A critical look at studies applying over-sampling on the TPEHGDB dataset

Gilles Vandewiele[1], Isabelle Dehaene[2], Olivier Janssens[1], Femke Ongenae[1], Femke De Backere[1], Filip De Turck[1], Kristien Roelens[2], Sofie Van Hoecke[1], and Thomas Demeester[1]

[1] IDLab, Ghent University – imec
Technologiepark-Zwijnaarde 126, Ghent, Belgium
{name}.{familyname}@ugent.be
[2] Department of Gynaecology and Obstetrics, Ghent University Hospital
Corneel Heymanslaan 10, Ghent, Belgium
{name}.{familyname}@uzgent.be

Abstract. Preterm birth is the leading cause of death among young children and has a large prevalence globally. Machine learning models, based on features extracted from clinical sources such as electronic patient files, yield promising results. In this study, we review similar studies that constructed predictive models based on a publicly available dataset, called the Term-Preterm EHG Database (TPEHGDB), which contains electrohysterogram signals on top of clinical data. These studies often report near-perfect prediction results, by applying over-sampling as a means of data augmentation. We reconstruct these results to show that they can only be achieved when data augmentation is applied on the entire dataset prior to partitioning into training and testing set. This results in (i) samples that are highly correlated to data points from the test set are introduced and added to the training set, and (ii) artificial samples that are highly correlated to points from the training set being added to the test set. Many previously reported results therefore carry little meaning in terms of the actual effectiveness of the model in making predictions on unseen data in a real-world setting. After focusing on the danger of applying over-sampling strategies before data partitioning, we present a realistic baseline for the TPEHGDB dataset and show how the predictive performance and clinical use can be improved by incorporating features from electrohysterogram sensors and by applying over-sampling on the training set.

Keywords: Preterm birth · Electrohysterogram (EHG) · Imbalanced data · over-sampling

1 Introduction

Giving birth before 37 weeks of pregnancy, which is referred to as preterm birth, has a significant negative impact on the expected outcome of the neonate. According to the World Health Organization (WHO), preterm birth is one of the

leading causes of death among young children, and its' prevalence ranges from 5% to 18% globally [23]. As preterm labor is currently not yet fully understood, gynecologists are experiencing difficulties in assessing whether a patient recently admitted to the hospital will deliver at term or not. In order to support experts in their assessment, several studies have already investigated the added value of a predictive model [24,35,6,13]. These models are based on a large number of variables extracted from clinical sources such as the electronic health record. These variables include the gestational age, results of a biomarker, cervical length, clinical history, and more. In this study, we provide a thorough and extensive overview of related work on a public dataset and discuss many of the overly optimistic results. These results are often obtained by introducing a large bias through over-sampling the dataset, before partitioning the data, in order to combat the class imbalance, i.e., the fact that it contains many more pregnancies with term deliveries than preterm. Afterwards, we set a realistic baseline and assess the impact of correct over-sampling and of incorporating features extracted from the electrohysterogram data.

2 The impact of over-sampling prior to data partitioning

In this section, we highlight the impact of applying over-sampling prior to the data partitioning on an artificially generated dataset. We generated a binary classification problem with 100 samples. Twenty samples were marked positive (red circles), and the others negative (blue squares). The generated dataset is depicted on the left of Figure 1 (step 0). We now compare the effect of over-sampling data after partitioning with the effect of over-sampling prior to partitioning. In the former approach, we first partition our data into two mutually exclusive sets (step 1). Then, we create artificial samples (red, unfilled circles) that are highly correlated to the training samples of the minority class (step 2) in order to have a similar number of samples for both classes in our training set. On the other hand, if we over-sample the data prior to partitioning, we generate train samples that are highly correlated with original data points that will end up in the test set (step 1). Moreover, some of the generated artificial samples will be distributed to the test set as well (step 2). These two consequences result in highly optimistic results that merely reflect the model's capability to memorize samples seen during training, rather than its predictive performance if it were applied in a real-world setting on unseen data.

3 A critical look on studies reporting near-perfect results on the TPEHGDB dataset

In 2008, a public dataset, called TPEHGDB (Term/Preterm ElectroHysteroGram DataBase), containing 300 records, which correspond to 300 pregnancies, has been released on PhysioNet [14,9]. Each record consists of three raw bipolar signals that express the difference in electric potentials, measured by four

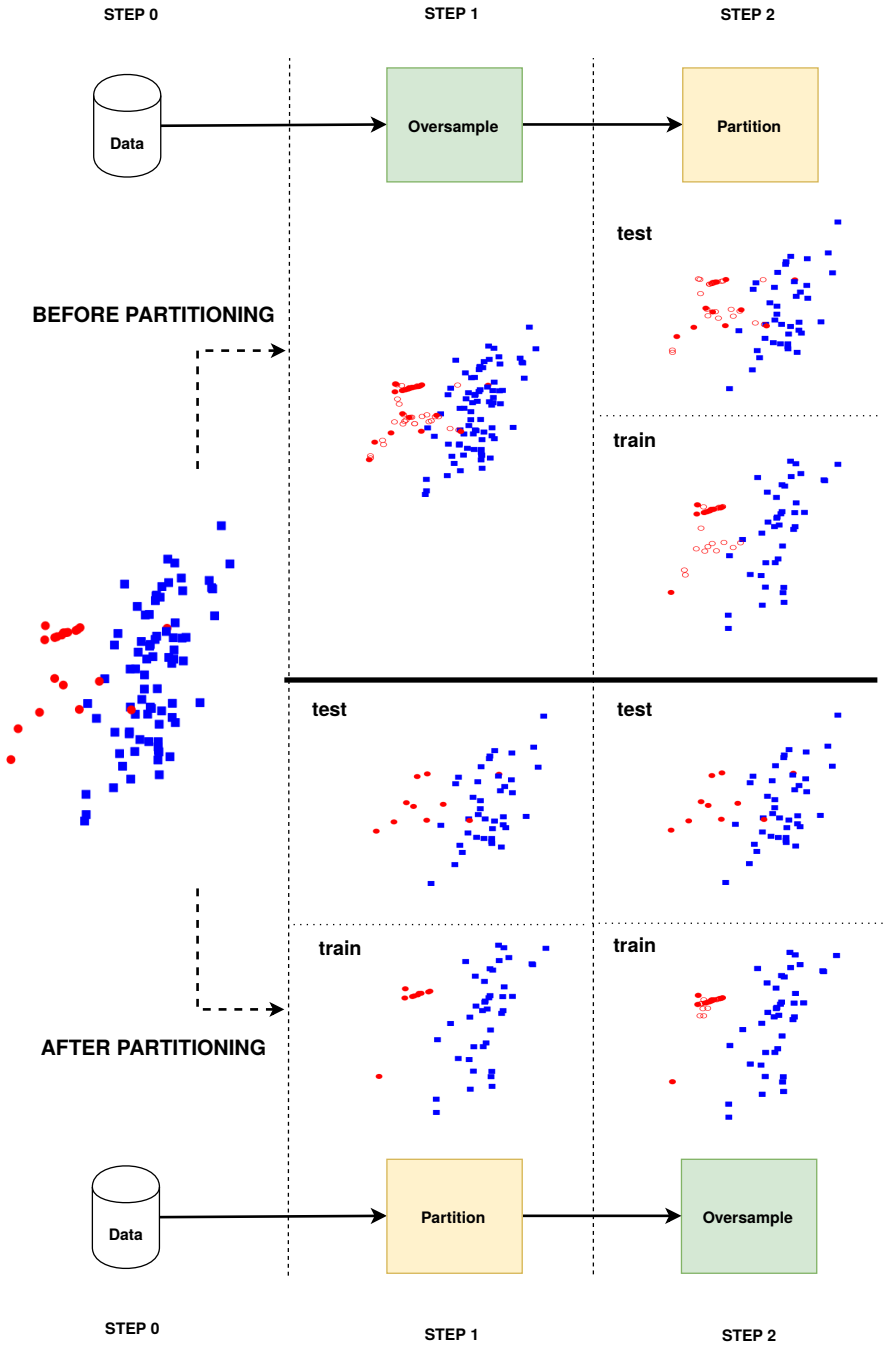

Fig. 1. Comparing the impact of applying over-sampling prior to data partitioning to applying over-sampling after data partitioning on a two-dimensional classification problem.

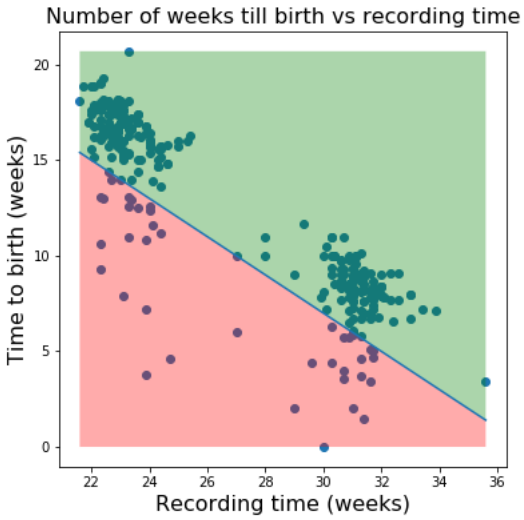

Fig. 2. The number of weeks till birth expressed in function of the gestational age in weeks at the time of recording. All data points within the red area correspond to preterm deliveries, while the ones within the green area correspond to term deliveries.

electrodes placed on the abdomen. In addition, each record is accompanied by clinical variables, such as the gestational age at recording time, the age and weight of the mother, and whether an abortion occurred in the patients' medical history. The recordings can be categorized as being captured at an early stage of pregnancy (gestational age of 23.11 ± 0.77 weeks) or at a later stage of pregnancy (31.09 ± 1.05 weeks). Recordings were captured at a frequency of 20 Hz for about 30 minutes. In Figure 2 the number of weeks till birth is plotted in function of the gestational age at the time of recording and displayed according to term or preterm delivery. Clearly, an imbalance is present in the dataset with more term (green area) than preterm (red area) deliveries (262 vs 38).

While impressive (near-perfect) results on the TPEHGDB dataset are reported in many studies [10,20,33,2,16,19,12,27,1,34,21,29,18,17,11,15], these results should be interpreted cautiously as their evaluation methodology is based on applying over-sampling strategies before data partitioning. All these studies apply over-sampling in order to make the distribution of classes more uniform. These over-sampling techniques are applied prior to partitioning the data into two mutually exclusive sets (referred to as the training and testing set). As discussed earlier, this causes the predictive performance metrics to be overly optimistic.

Nevertheless, a significant number of studies on the TPEHGDB dataset do not apply any over-sampling technique. However, in these studies, certain decisions concerning the evaluation were often made which raises serious questions concerning the credibility of the provided results [3,26,25,32,8,4]. In many of these studies, results were either not obtained through cross-validation, or cross-validation was applied on a subset of data subsampled from the original dataset. Performing this kind of pre-processing, in a machine learning context, without any kind of argumentation, raises doubts since it drastically increases the variance of the obtained results and avoids the problem of imbalanced data, which does not reflect reality in terms of potential applications. In other studies, segments are extracted from the original signals, which are highly correlated with each other, and then partitioned into training and testing set [7,31], which again results in highly optimistic results.

At the time of writing (December 2018), within all the 153 citations to the original paper, which introduced the TPEHGDB dataset, we have found three machine learning studies that were accessible and, to the best of our knowledge, had a sound evaluation methodology [28,22,30]. In the study of Sadi-Ahmed et al. [30], all records taken before 26 weeks of gestation were filtered away from the dataset, resulting in a dataset of 138 recordings taken after the 26th week of gestation. All of these signals were processed in order to detect contractions through Auto-Regressive Moving Averages (ARMA). From the detected contractions, features were extracted such as the total number of contractions, average duration and average time between contractions. Unfortunately, only an accuracy score of 0.89 to distinguish between term and preterm pregnancies was achieved within this study, making it hard to assess the clinical use of such a model. It is important to note that, on this filtered dataset, an accuracy score of 0.86 can be achieved by always predicting term birth, precisely because of the aforementioned class imbalance, with a fraction of 119 term deliveries on 138 records from 26 weeks onwards. Janjarasjitt et al. proposed a new type of feature, based on a wavelet decomposition of the signals [22]. The feature was evaluated by tuning a threshold on a single feature in a leave-one-out cross-validation scheme. A sensitivity and specificity of 0.6842 and 0.7133 are achieved. While these scores are very promising, it should be noted, that they are rather optimistic due to the fact that the evaluation happened in a leave-one-out scheme. As such, the performance of the sample entropy feature, provided along with the original data, closely matches, and sometimes even outperforms, that of the proposed feature. Nevertheless, the wavelet-based feature may be an interesting addition to the feature set. In the work of Ryu et al. [28] a similar study is performed in which a feature based on Multivariate Emperical Mode Decomposition (MEMD) is proposed. They evaluate the added value of their feature, by subsampling a balanced dataset of 38 term and 38 preterm records, 100 times, from the original dataset. They found that the AUC improved from 0.5698 to 0.6049 by adding their feature to the dataset. While this subsampling strategy again avoids the problem of imbalanced data, which is reflected in the

original dataset, it does show an improvement in AUC and thus indicates that adding the MEMD-based feature to the dataset could be beneficial for the predictive performance. Moreover, due to the many repetitions of the experiment, the sample mean better reflects the real mean.

4 Setting a realistic baseline for the TPEHGDB dataset

In this section, we will assess the effects of incorporating information from raw EHG signals, and of over-sampling the data after partitioning, on the predictive performance of the resulting model. Moreover, we will show that predictive performances similar to the aforementioned studies can be only be achieved through over-sampling before data partitioning.

Seven machine learning algorithms were trained on the original dataset consisting of clinical features and four features extracted from the raw EHG signals, i.e.: the root mean square value & entropy of the raw signal and the median and peak frequency from the spectral information of each signal. The seven different classification techniques, and their corresponding abbreviations, are: (1) Logistic Regression (LR), (2) Decision Trees (DT), (3) Linear Discriminant Analysis (LDA), (4) Quadratic Discriminant Analysis (QDA), (5) K-Nearest Neighbors (KNN), (6) Random Forests (RF), and (7) Support Vector Machines (SVM). All reported results are generated using five-fold stratified cross-validation. Hyperparameters were tuned using grid search. Moreover, to solve the issue of imbalanced data, and to improve the clinical use of the different classifiers, we apply over-sampling, using SMOTE [5], on the train set. We compare these results to when SMOTE is applied on the entire dataset, to show that near-perfect predictive performance can only be achieved by introducing label leakage.

In total, we evaluate four different approaches: (i) clinical features and no over-sampling, (ii) clinical and EHG features and no over-sampling, (iii) clinical and EHG features and over-sampling in a correct fashion, and finally (iv) clinical and EHG features and over-sampling in an incorrect fashion. The first two approaches are compared in Table 1. As can be seen, the AUC scores drastically improve when features, extracted from the EHG signals, are incorporated. Nevertheless, the clinical use of both approaches is very limited, as all the models almost always predict that someone will deliver at term (which is reflected in the low sensitivity scores), which is a typical problem that arises when dealing with imbalanced data. The performance for both over-sampling approaches is listed in Table 2. We can conclude that the near-perfect performances from the studies mentioned in Section 3 can only be closely matched by applying over-sampling prior to data partitioning. If we apply over-sampling on the training set, we see that the clinical use of a predictive model for preterm birth prediction, based on the TPEGHDB dataset, is still limited, with a maximum AUC score of 63.20%.

Algorithm	Sensitivity (%)		Specificity (%)		AUC (%)	
	clinical	all	clinical	all	clinical	all
LR	0 ± 0	0 ± 0	100 ± 0	100 ± 0	48 ± 6	58 ± 7
DT	3 ± 5	0 ± 0	96 ± 4	96 ± 3	47 ± 6	62 ± 9
LDA	0 ± 0	0 ± 0	97 ± 3	96 ± 4	54 ± 9	59 ± 5
QDA	28 ± 34	11 ± 11	67 ± 36	90 ± 7	48 ± 5	62 ± 4
KNN	0 ± 0	0 ± 0	100 ± 1	98 ± 2	50 ± 8	57 ± 7
RF	0 ± 0	0 ± 0	99 ± 2	95 ± 4	52 ± 8	58 ± 5
SVM	0 ± 0	0 ± 0	100 ± 0	100 ± 0	52 ± 8	56 ± 9

Table 1. The results obtained with seven different classifiers, on (i) a dataset constructed using solely clinical variables and (ii) a dataset with clinical variables concatenated to four features extracted from the EHG data. No over-sampling is applied for both approaches.

Algorithm	Sensitivity (%)		Specificity (%)		AUC (%)	
	correct	incorrect	correct	incorrect	correct	incorrect
LR	39 ± 26	74 ± 3	68 ± 19	66 ± 6	59 ± 6	78 ± 3
DT	40 ± 16	81 ± 3	71 ± 10	84 ± 5	59 ± 3	86 ± 4
LDA	53 ± 14	73 ± 1	59 ± 10	69 ± 7	59 ± 5	78 ± 3
QDA	51 ± 36	100 ± 0	58 ± 24	41 ± 7	61 ± 6	79 ± 2
KNN	48 ± 16	99 ± 1	64 ± 3	73 ± 5	58 ± 6	92 ± 2
RF	42 ± 13	91 ± 4	68 ± 6	95 ± 2	63 ± 5	98 ± 1
SVM	43 ± 31	99 ± 2	64 ± 23	86 ± 4	58 ± 6	98 ± 1

Table 2. The results obtained with seven different classifiers, on the entire TPEHGDB dataset, constructed using clinical features and features extracted from the 3 filtered EHG signals. Oversampling with SMOTE is applied before data partitioning (column *correct*) versus after data partitioning (column *incorrect*).

5 Conclusion and future work

This study tackles the problem of preterm birth risk prediction, based on the publicly available dataset TPEHGDB. Our contributions are two-fold. First, in the light of a significant body of recent literature, we show that applying over-sampling for data-augmentation purposes, prior to partitioning the data into separate parts for training and evaluation, leads to overly optimistic results. To evaluate a model's predictive performance, the data partitioning needs to be performed before applying over-sampling. Second, a realistic baseline was set in which it was shown how an increase in AUC score can be obtained by using features extracted from electrohysterogram recordings, besides clinical observations. This confirms the potential added value of such recordings. In future work we will investigate whether deep learning techniques can improve the predictive performance by directly training on the raw recordings, as opposed to manually extracting features. Unfortunately, for this, a larger dataset may be required.

6 Acknowledgements

Gilles Vandewiele is funded by a scholarship of FWO (1S31417N). This study has been performed in the context of the 'Predictive health care using text analysis on unstructured data project', funded by imec, and the PRETURN (PREdiction Tool for prematUre laboR and Neonatal outcome) clinical trial (EC/2018/0609) of Ghent University Hospital.

7 Reproducibility and dataset availability

In order to allow reproduction of the reported results on this public dataset, we host all code, required to reproduce the results reported in this paper, on a public GitHub repository[1]. The dataset is available from that repository, or from the original hosting location[2].

References

1. Acharya, U.R., Sudarshan, V.K., Rong, S.Q., Tan, Z., Lim, C.M., Koh, J.E., Nayak, S., Bhandary, S.V.: Automated detection of premature delivery using empirical mode and wavelet packet decomposition techniques with uterine electromyogram signals. Computers in biology and medicine **85**, 33–42 (2017)
2. Ahmed, M.U., Chanwimalueang, T., Thayyil, S., Mandic, D.P.: A multivariate multiscale fuzzy entropy algorithm with application to uterine emg complexity analysis. Entropy **19**(1), 2 (2016)
3. Baghamoradi, S.M.S., Naji, M., Aryadoost, H.: Evaluation of cepstral analysis of ehg signals to prediction of preterm labor. In: Biomedical Engineering (ICBME), 2011 18th Iranian Conference of. pp. 81–83. IEEE (2011)
4. Beiranvand, M., Shahbakhti, M., Eslamizadeh, M., Bavi, M., Mohammadifar, S.: Investigating wavelet energy vector for pre-term labor detection using ehg signals. In: Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), 2017. pp. 269–274. IEEE (2017)
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
6. De Silva, D.A., Lisonkova, S., von Dadelszen, P., Synnes, A.R., Magee, L.A.: Timing of delivery in a high-risk obstetric population: a clinical prediction model. BMC pregnancy and childbirth **17**(1), 202 (2017)
7. Despotović, D., Zec, A., Mladenović, K., Radin, N., Turukalo, T.L.: A machine learning approach for an early prediction of preterm delivery. In: 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY). pp. 000265–000270. IEEE (2018)
8. Far, D.T., Beiranvand, M., Shahbakhti, M.: Prediction of preterm labor from ehg signals using statistical and non-linear features. In: Biomedical Engineering International Conference (BMEiCON), 2015 8th. pp. 1–5. IEEE (2015)

[1] `https://github.com/IBCNServices/TPEHGDB-Experiments/`
[2] `https://physionet.org/physiobank/database/tpehgdb/`

9. Fele-Žorž, G., Kavšek, G., Novak-Antolič, Ž., Jager, F.: A comparison of various linear and non-linear signal processing techniques to separate uterine emg records of term and pre-term delivery groups. Medical & biological engineering & computing **46**(9), 911–922 (2008)
10. Fergus, P., Cheung, P., Hussain, A., Al-Jumeily, D., Dobbins, C., Iram, S.: Prediction of preterm deliveries from ehg signals using machine learning. PloS one **8**(10), e77154 (2013)
11. Fergus, P., Hussain, A., Al-Jumeily, D., Hamdan, H.: A machine learning system for automatic detection of preterm activity using artificial neural networks and uterine electromyography data. International Journal of Adaptive and Innovative Systems **2**(2), 161–179 (2015)
12. Fergus, P., Idowu, I., Hussain, A., Dobbins, C.: Advanced artificial neural network classification for detecting preterm births using ehg records. Neurocomputing **188**, 42–49 (2016)
13. García-Blanco, A., Diago, V., De La Cruz, V.S., Hervás, D., Cháfer-Pericás, C., Vento, M.: Can stress biomarkers predict preterm birth in women with threatened preterm labor? Psychoneuroendocrinology **83**, 19–24 (2017)
14. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (June 2000). https://doi.org/10.1161/01.CIR.101.23.e215, `http://circ.ahajournals.org/content/101/23/e215`
15. Hoseinzadeh, S., Amirani, M.C.: Use of electro hysterogram (ehg) signal to diagnose preterm birth. In: Electrical Engineering (ICEE), Iranian Conference on. pp. 1477–1481. IEEE (2018)
16. Hussain, A.J., Fergus, P., Al-Askar, H., Al-Jumeily, D., Jager, F.: Dynamic neural network architecture inspired by the immune algorithm to predict preterm deliveries in pregnant women. Neurocomputing **151**, 963–974 (2015)
17. Idowu, I.O.: Classification Techniques Using EHG Signals for Detecting Preterm Births. Ph.D. thesis, Liverpool John Moores University (2017)
18. Idowu, I.O., Fergus, P., Hussain, A., Dobbins, C., Al Askar, H.: Advance artificial neural network classification techniques using ehg for detecting preterm births. In: 2014 Eighth International Conference on Complex, Intelligent and Software Intensive Systems (CISIS). pp. 95–100. IEEE (2014)
19. Idowu, I.O., Fergus, P., Hussain, A., Dobbins, C., Khalaf, M., Eslava, R.V.C., Keight, R.: Artificial intelligence for detecting preterm uterine activity in gynecology and obstetric care. In: Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015 IEEE International Conference on. pp. 215–220. IEEE (2015)
20. Jager, F., Libensek, S., Gersak, K.: Characterization and automatic classification of preterm and term uterine records. bioRxiv p. 349266 (2018)
21. Janjarasjitt, S.: Evaluation of performance on preterm birth classification using single wavelet-based features of ehg signals. In: Biomedical Engineering International Conference (BMEiCON), 2017 10th. pp. 1–4. IEEE (2017)
22. Janjarasjitt, S.: Examination of single wavelet-based features of ehg signals for preterm birth classification. IAENG International Journal of Computer Science **44**(2) (2017)

23. Liu, L., Oza, S., Hogan, D., Chu, Y., Perin, J., Zhu, J., Lawn, J.E., Cousens, S., Mathers, C., Black, R.E.: Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the sustainable development goals. The Lancet **388**(10063), 3027–3035 (2016)
24. Meertens, L.J., van Montfort, P., Scheepers, H.C., van Kuijk, S.M., Aardenburg, R., Langenveld, J., van Dooren, I.M., Zwaan, I.M., Spaanderman, M.E., Smits, L.J.: Prediction models for the risk of spontaneous preterm birth based on maternal characteristics: a systematic review and independent external validation. Acta obstetricia et gynecologica Scandinavica (2018)
25. Naeem, S., Ali, A., Eldosoky, M.: Kl. comparison between using linear and nonlinear features to classify uterine electromyography signals of term and preterm deliveries. In: Radio Science Conference (NRSC), 2013 30th National. pp. 492–502. IEEE (2013)
26. Naeem, S.M., Seddik, A.F., Eldosoky, M.A.: New technique based on uterine electromyography nonlinearity for preterm delivery detection. Journal of Engineering and Technology Research **6**(7), 107–114 (2014)
27. Ren, P., Yao, S., Li, J., Valdes-Sosa, P.A., Kendrick, K.M.: Improved prediction of preterm delivery using empirical mode decomposition analysis of uterine electromyography signals. PloS one **10**(7), e0132116 (2015)
28. Ryu, J., Park, C.: Time-frequency analysis of electrohysterogram for classification of term and preterm birth. IEIE Transactions on Smart Processing & Computing **4**(2), 103–109 (2015)
29. Sadi-Ahmed, N., Kacha, B., Taleb, H., Kedir-Talha, M.: Relevant features selection for automatic prediction of preterm deliveries from pregnancy electrohysterograhic (ehg) records. Journal of medical systems **41**(12), 204 (2017)
30. Sadi-Ahmed, N., Kedir-Talha, M.: Contraction extraction from term and preterm electrohyterographic signals. In: Electrical Engineering (ICEE), 2015 4th International Conference on. pp. 1–4. IEEE (2015)
31. Shahrdad, M., Amirani, M.C.: Detection of preterm labor by partitioning and clustering the ehg signal. Biomedical Signal Processing and Control **45**, 109–116 (2018)
32. Sim, S., Ryou, H., Kim, H., Han, J., Park, K.: Evaluation of electrohysterogram feature extraction to classify the preterm and term delivery groups. In: The 15th International Conference on Biomedical Engineering. pp. 675–678. Springer (2014)
33. Smrdel, A., Jager, F.: Separating sets of term and pre-term uterine emg records. Physiological measurement **36**(2), 341 (2015)
34. Subramaniam, K., Iqbal, N.V., et al.: Classification of fractal features of uterine emg signal for the prediction of preterm birth. Biomedical and Pharmacology Journal **11**(1), 369–374 (2018)
35. Watson, H., Carter, J., Seed, P., Tribe, R., Shennan, A.: Quipp app: a safe alternative to a treat-all strategy for threatened preterm labor. Ultrasound in Obstetrics & Gynecology **50**(3), 342–346 (2017)