CONTRASTIVE PREDICTIVE CODING FOR ANOMALY DETECTION OF FETAL HEALTH FROM THE CARDIOTOCOGRAM

Ivar R. de Vries[1,2]*, Iris A.M. Huijben*[1]*, René D. Kok*[2]*, Ruud J.G. van Sloun*[1]*, and Rik Vullings*[1,2]

[1]Dept. of Electrical Engineering, Eindhoven University of Technology, The Netherlands
[2]Nemo Healthcare BV, The Netherlands

ABSTRACT

Fetal well-being during labor is currently assessed by medical professionals through visual interpretation of the cardiotocogram (CTG), a simultaneous recording of Fetal Heart Rate (FHR) and Uterine Contractions (UC). This method is disputed due to high inter- and intra-observer variability and a resulting increase in the number of unnecessary interventions. A method for computerized interpretation of the CTG, based on Contrastive Predictive Coding (CPC) is presented here. We hypothesize that the CPC framework, when trained on healthy fetuses only, can predict the FHR response of healthy fetuses to UC, but will provide significant prediction error in case of fetuses with compromised condition. To that end, we have extended the original CPC model by making stochastic, recurrent, and conditioned (upon Uterine Contractions) predictions. We, moreover, introduce a new training objective that was found more suitable for the task of anomaly detection. Based on the detection of out-of-distribution behaviour and deviations from subject-specific behaviour, the proposed model is capable of achieving promising results for identification of suspicious and anomalous FHR events in the CTG, with an average correlation coefficient of 0.80±0.13 with respect to expert annotations.

Index Terms— Cardiotocogram, Anomaly detection, Contrastive Predictive Coding, Conditional prediction

1. INTRODUCTION

During labor, uterine contractions can limit the placental blood flow towards the fetus which may lead to temporary oxygen deprivation. Even though this is common to all births, in around 4 out of 1000 cases this oxygen deprivation causes permanent damage or even death [1, 2]. While an oxygen deficit can be determined from the pH value of the fetal blood, a blood test is typically only performed on the placenta after birth to minimize harm to the newborn arising from fetal scalp blood tests. In an effort to obtain a method for continuous monitoring of the fetal health during labor, Cardiotocography (CTG) was developed. The CTG provides a temporal recording of both the Fetal Heart Rate (FHR) and Uterine Contractions (UC), and the fetal health status can be inferred from these recordings [3, 4]. Even though several guidelines for the visual interpretation of the CTG (by medical experts) exist, it is prone to high inter- and intra-observer variability and research suggests that neonatal well-being is not significantly influenced by the use of the CTG in practice [5, 6, 7]. Moreover, an increase in cesarean sections and other interventions during birth was even reported, which puts the risk-benefit trade-off in CTG monitoring to discussion [8].

Because of this, research has been devoted to obtaining an objective and adequate interpretation of the fetal health from the CTG. Whereas application of traditional signal processing techniques has been focused on automated extraction of clinical CTG features [9, 10], more recent efforts were made to apply machine learning for the classification of oxygen deprivation or fetal morbidity [11, 12]. Current machine learning approaches leverage supervised learning and are focused on classifying the fetal state in two or three classes [13, 14], automated annotation of the features presented in guidelines [15], or even estimation of the pH outcome from the CTG recording [16, 17]. Major challenges of data-hungry machine learning approaches are the low availability of pathological data along with the high variability in pathologies and a scarcity of available labels.

To solve these challenges, this research aims at using an unsupervised (i.e. not requiring labels) machine learning approach for detecting suspicious or anomalous events in the CTG measurements, by only training on the abundant recordings from healthy-born babies. Since the (visual interpretation of) CTG has been reported to have good test properties for assessment of healthy conditions (and poor properties for assessing compromised health) [7], the key aspect of our approach is that, other than attempting to detect signs of compromised health, our method aims to determine whether a fetus is still in good condition. The identification of an abnormal response of the FHR to the corresponding UC (i.e. suspicious or anomalous events) might aid medical experts in the interpretation of fetal health and has potential for clinical decision support.

Our model predicts FHR features based on an accumulated state of both FHR and UC features from the past. This prediction yields bad performance in case of an abnormal response of the fetal heart rate. Predictions are made in a data-

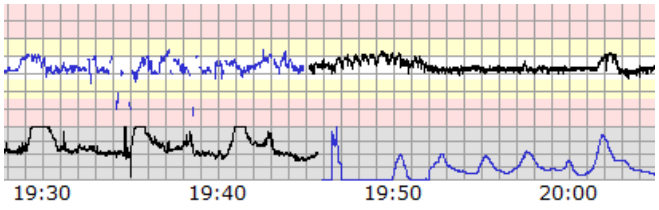

Fig. 1. A CTG recording showing the FHR (top) and UC/toco (bottom). Around 19:45 both external sensors are replaced by internal sensors. The typically occurring clipping of toco can be seen between 19:30 and 19:42.

driven feature space - rather than the data space itself - to make prediction performance independent of (unpredictable) measurement noise. We, moreover, speculate that this learned latent space provides a representation of the underlying physiological system regulating the heart rate, facilitating anomaly detection at this deeper level.

Our model extends the Contrastive Predictive Coding (CPC) framework proposed in [18] with a conditional, recurrent and stochastic prediction, and a custom loss function. Moreover, the full CPC architecture was used during inference, which is contrary to the original approach by [18], where only the learned compact latent space is used for a subsequent downstream (classification) task.

1.1. Cardiotocography data

The CTG data consist of simultaneous recordings of the FHR and a relative measurement of UC, referred to as *toco*. Both signals can be measured externally or internally and it is not uncommon for gynaecologists to switch from the non-invasive external to internal measurements during labor, to improve signal quality[19]. External FHR measurements leverage a Doppler ultrasound transducer, while internal measurements use an invasive fetal scalp electrode. Both sensors measure the same absolute signal (in beats per minute (BPM)), but differ primarily in signal quality and invasiveness.

External monitoring of Uterine Contractions uses a tocodynamometer, which measures the strain on an elastic band placed around the abdomen, providing a relative measurement. Internal monitoring uses an intrauterine pressure catheter, which provides quantitative information on the contraction intensity. However, despite the superior signal quality, intrauterine catheters are not associated with improved labor outcomes and due to their more expensive nature and small induced risk, the external tocodynamometer is preferred in guidelines for routine clinical use [19, 20].

An example CTG recording is shown in Fig. 1. In this recording, both external sensors are exchanged for internal sensors, of which the superior signal quality can be noticed.

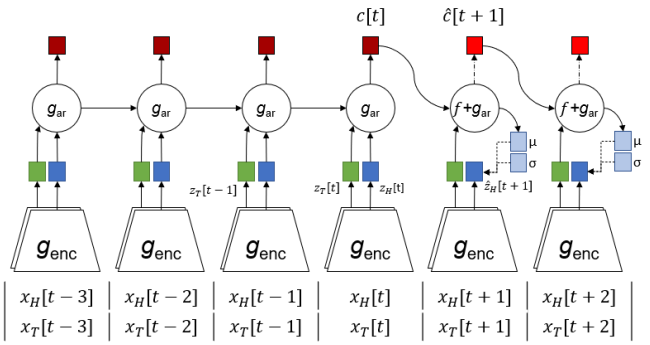

Fig. 2. The adapted conditional CPC architecture, which differs from the original model presented in [18] by the addition of a second independent encoder (one for FHR, and one for toco) and the implementation of a stochastic, recurrent predictor network. Moreover, the model was trained with a custom loss function.

2. METHODS

2.1. Data preprocessing

The data used in this research originate from the Dutch STAN trial [21]. A subset was composed, containing CTG recordings of 688 subjects with a gestation between 36 and 42 weeks. The following exclusion criteria were used: congenital abnormalities, stained amniotic fluid, low birth weight, pH < 7.05, base excess > 12, medication that might influence the CTG, maternal or neonatal fever, and incompleteness of relevant patient data.

Additionally, a dataset consisting of six healthy-born subjects was annotated by an experienced gynaecologist, where each moment in time was annotated as normal, suspicious or anomalous with the availability of all post birth assessments. These six measurements were used for model evaluation during inference.

Both the (fetal) Heart rate signals (H), and toco data (T) were pre-processed to yield a constant sampling frequency of 4 Hz by means of linear interpolation and subsequently normalized using the mean μ_i, and 98th percentile λ_i of the healthy dataset, with $i \in \{H, T\}$:

$$\tilde{x}_i = \frac{x_i - \mu_i}{\lambda_i}. \tag{1}$$

Before applying the normalization given in (1), the toco signal was filtered by a zero-phase, 4^{th} order Butterworth bandpass-filter with cut-off frequencies at 0.001 and 0.1 Hz. This was done to eliminate offset and high-frequency noise arising from clipping (see e.g. Fig. 1, 19:30).

2.2. Model architecture

The proposed model architecture is based on the CPC framework introduced by van den Oord *et al.* [18]. It uses two

independent encoders (indexed by i) to transform each non-overlapping 64-second data window $\boldsymbol{x}_i[t] \in \mathbb{R}^{256}$ into a latent representation $\boldsymbol{z}_i[t] \in \mathbb{R}^8$, with $i \in \{H, T\}$ (H refers to FHR, and T to toco). An autoregressive model subsequently summarises a concatenation of both latent representations in a single context representation $\boldsymbol{c}[t] \in \mathbb{R}^{12}$, which is then used together with the upcoming toco features $\boldsymbol{z}_t[t+1]$ to predict a future latent representation of FHR.

The original CPC framework uses the context vector to predict up to K future latent representations: $\hat{\boldsymbol{z}}[t+k] = f_k(\boldsymbol{c}[t])$, with $k \in \{1, \ldots, K\}$, which gives a prediction accuracy decreasing rapidly with increasing k (i.e. further in the future). This is typically not considered a problem, since common downstream tasks (e.g. classification) do not use said predictions during inference, but only the resulting latent representation.

Contrary to this, we do want to use these predictions during inference, with the aim of anomaly detection. We therefore extend the original model by adopting a recurrent prediction, where each predicted latent representation $\hat{\boldsymbol{z}}_H[t+k]$ is used to update the context representation, which is, in turn, used for the subsequent prediction $\hat{\boldsymbol{z}}_H[t+k+1]$. Additionally, to regularize training, the model was extended with a sampling module, where the predicted latent representation $\hat{\boldsymbol{z}}_H[t+k]$ is drawn from a Normal distribution with mean $\boldsymbol{\mu}[t+k] \in \mathbb{R}^8$ and standard deviation $\boldsymbol{\sigma}[t+k] \in \mathbb{R}^8$ predicted by a function $f(\cdot)$. The resulting conditional, recurrent and stochastic prediction can be summarised by:

$$\begin{aligned} \begin{bmatrix} \hat{\boldsymbol{\mu}}[t] \\ \hat{\boldsymbol{\sigma}}[t] \end{bmatrix} &= f(\hat{\boldsymbol{c}}[t-1], \boldsymbol{z}_T[t]), \\ \hat{\boldsymbol{z}}_H[t] &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}[t], \hat{\boldsymbol{\sigma}}[t]), \\ \text{with} \qquad \hat{\boldsymbol{c}}[t] &= g_{ar}(\hat{\boldsymbol{c}}[t-1], \left[\hat{\boldsymbol{z}}_H[t], \boldsymbol{z}_T[t]\right]). \end{aligned} \tag{2}$$

Figure 2 gives an overview of the architecture used in this research, where the $g_{ar}(\cdot)$ is a Gated Recurrent Unit (GRU) and $f(\cdot)$ is a 3-layer Multilayer perceptron (MLP) with leaky ReLU activations.

2.3. Contrastive losses

Originally, CPC uses the InfoNCE loss, which is the categorical cross-entropy loss for classifying a 'correct' (or positive) sample $\boldsymbol{z}^+$ from a set of N negative samples $\mathcal{Z}^- = \{\boldsymbol{z}_1^-, ..., \boldsymbol{z}_N^-\}$ expanded with the positive sample, all of which specific to a time instance t. The InfoNCE loss thus allows for unsupervised training, reducing the distance between the prediction $\hat{\boldsymbol{z}}_k$ and its corresponding positive sample, while increasing the distance to the negative samples.

While the InfoNCE loss has shown good performance for learning representations for various downstream tasks, it might not naturally yield future predictions suitable for anomaly detection since the prediction error itself is not necessarily minimized. We, therefore, introduce a new loss function for unsupervised anomaly detection. For K predictions

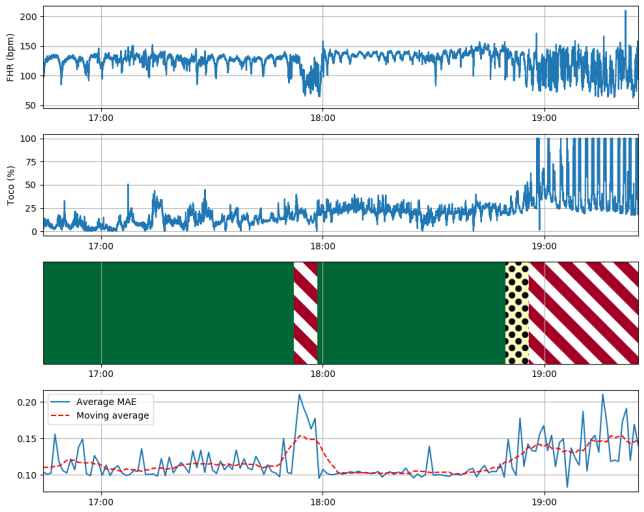

Fig. 3. An example of a measurement, showing FHR, toco, expert annotations and average MAE along with the moving average over 10 windows. Expert annotations are divided in normal (solid green), suspicious (dotted orange) and anomalous (striped red).

at time t, this proposed loss can be generalized as:

$$\mathcal{L}[t] = \frac{1}{K} \sum_{k=1}^{K} \Big(\mathcal{L}_{sim}[t+k] + \mathcal{L}_{contr}[t+k] \Big), \tag{3}$$

where $\mathcal{L}_{sim}$ is a regression loss, optimizing the prediction:

$$\mathcal{L}_{sim}[t] = 1 - cosSim(\hat{\boldsymbol{z}}_H[t], \boldsymbol{z}_H^+[t]),$$

with $cosSim(\cdot)$ the cosine similarity metric. Based on [22], the contrastive part was implemented as

$$\begin{aligned} &\mathcal{L}_{contr}[t] = \\ &\max\left(0, MSE(\hat{\boldsymbol{z}}_H[t], \boldsymbol{z}_H^+[t]) - \min_{\boldsymbol{z}^- \in \mathcal{Z}_H^-[t]} \{MSE(\hat{\boldsymbol{z}}_H[t], \boldsymbol{z}^-)\}\right), \end{aligned}$$

with $MSE(\cdot)$ the Mean Squared Error. This contrastive loss favors a prediction error for all of the negative samples greater than the error for the positive sample, which prevents trivial solutions (i.e. collapsing to a constant $\boldsymbol{z}_H$).

The model proposed in this research uses three past windows to make $K = 4$ predictions and was trained with $N = 4$ negative samples drawn from the same measurement at different time instances.

3. MODEL INFERENCE

The model was used for anomaly detection, where anomalies are characterized by unusual behaviour of the FHR with respect to the corresponding UC (i.e. they yield high prediction error). First of all, the model was trained on data

of healthy children only, and therefore primarily trained on non-anomalous data. Even though suspicious or anomalous events do occur in the training data, they have a much lower prevalence compared to healthy events and the corresponding prediction error made by the model will be higher for such suspicious or anomalous events.

Secondly, since the adapted CPC network is presumably capable of modelling the subject-specific UC-FHR interaction (modelled in $\boldsymbol{c}[t]$), deviations from the fetus's normal behaviour are considered suspicious or anomalous and will result in significant prediction error.

3.1. Anomaly metric

Following the two assumptions given above, either unhealthy behaviour or changes in behaviour are expected to lead to significant prediction errors, which we quantify trough the Mean Absolute Error (MAE). Since our model makes K predictions ahead, each FHR window is predicted K times and the average, standardized MAE for these predictions was used as a metric:

$$\mathcal{M}[t] = \frac{1}{\alpha K} \sum_{j=1}^{K} MAE(\hat{\boldsymbol{\mu}}_H^j[t], \boldsymbol{z}_H^+[t]), \tag{4}$$

with $\boldsymbol{\mu}_H^j[t]$ being the mean for the j$^{\text{th}}$ time $\boldsymbol{z}_H[t]$ has been predicted. Since the absolute value of MAE might differ between experiments (e.g. due to a difference in measurement quality), the scaling factor α, being the mean over the first 15 minutes, was added. This standardization technique was chosen since it allows for real-time implementation and none of the annotated measurements had any suspicious or anomalous behaviour during this initial period. This method was tested on the annotated dataset, for which Fig. 3 shows an example of a measurement along with the expert annotations and model output (un-standardized, i.e. $\alpha = 1$). These expert annotations were used to divide each measurement in events (of different duration) belonging to each of the three classes.

The model performance was evaluated by means of the correlation between the model output and the expert annotations. For this, a 10-window moving average filter was used on $\mathcal{M}[t]$ (see Fig. 3-bottom). Correlations were evaluated after assigning each window to a class based on the expert annotations and for each measurement separately.

3.2. Results

Figure 4 shows the performance for the filtered $\mathcal{M}[t]$ on all time windows corresponding to the expert annotations. Even though overlap between the Normal, Suspicious and Anomalous sets is present, the reported area under the ROC-curves is 0.81, 0.96 and 0.78 for the distinction between the Normal-Suspicious, Normal-Anomalous and Suspicious-Anomalous sets, respectively. The combined model output (as given in Fig. 4 left) yields a correlation coefficient of 0.70, where the six individual measurements, filtered by the same 10-window moving average filter, have an average correlation coefficient of 0.80±0.13 with respect to the unfiltered expert annotations.

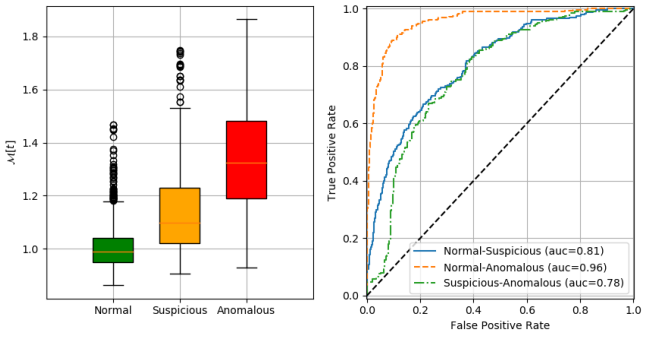

Fig. 4. $\mathcal{M}[t]$ filtered with a 10-window moving average filter for the three classes (left) along with the corresponding ROC curves for the three classes (right).

4. CONCLUSIONS

With the aim of anomaly detection in CTG data, anomalies were defined as events with an abnormal response of the FHR to the corresponding UC. With a model trained on data from healthy-born subjects, the MAE between the latent space and its predictions is used as a metric for anomaly detection.

The original Contrastive Predictive Coding model [18] was extended to provide conditional and stochastic predictions of future windows with a recurrent predictor architecture and was trained with a custom loss function. Even though the method of evaluation is sensitive to minor mismatches (e.g. a minute too early or too late) in the annotations, validation on six annotated recordings shows the model's capability of distinguishing between normal and anomalous events, and to a lesser extent between normal and suspicious or suspicious and anomalous events.

Future work should be focused on evaluation on a bigger dataset, preferably annotated by different experts. Furthermore, measurements with an unhealthy outcome (not used in this research due to their scarcity and diversity in pathologies) might be included in future work. Lastly, future work should be focused on the implementation of a standardization technique for the anomaly metric $\mathcal{M}[t]$ which generalizes to all measurements.

5. REFERENCES

[1] Anna-Karin Sundström, David Rosén, and KG Rosén, *Fetal surveillance*, Neoventa Medical AB, 2005.

[2] Nadia Badawi, Jennifer J. Kurinczuk, et al., "Antepartum risk factors for newborn encephalopathy: The West-

ern Australian case-control study," *British Medical Journal*, vol. 317, no. 7172, pp. 1549–1553, 1998.

[3] Julian T. Parer, Tekoa King, et al., "Fetal acidemia and electronic fetal heart rate patterns: Is there evidence of an association?," *Journal of Maternal-Fetal and Neonatal Medicine*, vol. 19, no. 5, pp. 289–294, 2006.

[4] Elizabeth S. Draper, Jenny J. Kurinczuk, et al., "A confidential enquiry into cases of neonatal encephalopathy," *Archives of Disease in Childhood: Fetal and Neonatal Edition*, vol. 87, no. 3, pp. 176–181, 2002.

[5] Zarko Alfirevic, Gillian M.L. Gyte, et al., "Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour," *Cochrane Database of Systematic Reviews*, vol. 2017, no. 2, 2017.

[6] Dick K. Donker, Herman P. van Geijn, and Arie Hasman, "Interobserver variation in the assessment of fetal heart rate recordings," *European Journal of Obstetrics and Gynecology and Reproductive Biology*, vol. 52, no. 1, pp. 21–28, 1993.

[7] Sahana Das, Himadri Mukherjee, et al., "Shortcoming of Visual Interpretation of Cardiotocography: A Comparative Study with Automated Method and Established Guideline Using Statistical Analysis," *SN Computer Science*, vol. 1, no. 3, pp. 1–18, 2020.

[8] Gayani Jayasooriya and Veronica Djapardy, "Intrapartum assessment of fetal well-being," *BJA Education*, vol. 17, no. 12, pp. 406–411, 2017.

[9] Vaclav Chudacek, Jiří Spilka, et al., "Evaluation of feature subsets for classification of cardiotocographic recordings," *Computers in Cardiology*, vol. 35, no. 21, pp. 845–848, 2008.

[10] Mario Cesarelli, Maria Romano, and Paolo Bifulco, "Comparison of short term variability indexes in cardiotocographic foetal monitoring," *Computers in Biology and Medicine*, vol. 39, no. 2, pp. 106–118, 2009.

[11] Paul Fergus, Carl Chalmers, et al., "Modelling Segmented Cardiotocography Time-Series Signals Using One-Dimensional Convolutional Neural Networks for the Early Detection of Abnormal Birth Outcomes," *IEEE Transactions on Emerging Topics in Computational Intelligence*, pp. 1–11, 2020.

[12] Jiří Spilka, George Georgoulas, et al., "Discriminating normal from "abnormal" pregnancy cases using an automated FHR evaluation method," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8445 LNCS, pp. 521–531, 2014.

[13] Philip A. Warrick, Emily F. Hamilton, et al., "Classification of normal and hypoxic fetuses from systems modeling of intrapartum cardiotocography," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 771–779, 2010.

[14] Ersen Yılmaz, "Fetal State Assessment from Cardiotocogram Data Using Artificial Neural Networks," *Journal of Medical and Biological Engineering*, vol. 36, no. 6, pp. 820–832, 2016.

[15] Robert D.F. Keith, Jenny Westgate, et al., "Suitability of artificial neural networks for feature extraction from cardiotocogram during labour," *Medical and Biological Engineering and Computing*, vol. 32, pp. 51–57, 1994.

[16] Tony K.H. Chung, Michele P. Mohajer, et al., "The prediction of fetal acidosis at birth by computerised analysis of intrapartum cardiotocography," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 102, no. 6, pp. 454–460, 1995.

[17] Sudip Kundu, Elna Kuehnle, et al., "Estimation of neonatal outcome artery pH value according to CTG interpretation of the last 60 min before delivery: a retrospective study. Can the outcome pH value be predicted?," *Archives of Gynecology and Obstetrics*, vol. 296, no. 5, pp. 897–905, 2017.

[18] Aaron Van Den Oord, Yazhe Li, and Oriol Vinyals, "Representation Learning with Contrastive Predictive Coding," *arXiv preprint arXiv:1807.03748*, 2018.

[19] Diogo Ayres-de Campos, Catherine Spong, et al., "FIGO consensus guidelines on intrapartum fetal monitoring: Cardiotocography," *International Journal of Gynecology and Obstetrics*, vol. 131, no. 1, pp. 13–24, 10 2015.

[20] Jannet J.H. Bakker, Petra F. Janssen, et al., "Internal versus external tocodynamometry during induced or augmented labour," *Cochrane Database of Systematic Reviews*, vol. 2013, no. 8, 2013.

[21] Michelle E.M.H. Westerhuis, Gerard H.A. Visser, et al., "Cardiotocography plus ST analysis of fetal electrocardiogram compared with cardiotocography only for intrapartum monitoring: A randomized controlled trial," *Obstetrics and Gynecology*, vol. 117, no. 2 PART 1, pp. 406–407, 2011.

[22] Sumit Chopra, Raia Hadsell, and Yann LeCun, "Learning a Similarity Metric Discriminatively, with Application to Face Verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, pp. 539–546.