: a review, Acta Physiol. 213 (2) (2015) 406–416.
[10] F. Jager, S. Libenšek, K. Geršak, Characterization and automatic classification of preterm and term uterine records, PLoS One 13 (8) (2018) e0202125.
[11] L. Lange, A. Vaeggemose, P. Kidmose, E. Mikkelsen, N. Uldbjerg, P. Johansen, Velocity and directionality of the electrohysterographic signal propagation, PLoS One 9 (1) (2014) e86775.
[12] C. Rabotti, M. Mischi, S.G. Oei, J.W. Bergmans, Noninvasive estimation of the electrohysterographic action-potential conduction velocity, IEEE Trans. Biomed. Eng. 57 (9) (2010) 2178–2187.
[13] E. Mikkelsen, P. Johansen, A. Fuglsang-Frederiksen, N. Uldbjerg, Electrohysterography of labor contractions: propagation velocity and direction, Acta Obstetricia Gynecol. Scandinavica 92 (9) (2013) 1070–1078.
[14] G. Fele-Žorž, G. Kavšek, Ž. Novak-Antolič, F. Jager, A comparison of various linear and non-linear signal processing techniques to separate uterine EMG records of term and pre-term delivery groups, Med. Biol. Eng. Comput. 46 (9) (2008) 911–922.
[15] J. Mas-Cabo, Y. Ye-Lin, C. Benalcazar-Parra, J. Alberola-Rubio, A. Perales, J. Garcia-Casado, G. Prats-Boluda, Electrohysterogram signals from patients with threatened preterm labor: Concentric ring electrode vs disk electrode recordings., in: BIOSIGNALS, 2017, pp. 78–83.
[16] O. Most, O. Langer, R. Kerner, G.B. David, I. Calderon, Can myometrial electrical activity identify patients in preterm labor? Am. J. Obstet. Gynecol. 199 (4) (2008) 378–e1.
[17] M. Lucovnik, R.J. Kuon, L.R. Chambliss, W.L. Maner, S.-Q. SHI, L. Shi, J. Balducci, R.E. Garfield, Use of uterine electromyography to diagnose term and preterm labor, Acta Obstetricia Gynecol. Scandinavica 90 (2) (2011) 150–157.
[18] K. Horoba, J. Jezewski, A. Matonia, J. Wrobel, R. Czabanski, M. Jezewski, Early predicting a risk of preterm labour by analysis of antepartum electrohysterograhic signals, Biocybern. Biomed. Eng. 36 (4) (2016) 574–583.
[19] M. Lucovnik, W.L. Maner, L.R. Chambliss, R. Blumrick, J. Balducci, Z. Novak-Antolic, R.E. Garfield, Noninvasive uterine electromyography for prediction of preterm delivery, Am. J. Obstet. Gynecol. 204 (3) (2011) 228–e1.
[20] U.R. Acharya, V.K. Sudarshan, S.Q. Rong, Z. Tan, C.M. Lim, J.E. Koh, S. Nayak, S.V. Bhandary, Automated detection of premature delivery using empirical mode and wavelet packet decomposition techniques with uterine electromyogram signals, Comput. Biol. Med. 85 (2017) 33–42.
[21] M. Mischi, C. Chen, T. Ignatenko, H. de Lau, B. Ding, S.G. Oei, C. Rabotti, Dedicated entropy measures for early assessment of pregnancy progression from single-channel electrohysterography, IEEE Trans. Biomed. Eng. 65 (4) (2017) 875–884.

[22] M.U. Ahmed, T. Chanwimalueang, S. Thayyil, D.P. Mandic, A multivariate multiscale fuzzy entropy algorithm with application to uterine EMG complexity analysis, Entropy 19 (1) (2016) 2.

[23] P. Fergus, P. Cheung, A. Hussain, D. Al-Jumeily, C. Dobbins, S. Iram, Prediction of preterm deliveries from EHG signals using machine learning, PLoS One 8 (10) (2013) e77154.

[24] A. Smrdel, F. Jager, Separating sets of term and pre-term uterine EMG records, Physiol. Meas. 36 (2) (2015) 341.

[25] W.L. Maner, R.E. Garfield, Identification of human term and preterm labor using artificial neural networks on uterine electromyography data, Ann. Biomed. Eng. 35 (3) (2007) 465–473.

[26] A.S. Alshehri, F. You, Paradigm shift: the promise of deep learning in molecular systems engineering and design, Front. Chem. Eng. 3 (2021) 26.

[27] R. Miotto, F. Wang, S. Wang, X. Jiang, J.T. Dudley, Deep learning for healthcare: review, opportunities and challenges, Brief. Bioinform. 19 (6) (2018) 1236–1246.

[28] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[29] S. Bai, J.Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, 2018, arXiv preprint arXiv: 1803.01271.

[30] I. Gandin, A. Scagnetto, S. Romani, G. Barbati, Interpretability of time-series deep learning models: A study in cardiovascular patients admitted to intensive care unit, J. Biomed. Inform. 121 (2021) 103876.

[31] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, N. Díaz-Rodríguez, Explainable artificial intelligence (xai) on timeseries data: A survey, 2021, arXiv preprint arXiv:2104.00950.

[32] L. Lin, B. Xu, W. Wu, T.W. Richardson, E.A. Bernal, Medical time series classification with hierarchical attention-based temporal convolutional networks: A case study of myotonic dystrophy diagnosis., in: CVPR Workshops, 2019, pp. 83–86.

[33] D. Zhang, C. Yin, K.M. Hunold, X. Jiang, J.M. Caterino, P. Zhang, An interpretable deep-learning model for early prediction of sepsis in the emergency department, Patterns 2 (2) (2021) 100196.

[34] H. Song, D. Rajan, J.J. Thiagarajan, A. Spanias, Attend and diagnose: Clinical time series analysis using attention models, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[35] A.A. Ismail, M. Gunady, H. Corrada Bravo, S. Feizi, Benchmarking deep learning interpretability in time series predictions, Adv. Neural Inf. Process. Syst. 33 (2020) 6441–6452.

[36] B. Lim, S.Ö. Arık, N. Loeff, T. Pfister, Temporal fusion transformers for interpretable multi-horizon time series forecasting, Int. J. Forecast. 37 (4) (2021) 1748–1764.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[38] N.K. Tepper, S.L. Farr, B.B. Cohen, A. Nannini, Z. Zhang, J.E. Anderson, D.J. Jamieson, M. Macaluso, Singleton preterm birth: risk factors and association with assisted reproductive technology, Mater. Child Health J. 16 (4) (2012) 807–813.

[39] R.L. Goldenberg, J.D. Iams, M. Miodovnik, J.P. Van Dorsten, G. Thurnau, S. Bottoms, B.M. Mercer, P.J. Meis, A.H. Moawad, A. Das, et al., The preterm prediction study: risk factors in twin gestations, Am. J. Obstet. Gynecol. 175 (4) (1996) 1047–1053.

[40] P.J. Meis, R.L. Goldenberg, B.M. Mercer, J.D. Iams, A.H. Moawad, M. Miodovnik, M.K. Menard, S.N. Caritis, G.R. Thurnau, S.F. Bottoms, et al., The preterm prediction study: risk factors for indicated preterm births, Am. J. Obstet. Gynecol. 178 (3) (1998) 562–567.

[41] G.S. Berkowitz, C. Blackmore-Prince, R.H. Lapinski, D.A. Savitz, Risk factors for preterm birth subtypes, Epidemiology (1998) 279–285.

[42] J.N. Robinson, E.R. Norwitz, Preterm birth: Risk factors, interventions for risk reduction, and maternal prognosis, 2018, UpToDate. Available on-Line at: https://www-uptodate-com Myaccess. Library. Utoronto. Ca/Contents/Preterm-Birth-Risk-Factors-Interventions-for-Risk-Reduction-and-Maternal-Prognosis.

[43] R.L. Goldenberg, J.F. Culhane, J.D. Iams, R. Romero, Epidemiology and causes of preterm birth, Lancet 371 (9606) (2008) 75–84.

[44] G. Vandewiele, I. Dehaene, G. Kovács, L. Sterckx, O. Janssens, F. Ongenae, F. De Backere, F. De Turck, K. Roelens, J. Decruyenaere, et al., Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling, Artif. Intell. Med. 111 (2021) 101987.

[45] I.O. Idowu, P. Fergus, A. Hussain, C. Dobbins, M. Khalaf, R.V.C. Eslava, R. Keight, Artificial intelligence for detecting preterm uterine activity in gynecology and obstetric care, in: 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, IEEE, 2015, pp. 215–220.

[46] P. Fergus, I. Idowu, A. Hussain, C. Dobbins, Advanced artificial neural network classification for detecting preterm births using EHG records, Neurocomputing 188 (2016) 42–49.

[47] D.K. Degbedzui, M.E. Yüksel, Accurate diagnosis of term–preterm births by spectral analysis of electrohysterography signals, Comput. Biol. Med. 119 (2020) 103677.

[48] J. Ryu, C. Park, Time-frequency analysis of electrohysterogram for classification of term and preterm birth, IEIE Trans. Smart Process. Comput. 4 (2) (2015) 103–109.

[49] S. Janjarasjitt, Examination of single wavelet-based features of EHG signals for preterm birth classification, IAENG Int. J. Comput. Sci. 44 (2) (2017).

[50] N. Sadi-Ahmed, M. Kedir-Talha, Contraction extraction from term and preterm electrohyterographic signals, in: 2015 4th International Conference on Electrical Engineering, ICEE, IEEE, 2015, pp. 1–4.

[51] M. Khalil, J. Duchêne, Uterine EMG analysis: a dynamic approach for change detection and classification, IEEE Trans. Biomed. Eng. 47 (6) (2000) 748–756.

[52] M. Shahrdad, M.C. Amirani, Detection of preterm labor by partitioning and clustering the EHG signal, Biomed. Signal Process. Control 45 (2018) 109–116.

[53] P. Ren, S. Yao, J. Li, P.A. Valdes-Sosa, K.M. Kendrick, Improved prediction of preterm delivery using empirical mode decomposition analysis of uterine electromyography signals, PLoS One 10 (7) (2015) e0132116.

[54] A.J. Hussain, P. Fergus, H. Al-Askar, D. Al-Jumeily, F. Jager, Dynamic neural network architecture inspired by the immune algorithm to predict preterm deliveries in pregnant women, Neurocomputing 151 (2015) 963–974.

[55] S. Hoseinzadeh, M.C. Amirani, Use of electro hysterogram (EHG) signal to diagnose preterm birth, in: Electrical Engineering (ICEE), Iranian Conference on, IEEE, 2018, pp. 1477–1481.

[56] M.U. Khan, S. Aziz, S. Ibraheem, A. Butt, H. Shahid, Characterization of term and preterm deliveries using electrohysterograms signatures, in: 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON, IEEE, 2019, pp. 0899–0905.

[57] J. Peng, D. Hao, L. Yang, M. Du, X. Song, H. Jiang, Y. Zhang, D. Zheng, Evaluation of electrohysterogram measured from different gestational weeks for recognizing preterm delivery: a preliminary study using random forest, Biocybern. Biomed. Eng. 40 (1) (2020) 352–362.

[58] P. Ivaturi, M. Gadaleta, A.C. Pandey, M. Pazzani, S.R. Steinhubl, G. Quer, A comprehensive explanation framework for biomedical time series classification, IEEE J. Biomed. Health Inf. 25 (7) (2021) 2398–2408.

[59] J. Wang, Z. Wang, J. Li, J. Wu, Multilevel wavelet decomposition network for interpretable time series analysis, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2437–2446.

[60] K. Siddiqui, T.E. Doyle, Trust metrics for medical deep learning using explainable-AI ensemble for time series classification, in: 2022 IEEE Canadian Conference on Electrical and Computer Engineering, CCECE, IEEE, 2022, pp. 370–377.

[61] C. Burns, J. Thomason, W. Tansey, Interpreting black box models via hypothesis testing, in: Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference, 2020, pp. 47–57.

[62] S. Tonekaboni, S. Joshi, K. Campbell, D.K. Duvenaud, A. Goldenberg, What went wrong and when? Instance-wise feature importance for time-series black-box models, Adv. Neural Inf. Process. Syst. 33 (2020) 799–809.

[63] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, B. Xu, Attention-based bidirectional long short-term memory networks for relation classification, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016, pp. 207–212.

[64] T.-Y. Hsieh, S. Wang, Y. Sun, V. Honavar, Explainable multivariate time series classification: a deep neural network which learns to attend to important variables as well as time intervals, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 2021, pp. 607–615.

[65] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) e215–e220, Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[66] M.P. Nageotte, W. Dorchester, M. Porto, K.A. Keegan Jr., R.K. Freeman, Quantitation of uterine activity preceding preterm, term, and postterm labor, Am. J. Obstet. Gynecol. 158 (6) (1988) 1254–1259.

[67] R. Jozefowicz, W. Zaremba, I. Sutskever, An empirical exploration of recurrent network architectures, in: International Conference on Machine Learning, PMLR, 2015, pp. 2342–2350.

[68] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[69] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[70] K. Subramaniam, N.V. Iqbal, et al., A review of significant researches on prediction of preterm birth using uterine electromyogram signal, Future Gener. Comput. Syst. 98 (2019) 135–143.

[71] R.E. Garfield, W.L. Maner, L.B. MacKay, D. Schlembach, G.R. Saade, Comparing uterine electromyography activity of antepartum patients versus term labor patients, Am. J. Obstet. Gynecol. 193 (1) (2005) 23–29.

[72] W.L. Maner, R.E. Garfield, H. Maul, G. Olson, G. Saade, Predicting term and preterm delivery with transabdominal uterine electromyography, Obstetrics Gynecol. 101 (6) (2003) 1254–1260.

[73] J. Bergstra, D. Yamins, D. Cox, Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures, in: International Conference on Machine Learning, PMLR, 2013, pp. 115–123.

[74] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.

[75] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, PLoS One 10 (3) (2015) e0118432.

[76] J. Davis, M. Goadrich, The relationship between precision-recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 233–240.