Research Article

Design and Assessment of a Robust and Generalizable ANN-Based Classifier for the Prediction of Premature Birth by means of Multichannel Electrohysterographic Records

J. Mas-Cabo,[1] G. Prats-Boluda ⓘ,[1] J. Garcia-Casado ⓘ,[1] J. Alberola-Rubio,[2] A. Perales,[2] and Y. Ye-Lin[1]

[1]*Centro de Investigación e Innovación en Bioingeniería, Universitat Politècnica de València, Camino de Vera s/n Ed.8B, 46022 Valencia, Spain*
[2]*Servicio de Obstetricia y Ginecología, Hospital Universitario y Politícnico La Fe de Valencia, Av. Fernando Abril Martorell 106, Edificio F, 3ª Planta, Valencia, Spain*

Correspondence should be addressed to G. Prats-Boluda; gprats@ci2b.upv.es

Preterm labor is one of the major causes of neonatal deaths and also the cause of significant health and development impairments in those who survive. However, there are still no reliable and accurate tools for preterm labor prediction in clinical settings. Electrohysterography (EHG) has been proven to provide relevant information on the labor time horizon. Many studies focused on predicting preterm labor by using temporal, spectral, and nonlinear parameters extracted from single EHG recordings. However, multichannel analysis, which includes information from the whole uterus and about coupling between the recording areas, may provide better results. The cross validation method is often used to design classifiers and evaluate their performance. However, when the validation dataset is used to tune the classifier hyperparameters, the performance metrics of this dataset may not properly assess its generalization capacity. In this work, we developed and compared different classifiers, based on artificial neural networks, for predicting preterm labor using EHG features from single and multichannel recordings. A set of temporal, spectral, nonlinear, and synchronization parameters computed from EHG recordings was used as the input features. All the classifiers were evaluated on independent test datasets, which were never "seen" by the models, to determine their generalization capacity. Classifiers' performance was also evaluated when obstetrical data were included. The experimental results show that the classifier performance metrics were significantly lower in the test dataset (AUC range 76-91%) than in the train and validation sets (AUC range 90-99%). The multichannel classifiers outperformed the single-channel classifiers, especially when information was combined into mean efficiency indexes and included coupling information between channels. Including obstetrical data slightly improved the classifier metrics and reached an AUC of $91.1 \pm 2.5\%$ for the test dataset. These results show promise for the transfer of the EHG technique to preterm labor prediction in clinical practice.

1. Introduction

Preterm labor (PL), defined by the World Health Organization as all deliveries before 37 weeks (259 days) of gestation [1], is one of the most urgent challenges in healthcare. It is associated with 75% of perinatal deaths [2], while those who survive have a greater risk of health issues and neurodevelopmental disabilities, and require strict monitoring by specialists in their early years [2, 3]. These situations entail high social and economic costs; a comprehensive study in this area estimated that the annual social economic burden attributed to preterm births in 2005 in the U.S. was $26.2 billion, more than $50,000 per premature newborn, or about 5 times that of a term birth [4].

In most European countries, the PL rate ranges between 6 and11% and has stagnated or even increased in recent years [5]. It has been reported that an advanced maternal age (40 and over) can be associated with a higher risk of

preterm birth. Maternal age has gradually increased worldwide, especially in high-income countries [6], while recent developments in artificial reproductive techniques have fomented pregnancies in women outside the usual biological reproductive age, thus increasing the PL risk [6].

Early diagnosis of PL is highly important for tocolytic drugs (which have to be administrated as soon as PL is detected [7]) to succeed in extending pregnancy long enough to allow corticosteroids to act and promote fetal maturation. Although great efforts have been made, and certain markers such as the Bishop Score, fetal fibronectin, cervical length, and tocodynamometry have been used for PL prediction, they have a limited prediction capacity and can be inaccurate or subjective [8, 9]. Indeed, the value of these tests lies mainly in their high negative predictive values, while their positive values are lower and do not always identify the patients who will give birth prematurely [9, 10].

The uterine electrical activity changes throughout pregnancy; it is scarce and barely synchronized in the early gestational ages and becomes more intense and coordinated as labor approaches [11, 12]. This phenomena is related to increased myometrial cell excitability [13, 14] and coupling [15–17], resulting in a large number of recruited cells and thus in effective contractions that end in labor. The contraction of uterine myometrial cells involves changes in electrical activity due to the flow of ionic currents. The noninvasive recording of this activity, known as the electrohysterogram (EHG), from the maternal abdominal wall has emerged as one of the most promising tools for PL prediction [18–20].

The temporal (root-mean-square amplitude, peak to peak amplitude) and spectral parameters (peak frequency of the power spectrum (PS), median frequency of the PS, or the ratio of high-frequency power/low-frequency power) are commonly computed from EHG signals. Some authors have shown that the EHG spectral content shifts towards higher frequencies with the approach of delivery [20–22]. Spectral parameters are more robust and less sensitive to interpatient variability than temporal parameters in distinguishing labor from nonlabor and term from preterm deliveries [11]. Like any other biological system, uterine electrophysiology involves nonlinear and complex processes, and several studies have worked out nonlinear and complex parameters, such as sample entropy, Lyapunov exponent, time reversibility, or Lempel-Ziv, to obtain additional information on physiological changes during pregnancy [23–25]. It has been reported that EHG signals become more organized or less "chaotic" [20, 26] and EHG synchronization indexes from multichannel recordings have been found to increase as labor approaches [16, 27].

Several authors have reported the development of classifiers for PL prediction using temporal, spectral, nonlinear, and complex EHG parameters [25, 28–30]. Although the results are promising, they still have certain limitations. Firstly, they attempted to predict PL from information from single-channel EHG recordings, even when multichannel recordings were available [31–34]. The channel with the highest signal-to-noise ratio was commonly used and information from other channels ignored. To enhance classifier robustness, uterine electrophysiological information from different recording sites and coupling indexes should be included in preterm birth predictive models.

On the other hand, the cross validation method is generally used to design and evaluate classifier performance [31–34], i.e., the validation set is used to "fine-tune" the model's hyperparameters, such as pruning parameters for the decision trees, the value of k for the nearest neighbor algorithm, and the learning features (learning rate, momentum, early stopping, and initial conditions) for the neural networks [35, 36]. The true generalization capacity of the classifiers therefore cannot be assessed from these validation groups, and a test dataset "unseen" by the classifiers would be needed to further evaluate their performance [31, 32, 34, 37, 38]. Moreover, only few studies consider the use of obstetrical data in addition to EHG parameters [32, 38].

With respect to the classification method, since the physiological mechanisms of biological systems often consist of nonlinear processes [39], nonlinear classifiers such as support vector machines (SVMs), k-nearest neighbor (KNN), and artificial neural networks (ANNs), the most common options in forecasting applications [40–43], are usually preferred. In this regard, when the sample size is limited, KNN often present an inferior performance metrics, since it is highly dependent on the training dataset and the dimensionality of the input features [40, 43].

Although SVM and ANNs are both universal nonlinear function approximators [44, 45], they present differences in the nonlinear data classification: SVM employs nonlinear mapping to make the linear data separable in which selecting the kernel is a key factor in classifier performance [41, 43, 45], while ANNs use multilayer connections and several activation functions to deal with nonlinear problems [40]. Both algorithms are powerful tools for pattern classification and recognition and have been widely used for forecasting tasks due to their ability to learn from experience and generalize [30, 33, 44]. From a general point of view, both algorithms usually provide similar performance, although ANNs seem to be more accurate at solving classification problems [45] and have been successfully used for labor prediction applications, reporting relatively high accuracy values on training and validation datasets [30, 33]. In this work, ANNs were selected for preterm labor prediction because of their capacity to learn from examples and extract functional relationships, even when the underlying relationships are unknown or hard to describe [44].

In this context, our aim was therefore to develop robust and generalizable classifiers for predicting PL based on ANN. For this purpose, both single-channel EHG feature and multichannel EHG using novel uterine contractile efficiency indexes, which take into account the EHG synchronization, were fed to the classifiers.

The improved performance achieved by adding obstetrical information to the preterm labor prediction classifiers was also evaluated. Since the objective was to obtain generalizable predictive models, the performance of the different classifiers was compared, not only on the training and validation dataset but also on an independent set of test data.

The following steps were performed to develop generalizable ANN-based classifiers: firstly, as the sample of preterm
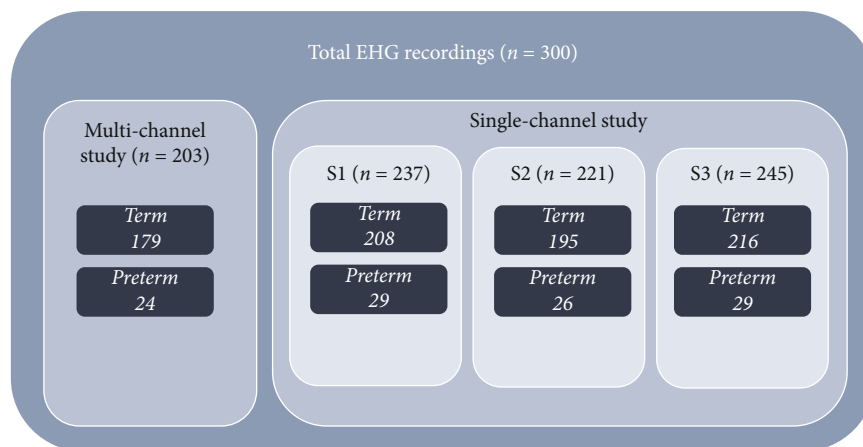
Figure 1: Distribution of the EHG recordings according to analysis type.

labor women was much smaller than that of the term labor sample, it could have made the classifier "learn" that the majority group had a high specificity but low sensitivity in preterm labor detection. To overcome this problem, the SMOTE technique was applied to oversample the minority class. The data was then randomly divided into 30 trials to reduce the bias, and principal component analysis was then performed to reduce data dimensionality and thus try to avoid overfitting. ANN-based classifiers were then trained and evaluated in each trial on the training, validation, and independent test dataset.

2. Materials and Methods

2.1. Database. Multichannel EHG signals recorded during regular check-ups from women delivering at term and preterm were used in this study from the open access database "Term-Preterm EHG Data base" (TPEHGDB) available on PhysioNet [20]. This database contains 300 EHG recordings from pregnant women, 262 from those who delivered at term and the remaining 38 from women who delivered prematurely. All EHG recordings were from individual women between 22 and 32 days of gestational age. The selected database includes three bipolar recordings (S1, S2, and S3) from four disposable electrodes symmetrically placed on the abdominal surface in two horizontal rows and horizontally and vertically separated by a distance of 7 cm [20]. The TPEHGDB also includes the following obstetrical data: maternal age, parity, previous abortions, gestational age, and fetal weight at the time of the recording, hypertension, diabetes, placental position, funneling, smoker, and bleeding in the first and second trimester. Only the first five obstetrical data were included as classifier input features. The remaining features consisted of categorical variables and missing data, since there were very few patients in the positive class (e.g., 2 for hypertension and 3 for diabetes).

2.2. Data Analysis. Since the main content of the EHG signals distributes in the range of 0.1 to 4 Hz [46, 47], bipolar signals were digitally filtered in that range (5th-order Butterworth band-pass in forward and backward direction to obtain zero-phase shift). Instead of traditional EHG-burst analysis,

we preferred to perform the whole EHG window analysis, which has been shown to provide relevant information for predicting preterm labor and is more easily integrated in real-time applications [38, 48]. In this respect, signal sections with evident motion artifacts were discarded from the single channels by visual inspection. Indeed, some EHG recordings had to be removed because of poor signal quality. Multichannel analysis was performed only in signal sections in which all the channels were artifact-free. Figure 1 shows the sample size of preterm and term labor records for each single-channel and multichannel analysis.

EHG signals were divided into 120 s analysis windows with a 50% overlap in order to include representative sections of the recording at a reasonable computational cost [48].

For each analysis window, several EHG characteristics of different nature (temporal, spectral, and nonlinear parameters) were computed for each single channel. Peak to peak amplitude was computed because it is directly related to the intensity of uterine electrical activity. A set of spectral parameters including mean frequency [48], dominant frequency [20, 37], H/L ratio, which represents the relation between the energy computed in (0.34-1 Hz) with respect to the energy computed in (0.2-0.34 Hz), and the deciles of the power spectrum density were selected since they are related to cell excitability [19, 49]. We also included a set of nonlinear parameters that have been widely used to characterize the electrophysiological state of the uterus and other bioelectrical signals such as EEG, including sample entropy [26], spectral entropy [50], fuzzy entropy [51], Lempel-Ziv complexity (binary and multistate versions) [25], time reversibility [23], Poincaré plot metrics (SD1, SD2, SDRR, and SD1/SD2) [52], and Higuchi's fractal dimension [53].

For the multichannel analysis, two approaches were adopted; in the first, the different temporal, spectral, and nonlinear parameters computed from the single channels were fed to the classifier. In the second, to estimate the synchronization degree of the different areas of the uterus, a bivariate method based on normalized permutation cross mutual information (NPCMI) [54], which has been proven to better discriminate imminent term labor [55], was computed from the different pairs of EHG channels. We then computed the mean efficiency index (MEI) of the different

TABLE 1: Summary of the different computed parameters grouped according to type.

Temporal	Spectral	Nonlinear	Synchronization
	Mean frequency	Sample entropy	NPCMI
	Dominant frequency	Spectral entropy	
	H/L ratio	Fuzzy entropy	
	Frequency deciles	LZ-binary ($n = 2$)	
Peak to peak amplitude		LZ-multi ($n = 6$)	
		Time reversibility	
		SD1 SD2 SDs SDRR (Poincaré)	
		Higuchi fractal dimension	

parameters proposed in a previous work to define a more robust indicator of uterine electrical activity efficiency from multichannel recordings [55]. MEI was defined according to the formula described in:

$$\text{MEI(Ft)} = \frac{\sqrt{\text{Ft}_{S1} * \text{Ft}_{S2}} + \sqrt{\text{Ft}_{S1} * \text{Ft}_{S3}} + \sqrt{\text{Ft}_{S2} * \text{Ft}_{S3}}}{3}, \quad (1)$$

where Ft_{S1}, Ft_{S2}, and Ft_{S3} are any EHG characteristics (temporal, spectral, and nonlinear) estimated from single-channel S1, S2, and S3, respectively. In the case of NPCMI, the product of each pair of single-channel parameter in the formula is replaced by the coupling index between this pair of channels.

A brief summary of the computed parameters in this work, grouped by families, can be found in Table 1.

2.3. Data Balancing.

The dataset included in the TPEHGDB is clearly unbalanced, since it contains 262 recordings from women who delivered at term and only 38 from women who delivered preterm. This situation means that classifiers are more susceptible to detecting term situations and reduce the probability of correctly detecting PL situations. To overcome this problem, one of the solutions proposed in the literature is the use of oversampling techniques to increase the number of minority class samples [31, 32]. In this work, we oversampled the minority class up to the number of the majority class samples by the synthetic minority oversampling technique (SMOTE) [56]. This technique (based on k-nearest neighbor interpolation) has been used in previous studies on TPEHGDB to solve the imbalance class problem [33, 34, 38] and has been reported to outperform other techniques, such as downsampling [45]. In this work, 5 neighbors were selected to interpolate and obtain new minority class samples. In order to check the deviation of the accuracy of the models due to SMOTE variability when generating new samples, we also oversampled the minority class ten times.

2.4. Dimensionality Reduction.

So as to avoid generalization errors of classifiers developed due to overfitting, which occurs when the number of parameters of the classifier is very high with respect to the number of training samples, dimensionality reduction techniques are applied [57]. Specifically, in the present work, PCA was selected to perform dimensionality reduction of the input features, since a large number of
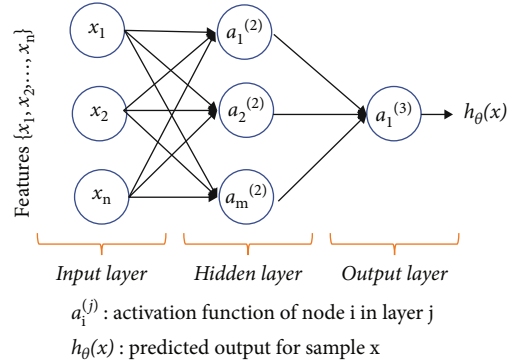

FIGURE 2: Multilayer perceptron artificial neural network scheme.

input features could cause overfitting, while the relatively small size of the database could lead to overfitting [57, 58]. We thus decided to perform PCA, since it retains a relatively high value for the initial variance and significantly reduces the number of features and has been already used by other authors in the EHG field [51, 59]. It involves an orthogonal linear transformation of the original data projected onto a new set of coordinates of decreasing variance [57, 59]. After PCA, the resulting components were selected sequentially until 98% of the original variance was reached, in order to maintain a trade-off between dimensionality reduction and the amount of retained information.

2.5. Classifier Design and Evaluation.

The artificial neural network (ANN) classification algorithm was selected to build the classifiers, due to its performance when dealing with nonlinear problems [36]. Multilayer perceptron (MLP) classifiers were used, setting the hyperbolic tangent as activation function for all neurons. MLP is a class of feed-forward artificial neural network, which consists of at least three layers (see Figure 2) of neurons in which each layer is fully connected to the next: the input layer, the hidden layer, and the output layer [34, 44]. MLP training is supervised in that the real desired class for each input is always available [44]. The input weights of each classifier were adjusted iteratively by the backpropagation training algorithm, which is conceptually simple and computationally efficient [60].

Ten different classifiers were developed based on feed-forward neural networks trained by the backpropagation algorithm. The different classifiers included three versions
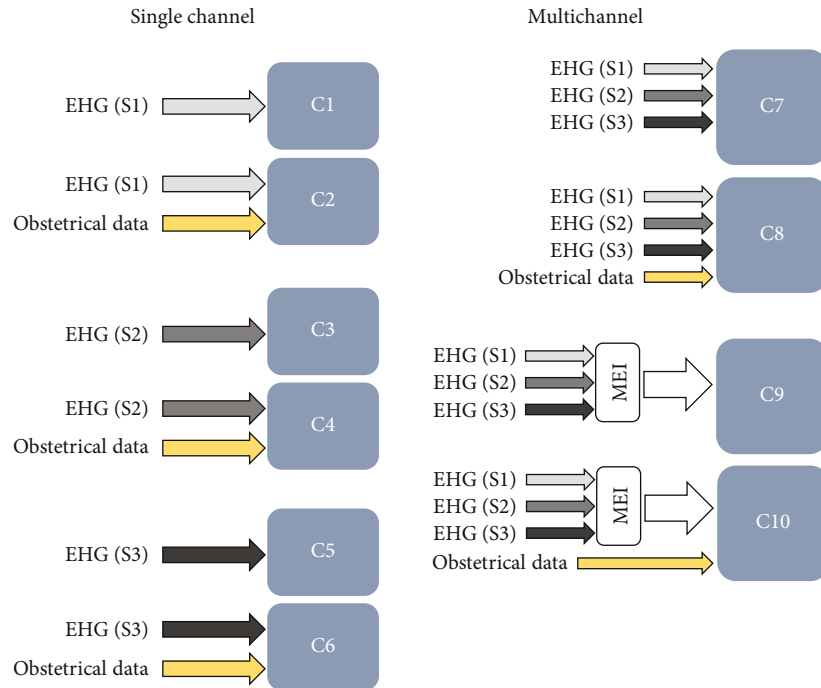
FIGURE 3: Scheme of the different developed classifiers (C1…C10).

with single-channel information from the bipolar channels S1, S2, and S3 (classifiers: C1, C3 and C5); one multichannel version which considered the information from the three individually computed bipolar channels (C7); and a last multichannel version that made use of the mean parameter efficiency index (C9). Another five modified versions of the previously described classifiers were developed to include obstetrical data in addition to the EHG parameters (C2, C4, C6, C8, and C10). A visual representation of the different developed classifiers and the parameters involved is shown in Figure 3.

To address the possible overfitting problem due to an excess of hidden units or overparameterized training data [58, 60], we decided to use only one hidden layer and performed a grid search of the neuron number from 2 to 8 so as to determine the optimal topology for predicting preterm labor in each classifier. In this regard, the number of hidden neurons in the ANN hidden layer was gradually increased from 2 to 10, selecting the best topology according to the improved performance over the training and validation datasets. We also used the early stopping methodology for regularization, because it has been reported to significantly reduce overfitting when combined with back-propagation [60, 61].

In order to assess the generalization capacity of the trained ANNs, the dataset swas randomly divided 30 times, splitting the data into three equal parts in each iteration (holdout methodology): 1/3 for training, 1/3 for validation, and 1/3 for testing (see Figure 4). Since the initial weights were set randomly, ANN training with the train and validation partitions was performed 30 times in each iteration to avoid stacking in a local minimum. Only the best of the 30 neural networks per topology and iteration was finally selected. Table 2 shows the optimal number of neurons of

the hidden layer and features included in each classifier prior and after performing the PCA. Classifiers C9 and C10 include an additional EHG NPCMI feature.

When the ANNs are trained, different metrics including the accuracy, sensibility, specificity, predictive positive value (PPV), the negative predictive value (NPV), and area under curve (AUC) of the receiver operating characteristic (ROC) are evaluated for each dataset: training, validation, and test.

$$
\begin{aligned}
\text{Accuracy (\%)} &= \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}, \\
\text{Sensibility (\%)} &= \frac{\text{TP}}{(\text{TP} + \text{FN})}, \\
\text{Specificity (\%)} &= \frac{\text{TN}}{(\text{TN} + \text{FP})}, \\
\text{PPV (\%)} &= \frac{\text{TP}}{(\text{TP} + \text{FP})}, \\
\text{NPV (\%)} &= \frac{\text{TN}}{(\text{TN} + \text{FN})},
\end{aligned}
\tag{2}
$$

where TP represents the true positives, TN the true negatives, and FP and FN constitute the false positives and false negatives, respectively. All the metrics were evaluated for the 30 iterations in each of the 10 different balanced datasets using SMOTE, computing their mean value and standard deviation, and finally displaying the values associated to the topology with the best results. An ANOVA analysis was also performed to determine whether statistically significant differences were found on the AUC obtained by the different classifiers on the test dataset. The Shapiro-Wilk test had
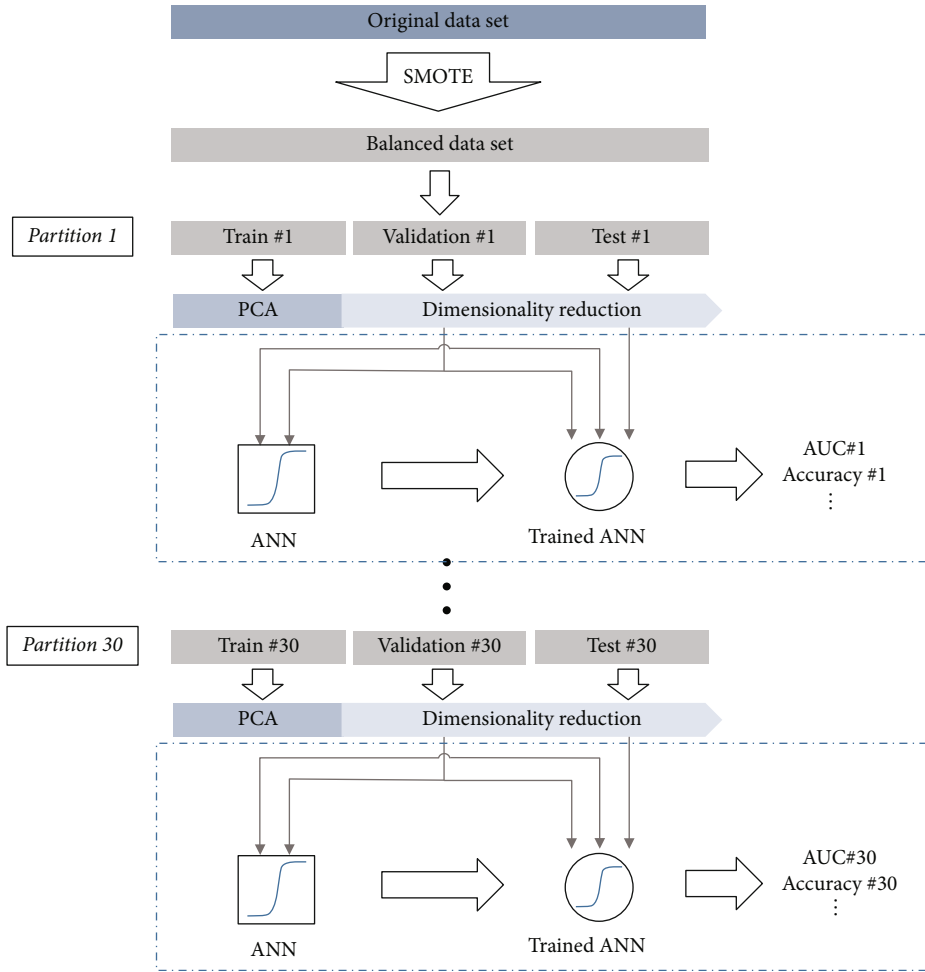
Figure 4: Diagram of the method used to train and validate the neural networks.

Table 2: Number of hidden layer nodes, initial features, and principal components included after performing PCA associated with each classifier (C1…C10).

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Hidden layer neurons	7	6	7	7	8	7	6	5	7	7
EHG features	24	24	24	24	24	24	72	72	25	30
Obstetrical features	0	5	0	5	0	5	0	5	0	5
Features after PCA	14	16	14	16	14	16	22	26	14	16

previously been applied to check that all AUC values came from normal distributions.

3. Results

Tables 3 and 4 show the mean and standard deviation of the different classifiers' performance on the training, validation, and test sets, obtained from the 10 balanced datasets by the SMOTE technique. Regardless of the feature set used for designing the ANN classifiers (single channel, multichannel, and with or without obstetrical data), their performance (accuracy and AUC) was better than 90% for both training and validation data (Table 3). No noticeable differences were found in the classifier metrics for training and validation

data. These results suggest that the ANN-based predictive models were able to learn from the underlying structure of the input features.

However, the classifiers' performance was significantly worse for the test datasets (Table 4) than the training and validation data. Indeed, there was a great variability in their accuracy (from $73.2 \pm 1.3\%$ to $87.9 \pm 1.6\%$) and AUC (from $76.5 \pm 2.1\%$ to $91.1 \pm 2.5\%$), according to their input features. As for single-channel classifiers' and EHG features only, S3 (C5) seems to contain better information for PL prediction than channel S1 (C1). When EHG features extracted from the three single channels were fed to the classifier (C7), there was no noticeable improvement over the C5 classifier metrics. By contrast, a considerable

Table 3: Mean and standard deviation (between brackets) of different classifier metrics obtained for the training and validation datasets for the 10 balanced data obtained using SMOTE techniques. For each balanced dataset, a cross validation method was carried out 30 times.

	Accuracy (%)		AUC (%)		Sensibility (%)		Specificity (%)		PPV (%)		NPV (%)	
	Train	Validation	Train	Validation	Train	Validation	Train	Validation	Train	Validation	Train	Validation
C1	91.4 (1.1)	90.2 (1.8)	91.3 (1.3)	90.4 (2.1)	91.4 (1.1)	89.7 (1.6)	91.4 (1.1)	90.7 (1.9)	91.4 (1.1)	90.7 (1.9)	91.4 (1.1)	89.7 (1.6)
C2	95.5 (0.9)	94.2 (1.4)	95.4 (1.1)	94.5 (1.9)	95.6 (0.7)	93.7 (1.2)	95.5 (1.1)	94.7 (1.8)	95.5 (1.0)	94.7 (1.8)	95.6 (0.8)	93.8 (1.2)
C3	98.7 (1.0)	98.1 (1.6)	99.8 (1.1)	99.6 (1.9)	99.5 (0.8)	98.8 (1.3)	99.4 (1.3)	99.5 (0.021)	99.6 (1.2)	99.5 (2.0)	99.2 (0.8)	98.8 (1.3)
C4	98.6 (0.6)	98.3 (1.2)	99.6 (0.7)	98.9 (1.5)	99.8 (0.5)	97.9 (1.0)	99.5 (0.8)	98.8 (1.6)	99.5 (0.8)	98.8 (1.6)	99.8 (0.5)	97.9 (1.0)
C5	98.8 (1.3)	98.4 (2.0)	99.5 (1.5)	98.8 (2.2)	99.7 (1.3)	98.8 (1.9)	99.4 (1.3)	98.9 (2.2)	99.6 (1.3)	99.1 (2.1)	99.8 (1.3)	98.9 (1.9)
C6	98.8 (1.0)	98.7 (1.6)	99.6 (1.3)	99.2 (2.1)	98.9 (0.8)	98.5 (1.3)	99.6 (1.2)	99.1 (2.0)	99.5 (1.2)	98.9 (2.0)	99.6 (0.8)	98.8 (1.3)
C7	98.7 (1.0)	98.5 (1.7)	99.8 (1.2)	99.5 (1.9)	99.3 (0.8)	99.0 (1.3)	99.7 (0.8)	99.4 (2.1)	99.7 (1.3)	99.1 (2.1)	99.7 (0.8)	99.2 (1.3)
C8	99.1 (0.7)	98.8 (1.3)	99.7 (0.7)	99.0 (1.5)	99.7 (0.5)	99.1 (1.5)	99.5 (1.3)	99.2 (1.7)	99.5 (0.8)	98.8 (1.7)	99.8 (0.5)	99.5 (1.0)
C9	98.8 (1.3)	98.6 (2.0)	99.7 (1.5)	99.3 (2.3)	99.5 (1.3)	99.0 (1.8)	99.5 (1.3)	99.5 (2.1)	99.8 (1.3)	96.5 (2.1)	99.8 (1.3)	99.3 (1.8)
C10	98.7 (0.9)	98.2 (1.5)	99.8 (1.2)	99.4 (2.0)	99.7 (0.8)	99.8 (1.2)	99.7 (1.1)	99.5 (1.9)	99.7 (1.1)	99.4 (1.9)	99.8 (0.8)	99.8 (1.3)

TABLE 4: Mean and standard deviation (between brackets) of different classifier metrics obtained for the test datasets for the 10 balanced data obtained using SMOTE techniques. For each balanced dataset, a cross validation method was carried out 30 times.

	Accuracy (%)	AUC (%)	Sensibility (%)	Specificity (%)	PPV (%)	NPV (%)
C1	73.2 (1.3)	76.5 (2.1)	70.5 (2.0)	75.3 (2.2)	74.8 (2.7)	73.6 (2.1)
C2	75.2 (1.5)	79.7 (2.3)	72.1 (2.1)	76.2 (2.4)	75.2 (2.7)	73.9 (2.2)
C3	80.2 (1.5)	83.7 (2.7)	76.7 (2.3)	82.6 (2.7)	79.8 (3.2)	77.6 (1.7)
C4	81.6 (1.5)	84.8 (2.8)	77.3 (1.9)	83.4 (2.5)	81.6 (3.3)	78.7 (1.8)
C5	82.8 (1.5)	84.7 (2.4)	79.3 (1.8)	84.1 (2.2)	83.6 (3.2)	80.1 (1.5)
C6	83.1 (1.6)	85.2 (2.6)	79.8 (1.9)	84.7 (2.0)	83.4 (2.9)	79.8 (2.2)
C7	81.7 (1.6)	86.6 (2.8)	79.7 (1.8)	85.2 (2.8)	84.5 (3.1)	80.1 (2.0)
C8	83.5 (1.6)	88.2 (2.9)	80.8 (1.8)	86.3 (3.1)	85.3 (2.7)	81.6 (1.9)
C9	85.8 (1.4)	88.7 (2.3)	81.8 (2.0)	87.3 (2.8)	86.3 (3.2)	81.2 (1.7)
C10	87.9 (1.6)	91.1 (2.5)	84.4 (2.1)	89.2 (2.6)	88.7 (2.7)	84.3 (1.9)

improvement of the classifier performance was obtained when using the MEI estimated from the three channels as input features (C9). The accuracy and AUC of the ROC of C9 were about $85.8 \pm 1.4\%$ and $88.7 \pm 2.3\%$, respectively, for the test dataset, presenting a sensitivity of $81.8 \pm 2.0\%$ and a specificity of $87.3 \pm 2.8\%$. Adding obstetrical data as input features to the classifiers slightly improved their metrics. The mean AUC of the classifier improvement ranged from 0.5% to 3.2%, depending on the input features. In this respect, no relevant improvement of any classifier metric was found between C5 (using only EHG characteristics of the channel S3 as input feature) and its corresponding C6 model (see Figure 5). The best performance was achieved when using the MEI estimated from the three channels together with obstetrical data (C10), obtaining a sensitivity and specificity of $84.4 \pm 2.1\%$ and $89.2 \pm 2.6\%$, respectively, the accuracy and AUC being $87.9 \pm 1.6\%$ and $91.1 \pm 2.5\%$, respectively.

Figure 5 shows box and whisker plots (a) of the AUC of the ROC curve of the different classifiers for the test dataset and the statistical significance between the different classifiers' performance in the bottom trace (b). Only AUC was selected to be represented, since the other metrics showed similar trends. When only EHG features were fed to the classifiers, the performance of C1 (on the information extracted from channel S1) was significantly lower than that using the EHG features from channels S2 and S3 (C3 and C5). No significant difference was found between the performance of C3 and C5. Classifier C7, which used the EHG features extracted from each single channel, did not provide a significantly higher AUC of the ROC curve than C5. By contrast, the C9 performance, which includes the MEI computed from multichannel recordings, was significantly higher than those of C5 and C7. The performance of the classifiers that used only the EHG information embedded in single-channel and multichannel recordings was also compared with their corresponding paired classifier to determine whether the obstetric data provided relevant information for predicting PL. A significant improvement was only obtained in classifier performance between the following pairs: C1-C2, C3-C4, and C9-C10 (see Figure 5(b)). The C10 performance, which
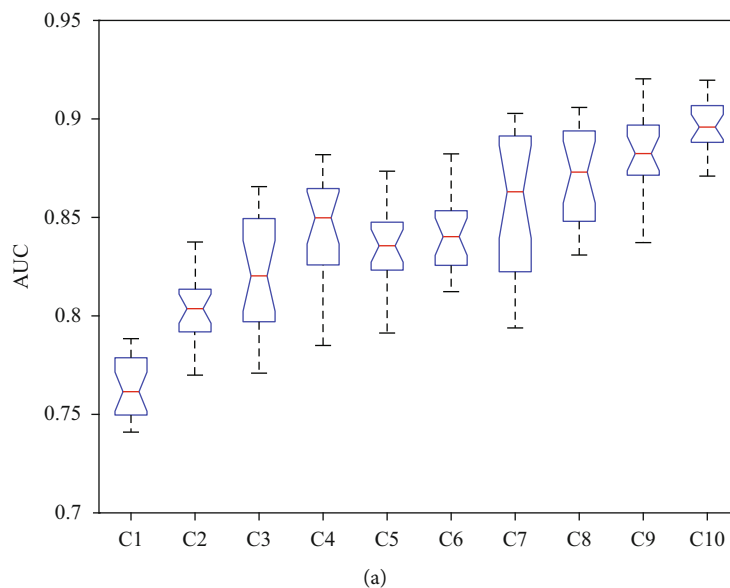
included the MEI computed from multichannel recordings and obstetrical data, significantly outperformed all other classifiers.

The sensitivity of the SMOTE technique to the developed classifiers' generalization capability was also analyzed. For this, 10 balanced datasets were obtained by oversampling the minority class by SMOTE. Figure 6 shows the mean and standard deviation of the average AUC of the ROC curve of the different classifiers for the ten balanced test datasets. Again, the highest average AUC was achieved when using the MEI extracted from the three channels together with obstetrical data. Regardless of the input features used to design the classifier, only small variations in the classifier performance were found between the different SMOTE datasets. In general, the variability of the classifier performance ranged from 0.8% to 1.7%, suggesting that the proposed models were insensitive to the artificial data added to the database.

4. Discussion

The present work focused on the development of ANN-based classifiers for predicting PL, one of the greatest challenges in the field of obstetrics. For this purpose, a set of temporal, spectral, nonlinear, and synchronization parameters were extracted from EHG recordings. Our results showed that the different classifiers developed reached an AUC of over 90% for both the training and validation datasets, comparable to those reported by other authors that attempted to predict PL using the same database [31]–[34, 51]. Different works have proven that, regardless of the input features (temporal, spectral, and nonlinear parameters) extracted from the raw EHG record and from the intrinsic mode functions after the application of empirical mode decomposition using different classifier algorithms (ANN, SVM, AdaBoost, polynomial classifiers, etc.), are able to learn the underlying data structure of the data and therefore to reach similar results for training and validation data [31–34, 51]. Nevertheless, the results obtained in the present work for the different classifiers reveal that the performance on the test dataset was significantly worse than those obtained on the training and validation datasets, which may reveal a possible overfitting phenomenon. These results highlight

(a)

(b)

FIGURE 5: Box and whisker plots of the AUCs obtained on the test datasets by the different classifiers in the 30 iterations on each of the 10 SMOTE-balanced datasets (a). P value associated with each comparison between AUC distributions of the different classifiers when ANOVA analysis is performed (b) (bold figures indicate statistically significant difference, $P < 0.05$).

the importance of preserving an independent data test partition to determine the generalization capability of the model when facing new data that the classifier has never "seen" [35, 36], which is of special relevance for the transfer of the EHG technique to clinical settings.

We also compared the different classifier metrics using EHG features extracted from several channels. The results showed that the EHG S3 recording channel contains better information for accurately predicting PL than the other single channels, which is in agreement with the findings reported by other authors [20, 32]. This could be related to the electrode positions when acquiring uterine myoelectrical activity. In this regard, the S3 sensing electrodes were positioned halfway between the fundus and symphysis [20] and could therefore pick up the electrical activity from more uterine muscle cells than other channels, especially in recordings made prior to week 26 of gestation.

Unlike other authors who discarded the information from other channels [31, 33, 34], we tested and compared two ways of combining the information from multichannel

recordings. EHG features extracted from each of the 3 single channels were first fed to the classifiers C7 and C8, and then, the MEI of EHG features computed on multichannel recordings were also used as input features (C9 and C10). Our results showed that the C9 and C10 classifiers' performance were significantly better than those of C7 and C8, respectively, which may be related to different factors. First, the MEI from multichannel recording "averages" the information of the whole uterus, which can be more reliable and robust than the same parameter obtained from a single channel, which provides information on the activity adjacent to the sensing electrode [55]. Moreover, including synchronization parameters significantly contributes to more accurate labor predictions. These results agree with a previous study that showed that the use of intensity, excitability, and synchronization MEIs from multichannel EHG recordings and their combination into a global efficiency index improves the ability to discriminate between women who will deliver in less than 7/14 days and those who give birth in a longer period [55].
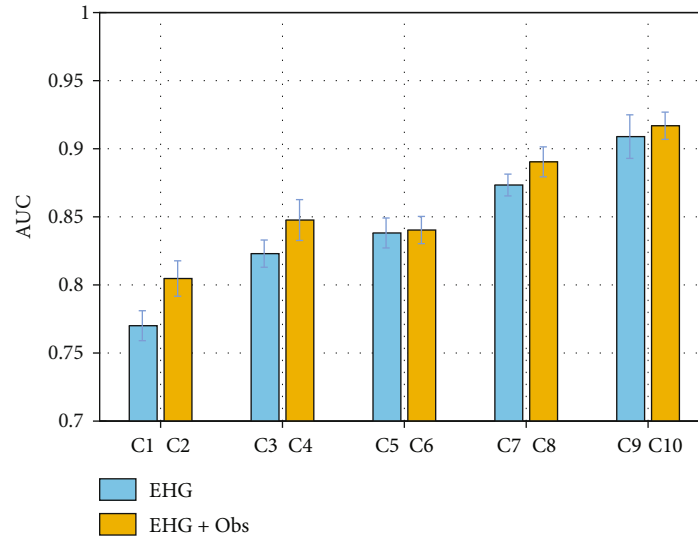
FIGURE 6: Mean AUC values obtained for the test partitions for each classifier when computed over the 10 SMOTE datasets. Error bars represent the deviation between the different SMOTE datasets.

Unlike this previous work, in which the different information was combined by a fixed formula [55], in the present work, the ANN was responsible for carrying out the nonlinear transformation and assigning the appropriate weight to each EHG feature (intensity, excitability, and synchronization) to achieve the best classifier performance. This is more suitable in the context of classifier designs than providing a single global feature that combines the information according to a predefined formula. On the other hand, instead of computing the MEI of a nonlinear parameter such as sample entropy to obtain more reliable information from multichannel recordings, other authors propose the use of multivariate sample entropy [51]. Further work needs to be done on comparing their ability to discriminate PL from term labor women and their computational cost.

Our results also show that including obstetrical data slightly improved classifier performance. The increase in the AUC of the ROC curve was only about 2%, which is considerably less than those reported by Fergus et al., who used different classifiers for predicting PL with the same database [32]. In this latter case, the AUC of the polynomial classifier ROC curve improved from 86% to 95% when additional obstetrical data was added to the input features, while the AUC improvement was only about 4% (from 89% to 93%) for the decision tree classifier [32]. This difference may be due to different factors. Firstly, the authors used 5-fold cross validation to evaluate the developed predictive model, and no testing data was preserved to determine their generalization capability. Furthermore, we believe that the slight improvement in classifier metrics associated with obstetrical data obtained in this work is related to the fact that the obstetrical data provided in the database only contains some of the premature labor risk factors [2, 62] so that no direct measurement of labor proximity was included. Adding other obstetrical measurements, such as cervical length, fetal

fibronectin, and/or interleukin 6, which have been shown to provide relevant information for PL prediction [4, 9, 10], could significantly improve classifier performance when used together with EHG features.

On the other hand, one of the common problems in the development of classifiers in biomedical engineering is a data imbalance between different classes. In our case, only 11-13% of women who had routine controls delivered prematurely. Data imbalance may give rise to a bias in the classifier learning algorithm to strongly learn from the majority class but to a lesser degree from the minority class [61]. Naeem et al. attempted to predict PL using EHG features fed to ANN and achieved a classifier accuracy of 92.3%, while the positive predictive value was about 42.1% [30]. In other words, its diagnosis value lies in its negative predictive value. In this work, we used the SMOTE technique to mitigate the unbalanced data learning problem, while other authors preferred to use other oversampling techniques such as ADASYN for this purpose [31, 51]. We believe that similar results would be obtained if using ADASYN for oversampling the minority class [63]. The classifier performance sensitivity to the oversampling technique was also analyzed. We found that the proposed models were insensitive to the artificial data added to the database, suggesting that the models can be generalizable as long as the real data of the minority class are statistically representative of this class. Nevertheless, the number of the women who finally delivered prematurely in the TPEHG database is relatively small, and therefore, the question of whether the provided sample is statistically representative of the population of preterm births remains unknown. The scientific community must therefore make a greater effort to create a database large enough to obtain reliable results for PL prediction and improve the transferability of the EHG technique to clinical practice. At the same time, other ways of mitigating the influence of unbalanced datasets can be

tested and compared with the oversampling technique. In this respect, the weighted classifiers, which assign more weight to the minority class in their cost function, which would force them to learn not only the underlying data structure of the majority class but also that of the minority class, would contribute to the development of more reliable classifiers for predicting PL [61].

ANN output currently consists of a value between [-1 and 1] (since the hyperbolic tangent is used as an activation function). This can be converted to a discrete value by setting a threshold to maximize sensitivity and specificity, which in clinical practice could be turned into a simple probability with a confidence interval. Once the ability of the classifiers that include multichannel EHG features and obstetrical information has been assessed in a large database, the next step in the development of a clinically viable decision support system for preterm prediction will be to implement the one with the best performance on an embedded system such as DSP or FPGA (software for EHG feature extraction and dimensionality reduction by PCA and trained ANN) and a user-friendly interface for clinicians. The use of such a decision support system will also require a specific protocol that includes multichannel EHG acquisition and obstetrical information.

5. Conclusions

Predicting PL is still a major challenge in obstetrics, and reliable tools that improve actual prediction capacity are required. We developed and compared the performance of ANN-based classifiers for PL prediction using EHG parameters extracted from single- and multichannel recordings together with obstetrical data.

Firstly, all the classifiers developed, regardless of their input features, reached high metrics for the train and validation datasets (AUCs over 90%). However, the results of the test datasets showed that generalization capacity varies remarkably.

As far as we know, this is the first time that mean efficiency indexes and information on signal synchronization estimated by NPCMI have been used to predict preterm labor. Since the objective was to obtain generalizable predictive models, the performance of the different classifiers was compared not only on the training and validation data set but also on an independent set of test data. The performance of ANN-based classifiers for preterm birth prediction using single-channel and multichannel EHG information, as well as uterine contractile efficiency indexes, has also been compared for the first time. Single-channel classifier performance was highly sensitive to electrode location, while those which combined the information from three EHG channels provided better AUC values. The classifier that used the mean efficiency indexes and obstetrical information yielded the best classifier performance metrics, achieving an AUC value in the test datasets of $91.1 \pm 2.5\%$. These results show that mean efficiency indexes computed from multichannel EHG recordings and obstetrical information could be powerful tools for obtaining generalizable and accurate PL classifiers and could be applied in clinical practice.

References

[1] S. Beck, D. Wojdyla, L. Say et al., "The worldwide incidence of preterm birth: a systematic review of maternal mortality and morbidity," *Bulletin of the World Health Organization*, vol. 88, no. 1, pp. 31–38, 2010.

[2] R. L. Goldenberg, J. F. Culhane, J. D. Iams, and R. Romero, "Epidemiology and causes of preterm birth," *The Lancet*, vol. 371, no. 9606, pp. 75–84, 2008.

[3] S. Petrou, "The economic consequences of preterm birth duringthe first 10 years of life," *BJOG*, vol. 112, pp. 10–15, 2005.

[4] R. E. Behrman and A. S. Butler, *Preterm Birth: Causes, Consequences, and Prevention*, National Academies Press, Washington DC, 2007.

[5] J. Zeitlin, K. Szamotulska, N. Drewniak et al., "Preterm birth time trends in Europe: a study of 19 countries," *BJOG*, vol. 120, no. 11, pp. 1356–1365, 2013.

[6] K. E. Fitzpatrick, D. Tuffnell, J. J. Kurinczuk, and M. Knight, "Pregnancy at very advanced maternal age: a UK population-based cohort study," *BJOG*, vol. 124, no. 7, pp. 1097–1106, 2017.

[7] D. M. Haas, T. Benjamin, R. Sawyer, and S. K. Quinney, "Short-term tocolytics for preterm delivery – current perspectives," *International Journal of Women's Health*, vol. 6, no. 1, pp. 343–349, 2014.

[8] R. E. Garfield, W. L. Maner, H. Maul, and G. R. Saade, "Use of uterine EMG and cervical LIF in monitoring pregnant patients," *BJOG*, vol. 112, no. 1, pp. 103–108, 2005.

[9] T. Y. Euliano, M. T. Nguyen, S. Darmanjian et al., "Monitoring uterine activity during labor: a comparison of 3 methods," *American Journal of Obstetrics and Gynecology*, vol. 208, no. 1, pp. 66.e1–66.e6, 2013.

[10] R. E. Garfield and W. L. Maner, "Biophysical methods of prediction and prevention of preterm labor: uterine electromyography and cervical light-induced fluorescence – new obstetrical diagnostic techniques," in *Preterm Birth*, pp. 131–144, 2006, https://www.semanticscholar.org/paper/Biophysical-methods-of-prediction-and-prevention-of-Garfield-Maner/4a3beef091718708a38cb6a4128a0de096b7b6e1.

[11] D. Devedeux, C. Marque, S. Mansour, G. Germain, and J. Duchêne, "Uterine electromyography: a critical review," *American Journal of Obstetrics and Gynecology*, vol. 169, no. 6, pp. 1636–1653, 1993.

[12] A. Chkeir, M. J. Fleury, B. Karlsson, M. Hassan, and C. Marque, "Patterns of electrical activity synchronization in the pregnant rat uterus," *BioMedicine*, vol. 3, no. 3, pp. 140–144, 2013.

[13] A. R. Fuchs, F. Fuchs, P. Husslein, and M. S. Soloff, "Oxytocin receptors in the human uterus during pregnancy and parturition," *American Journal of Obstetrics and Gynecology*, vol. 150, no. 6, pp. 734–741, 1984.

[14] N. Tezuka, M. Ali, K. Chwalisz, and R. E. Garfield, "Changes in transcripts encoding calcium channel subunits of rat myometrium during pregnancy," *American Journal of Physiology-Cell Physiology*, vol. 269, no. 4, pp. C1008–C1017, 1995.

[15] H. Honest, L. M. Bachmann, R. Sundaram, J. K. Gupta, J. Kleijnen, and K. S. Khan, "The accuracy of risk scores in predicting preterm birth—a systematic review," *Journal of Obstetrics and Gynaecology*, vol. 24, no. 4, pp. 343–359, 2004.

[16] M. Hassan, J. Terrien, A. Alexandersson, C. Marque, and B. Karlsson, "Improving the classification rate of labor vs. normal pregnancy contractions by using EHG multichannel recordings," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 4642–4645, Buenos Aires, Argentina, 2010.

[17] R. E. Garfield and R. H. Hayashi, "Appearance of gap junctions in the myometrium of women during labor," *American Journal of Obstetrics and Gynecology*, vol. 140, no. 3, pp. 254–260, 1981.

[18] J. Garcia-Casado, Y. Ye-Lin, G. Prats-Boluda, J. Mas-Cabo, J. Alberola-Rubio, and A. Perales, "Electrohysterography in the diagnosis of preterm birth: a review," *Physiological Measurement*, vol. 39, no. 2, article 02TR01, 2018.

[19] W. L. Maner, R. E. Garfield, H. Maul, G. Olson, and G. Saade, "Predicting term and preterm delivery with transabdominal uterine electromyography," *Obstetrics and Gynecology*, vol. 101, no. 6, pp. 1254–1260, 2003.

[20] G. Fele-Žorž, G. Kavšek, Ž. Novak-Antolič, and F. Jager, "A comparison of various linear and non-linear signal processing techniques to separate uterine EMG records of term and preterm delivery groups," *Medical & Biological Engineering & Computing*, vol. 46, no. 9, pp. 911–922, 2008.

[21] K. Horoba, J. Jezewski, A. Matonia, J. Wrobel, R. Czabanski, and M. Jezewski, "Early predicting a risk of preterm labour by analysis of antepartum electrohysterograhic signals," *Biocybernetics and Biomedical Engineering*, vol. 36, no. 4, pp. 574–583, 2016.

[22] M. P. G. C. Vinken, C. Rabotti, M. Mischi, and S. G. Oei, "Accuracy of frequency-related parameters of the electrohysterogram for predicting preterm delivery," *Obstetrical & Gynecological Survey*, vol. 64, no. 8, pp. 529–541, 2009.

[23] A. Diab, M. Hassan, C. Marque, and B. Karlsson, "Performance analysis of four nonlinearity analysis methods using a model with variable complexity and application to uterine EMG signals," *Medical Engineering & Physics*, vol. 36, no. 6, pp. 761–767, 2014.

[24] M. Hassan, J. Terrien, C. Marque, and B. Karlsson, "Comparison between approximate entropy, correntropy and time reversibility: application to uterine electromyogram signals," *Medical Engineering & Physics*, vol. 33, no. 8, pp. 980–986, 2011.

[25] A. Lemancewicz, M. Borowska, P. Kuć et al., "Early diagnosis of threatened premature labor by electrohysterographic recordings – the use of digital signal processing," *Biocybernetics and Biomedical Engineering*, vol. 36, no. 1, pp. 302–307, 2016.

[26] J. Vrhovec, A. Macek-Lebar, and D. Rudel, "Evaluating uterine electrohysterogram with entropy," in *11th Mediterranean Conference on Medical and Biomedical Engineering and Computing 2007. IFMBE Proceedings*, T. Jarm, P. Kramar, and A. Zupanic, Eds., pp. 144–147, Springer, Berlin, Heidelberg, 2007.

[27] N. Nader, M. Hassan, W. Falou et al., "Classification of pregnancy and labor contractions using a graph theory based analysis," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2876–2879, Milan, Italy, 2015.

[28] M. Lucovnik, W. L. Maner, L. R. Chambliss et al., "Noninvasive uterine electromyography for prediction of preterm delivery," *American Journal of Obstetrics and Gynecology*, vol. 204, no. 3, pp. 228.e1–228.e10, 2011.

[29] O. Most, O. Langer, R. Kerner, G. Ben David, and I. Calderon, "Can myometrial electrical activity identify patients in preterm labor?," *American Journal of Obstetrics and Gynecology*, vol. 199, no. 4, pp. 378.e1–378.e6, 2008.

[30] S. M. Naeem, A. F. Seddik, and M. A. Eldosoky, "New technique based on uterine electromyography nonlinearity for preterm delivery detection new technique based on uterine electromyography nonlinearity for preterm delivery detection," *Journal of Engineering and Technology Research*, vol. 6, no. 7, pp. 107–114, 2014.

[31] U. R. Acharya, V. K. Sudarshan, S. Q. Rong et al., "Automated detection of premature delivery using empirical mode and wavelet packet decomposition techniques with uterine electromyogram signals," *Computers in Biology and Medicine*, vol. 85, pp. 33–42, 2017.

[32] P. Fergus, P. Cheung, A. Hussain, D. Al-Jumeily, C. Dobbins, and S. Iram, "Prediction of preterm deliveries from EHG signals using machine learning," *PLoS One*, vol. 8, no. 10, article e77154, 2013.

[33] P. Fergus, I. Idowu, A. Hussain, and C. Dobbins, "Advanced artificial neural network classification for detecting preterm births using EHG records," *Neurocomputing*, vol. 188, pp. 42–49, 2016.

[34] P. Ren, S. Yao, J. Li, P. A. Valdes-Sosa, and K. M. Kendrick, "Improved prediction of preterm delivery using empirical mode decomposition analysis of uterine electromyography signals," *PLoS One*, vol. 10, no. 7, 2015.

[35] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[36] S. M. Weiss and C. A. Kulikowski, *Computer Systems that Learn: Classification and Prediction Methods from Statistics Nets, Machine Learning, and Expert Systems*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1991.

[37] W. L. Maner and R. E. Garfield, "Identification of human term and preterm labor using artificial neural networks on uterine electromyography data," *Annals of Biomedical Engineering*, vol. 35, no. 3, pp. 465–473, 2007.

[38] A. Smrdel and F. Jager, "Separating sets of term and pre-term uterine EMG records," *Physiological Measurement*, vol. 36, no. 2, pp. 341–355, 2015.

[39] M. Akay and IEEE/Engineering in Medicine and Biology Society, *Nonlinear Biomedical Signal Processing, vol. I*, the Institute of Electrical and Electronics Engineers, Inc., 2000.

[40] S. Aditya and D. N. Tibarewala, "Comparing ANN, LDA, QDA, KNN and SVM algorithms in classifying relaxed and stressful mental state from two-channel prefrontal EEG data,"

International Journal of Artificial Intelligence and Soft Computing, vol. 3, no. 2, p. 143, 2012.

[41] Q. Li, C. Rajagopalan, and G. D. Clifford, "A machine learning approach to multi-level ECG signal quality classification," *Computer Methods and Programs in Biomedicine*, vol. 117, no. 3, pp. 435–447, 2014.

[42] Z. Li, Q. Zhang, and X. Zhao, "Performance analysis of K-nearest neighbor, support vector machine, and artificial neural network classifiers for driver drowsiness detection with different road geometries," *International Journal of Distributed Sensor Networks*, vol. 13, no. 9, 2017.

[43] H. S. N. Murthy and D. M. Meenakshi, "ANN, SVM and KNN classifiers for prognosis of cardiac ischemia- a comparison," *Bonfring International Journal of Research in Communication Engineering*, vol. 5, no. 2, pp. 7–11, 2015.

[44] G. Zhang, B. E. Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks:: The state of the art," *International Journal of Forecasting*, vol. 14, no. 1, pp. 35–62, 1998.

[45] J. Ren, "ANN vs. SVM: which one performs better in classification of MCCs in mammogram imaging," *Knowledge-Based Systems*, vol. 26, pp. 144–153, 2012.

[46] H. Maul, W. L. Maner, G. Olson, G. R. Saade, and R. E. Garfield, "Non-invasive transabdominal uterine electromyography correlates with the strength of intrauterine pressure and is predictive of labor and delivery," *The Journal of Maternal-Fetal & Neonatal Medicine*, vol. 15, no. 5, pp. 297–301, 2004.

[47] J. Terrien, C. Marque, and B. Karlsson, "Spectral characterization of human EHG frequency components based on the extraction and reconstruction of the ridges in the scalogram," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1872–1875, Lyon, France, 2007.

[48] J. Mas-Cabo, G. Prats-Boluda, A. Perales, J. Garcia-Casado, J. Alberola-Rubio, and Y. Ye-Lin, "Uterine electromyography for discrimination of labor imminence in women with threatened preterm labor under tocolytic treatment," *Medical & Biological Engineering & Computing*, vol. 57, no. 2, pp. 401–411, 2019.

[49] R. E. Garfield and W. L. Maner, "Physiology and electrical activity of uterine contractions," *Seminars in Cell & Developmental Biology*, vol. 18, no. 3, pp. 289–295, 2007.

[50] X. S. Zhang, R. J. Roy, and E. W. Jensen, "EEG complexity as a measure of depth of anesthesia for patients," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 12, pp. 1424–1433, 2001.

[51] M. U. Ahmed, T. Chanwimalueang, S. Thayyil, and D. P. Mandic, "A Multivariate multiscale fuzzy entropy algorithm with application to uterine EMG complexity analysis," *Entropy*, vol. 19, no. 1, p. 2, 2017.

[52] M. Brennan, M. Palaniswami, and P. Kamen, "Do existing measures of Poincaré plot geometry reflect nonlinear features of heart rate variability?," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 11, pp. 1342–1347, 2001.

[53] W. Klonowski, "Application of new non-linear dynamics methods in biosignal analysis," in *Proceedings of the World Medical Conference*, pp. 180–187, Prague, Czech Republic, 2007.

[54] D. Cui, W. Pu, J. Liu et al., "A new EEG synchronization strength analysis method: S-estimator based normalized weighted-permutation mutual information," *Neural Networks*, vol. 82, pp. 30–38, 2016.

[55] J. Mas-Cabo, Y. Ye-Lin, J. Garcia-Casado, J. Alberola-Rubio, A. Perales, and G. Prats-Boluda, "Uterine contractile efficiency indexes for labor prediction: a bivariate approach from multi-channel electrohysterographic records," *Biomedical Signal Processing and Control*, vol. 46, pp. 238–248, 2018.

[56] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[57] J. P. Cunningham and Z. Ghahramani, "Linear dimensionality reduction: survey, insights, and generalizations," *Journal of Machine Learning Research*, vol. 16, pp. 2859–2900, 2014.

[58] R. Caruana, S. Lawrence, and L. Giles, "Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping," in *NIPS'00 Proceedings of the 13th International Conference on Neural Information Processing Systems*, pp. 381–387, Denver, CO, USA, 2000.

[59] S. M. Naeem, A. F. Ali, and M. A. Eldosoky, "Kl. Comparison between using linear and non-linear features to classify uterine electromyography signals of term and preterm deliveries," in *2013 30th National Radio Science Conference (NRSC)*, pp. 492–502, Cairo, Egypt, 2013.

[60] L. Bottou, "Stochastic Gradient Descent Tricks," in *Neural Networks: Tricks of the Trade*, G. Montavon, G. B. Orr, and K. R. Müller, Eds., vol. 7700 of Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2012.

[61] G. Weiss, K. McCarthy, and B. Zabar, "Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?," *International Conference on Data Mining*, vol. 7, pp. 35–41, 2007.

[62] K. Lim, K. Butt, and J. M. Crane, "No. 257-Ultrasonographic cervical length assessment in predicting preterm birth in singleton pregnancies," *Journal of Obstetrics and Gynaecology Canada*, vol. 40, no. 2, pp. e151–e164, 2018.

[63] G. Goos, L. Maguire, Y. Li, and S. McLoone, "Evaluation of Sampling Methods for Learning from Imbalanced Data," in *Intelligent computing theories*, D. S. Huang, V. Bevilacqua, J. C. Figueroa, and P. Premaratne, Eds., pp. 392–401, Springer, 2013.