# Realistic Preterm Prediction Based on Optimized Synthetic Sampling of EHG Signal

Jinshan Xu, Zhenqin Chen, Yanpei Lu, Xi Yang, Jinpeng Zhang, Alain Pumir

# Realistic Preterm Prediction Based on Optimized Synthetic Sampling of EHG Signal

Jinshan XU[a,b], Zhenqin Chen[a], Jinpeng Zhang[a], Yanpei Lu[a], Xi Yang[a,*] and Alain Pumir[c]

[a]College of Computer Science, Zhejiang University of Technology, Hangzhou 310023, China
[b]Research Center for AI Social Experiment, Zhejiang Lab, Hangzhou 311321, China
[c]Laboratoire de Physique, ENS-Lyon, Lyon 69007, France

ARTICLE INFO

*Keywords*:
preterm prediction
uterine electrohysterogram
synthetic sampling
sample balance coefficient

ABSTRACT

Preterm labor is the leading cause of neonatal morbidity and mortality and has attracted research efforts from many scientific areas. The relationship between uterine contraction and the underlying electrical activity makes the uterine electrohysterogram (EHG) a promising direction for detecting and predicting preterm birth. Due the scarcity of EHG signals, especially those leading to preterm birth, synthetic algorithms are used to generate artificial samples of preterm birth type to eliminate bias in prediction towards normal delivery, at the expense of reducing feature effectiveness in automatic preterm birth detection based on machine-learning. To address this issue, we quantify the effect of synthetic samples (balance coefficient) on feature effectiveness and form a general performance metric by utilizing several feature scores with relevant weights describing their contributions to class segregation. In combination with the activation/inactivation functions that characterize the effect of the abundance of training samples on the accuracy of the prediction of on- or off-time delivery, we obtain an optimal sampling balance coefficient that optimizes the effect of synthetic samples in removing the bias towards the majority and minimizes the side-effect of reducing the importance of features. More realistic prediction accuracy was achieved through a series of numerical tests, utilizing the publicly available TPEHG database, demonstrating the effectiveness of the proposed method.

## 1. Introduction

Preterm births, defined as babies born at gestational age of less than 37 weeks, represent a major and growing challenge for public health systems. Every year, nearly 15 millions babies, or about 10% of all births worldwide, are born prematurely. Approximately one million of these premature babies die from complications that follow preterm birth [1]. Currently, the lack of comprehensive knowledge of the mechanisms that trigger uterine contraction prevents effective early-stage treatment of preterm birth. Once delivery has started, it can no longer be interrupted. Therefore, early detection and preventive treatments are a promising direction to prevent preterm births. Commonly used methods of preterm diagnosis include Toco-grametry, intra-uterine pressure catheter, fetal fibronectin, cervical length measurement etc, but none of these methods provides reliable results [2].

The expulsion of a fetus is a direct consequence of the strong periodic uterine contractions, which are the result of the generation and propagation of action potentials [3]. The corresponding electric signals can be recorded by electrodes placed on the abdomen of pregnant women, using the electrohysterogram (EHG) technique. Because of the close relationship between uterine contraction and the underlying electrical activities, EHG provides a new direction for the development of preterm diagnosis method [4, 5]. For this reason, several databases of carefully annotated EHG signals recorded from pregnant women at different stages of pregnancy have been made publicly available [6, 7, 8]. For example, the PhysioNet TPEHG database contains recordings from 300 pregnant women at the gestation age of about 26 weeks. Unlike the Icelandic Electrohysterogram database, which uses 16 electrodes to record the electrical signals, a 4-electrode configuration (E1,E2,E3,E4) was adopted for TPEHG. To avoid common noises from external sources like undesired body movements, only differences between adjacent two electrodes, i.e., S1=E2-E1, S2=E2-E3, S3=E4-E3 were stored with a sampling frequency of 20 Hz and filtered with different bandwidths to reduce noises.

The availability of datasets and recent progress in machine learning have led to a number of new methods of preterm diagnosis [9, 10, 11]. Overall, preterm diagnosis can be categorized as a classification problem, i.e. deciding or predicting whether a patient (pregnant woman) is at risk for preterm birth based on a set of physical examination data (sample) and the features contained therein. It is natural to expect that both the abundance of the sample in terms of different classes and the quality of the features (difference between classes) are essential to obtain satisfactory classification results.

In addition to the sample abundance and feature quality already mentioned, the distribution pattern of samples among classes is also an important factor affecting classification performance [12, 13]. Although there are millions of preterm deliveries worldwide, the proportion of preterm birth is quite small compared to the total number of births. This fact is reflected in the composition of the TPEHG database, in which there are only 38 EHG samples from premature deliveries and 262 samples from normal term births. Due to the large difference between the number of samples from preterm and normal term infants, the application of conventional machine learning algorithms with such extremely im-

*Corresponding author: X. Yang (xyang@zjut.edu.cn)
ORCID(s):

balanced data will tend to classify the minority of patients into the majority class, i.e., there is a bias towards the majority [14], which is likely to lead to an inaccurate diagnosis result [15].

Learning from imbalanced dataset is a very active research topic in the field of machine learning [16, 17]. The state-of-the-art research methods to deal with imbalanced data mining problem can be categorized into two directions: 1) over-sampling the minority class or under-sampling the majority one in order to compensate the imbalance of samples between classes to be identified; 2) synthesizing artificial samples from the minority class [18]. The idea behind under-sampling is quite simple. Only a fraction of the majority data is used. In practical applications, special attention must be paid to selecting the right samples so that the distribution pattern in the feature space remains unchanged [13]. Although many studies have documented the effectiveness of under-sampling [19], its use for small datasets is excluded. To be more precise, under-sampling could significantly reduce the number of samples to be used in training the learning model, which may lead to underestimation. Oversampling simply uses minority samples multiple times. This makes the data set highly correlated, resulting in magnified feature variations. Consequently, high computational costs are required during training process and the trained classifiers might have the overstimation issue [20]. In particular, it is inappropriate to use very correlated signals, recorded simultaneously from different electrodes .

On the other hand, synthetic sampling with data generation methods consists in generating synthetic data that originate from the minority class. The synthesis method mimics the random distribution of the sample data in the minority's feature space, so that the generated samples are assumed to be close to the actual distribution of the minority in its feature space. Including these samples in the minority training set eliminates the original imbalance problem, and therefore removes the classification bias towards the majority. Frequently used synthetic algorithms such as SMOTE [21] and ADASYN [22], have exhibited some advantages in real-world applications of preterm diagnosis [10] and others [17, 23].

In addition to the abundance of training examples, the quality of the features is another key factor that contributes to the precision of the trained classifiers [24]. For this reason, algorithms are proposed to extract new features that improve the classification performance [25, 26, 27, 28]. It should be noted that the effect of imbalance may worsen when new features are adopted in the training process [29]. Also, when more synthetic/artificial minority class data is generated, the representation ability of the features may change. In addition, as the synthetic samples increase, the noise in the original samples may increase. When the classifier is trained with these datasets, it would overfit [30]. Therefore, adding synthetic samples may affect the feature quality in a complex way. As a result, the classification performances may change. Although there has been work on optimizing synthetic algorithms [31, 32], to the best of our knowledge, few studies address the optimal number of synthetic samples on classifier accuracy. For this reason, it is necessary to explore the relationship between the amount of synthetic data and the quality of features and to find a unified formulation.

In this paper, we examine the relationship between bias elimination and feature reduction when synthetic samples are introduced in the training process. We propose an optimal synthesis strategy for minority samples by quantifying the variation in feature importances in terms of the ratio between available samples in different classes. Our manuscript is organized as follows: In Section 2, we present some basic elements that are important for the problem, and analyze the underlying principles of the widely used synthetic sample generation algorithms, SMOTE and ADASYN, which we use to show the importance of finding an optimal sampling ratio between classes when learning from imbalanced dataset. In Section 3, we further quantify the effect of synthetic samples and formulate the problem of determining the optimal sample balance coefficient. In Section 4, we verify the effectiveness of the proposed method by applying it to EHG based preterm birth prediction numerically. Section 5 concludes the paper.

## 2. Problem Statement

As explained in the introduction, the strong imbalance between samples in pathological and normal classes in datasets leads to an unsatisfactory classification performance, especially for the pathological class. To avoid this, the preferable machine-learning-based algorithms typically introduce a certain amount of synthetic preterm sample data to mitigate the bias towards majority (normal delivery). At the same time, however, the possibility of misclassification of term samples increases. To ensure the representativeness of the artificial samples, synthetic algorithms are applied to the entire original minority, i.e., before partitioning into training and testing subsets. This leads to a high correlation between training and testing samples, resulting in higher values of preterm predictive performance than for the original sample. In this scenario, determining the optimal number of synthetic samples is crucial, not only for better performance of a machine-learning based diagnosis method, but also for more accurate validation of classifier performance.

### 2.1. Sample balance coefficient and feature score

As stated in Section 1, the balance and abundance of training samples of different classes are essential to the performance of the classifier. The TPEHG database contains many more samples corresponding to normal term than to preterm births, in the ratio (262:38). To train classifiers, it is essential to generate synthetic (artificial) preterm samples from the original minority class.

The enrichment of minority samples by applying data synthesis techniques can improve the classification performance to some extent. However, it could also degrade the classifier performance, since the synthetic data could change the original pattern of sample distribution in the feature space, i.e., the boundary in feature space between different classes

could be blurred. To better quantify the separability of classes in the different feature spaces, we introduce the following feature score $f_s^i$ defined as in [33],

$$f_s^i = \frac{\left(\bar{x}_i^+ - \bar{x}_i\right)^2 + \left(\bar{x}_i^- - \bar{x}_i\right)^2}{\frac{1}{n^+-1}\sum_{k=1}^{n^+}\left(x_{k,i}^+ - \bar{x}_i^+\right)^2 + \frac{1}{n^--1}\sum_{k=1}^{n^-}\left(x_{k,i}^- - \bar{x}_i^-\right)^2} \tag{1}$$

where $x_{k,i}^+$ and $x_{k,i}^-$ denote the measured physical value of the feature $i$ of sample $k$, which is in the positive (minority or preterm birth) class and negative (majority or term ) class, respectively. The quantity $\bar{x}_i$ denotes the average value among all samples, $\bar{x}_i^+$ is the average value of all $n_+$ positive (minority or preterm birth) samples, and $\bar{x}_i^-$ is the average value of $n_-$ negative samples. The numerator of Eq.(1) measures the distance between the centers of the distributions of the two classes, while the denominator gives the dispersion of the samples within each class. In general, $f_s^i$ expresses the separation between minority class (+) and majority class (-) in the space corresponding to the specific feature, $i$. The larger $f_s^i$ is, the more likely it is that the feature contributes to class separation, and samples of non-separable classes gives $f_s^i$ =0. A small value of $f_s^i$, however, does not imply that classes are non-separable, as classifiers can map the current feature in a higher dimensional space.

It is worth mentioning that the feature score explicitly depends on the size of testing samples. We introduce the sample balance coefficient $\alpha$ by,

$$\alpha = \frac{n_+}{n_-} \tag{2}$$

where $n_+$ and $n_-$ are the numbers of samples in the minority and majority class, respectively, after generating additional samples with the synthetic algorithms. Note that $n_+$ includes the number of synthesized samples. In addition, for each specified classification problem, different features that are abstracted from samples jointly contribute to the final classification result. According to (2) and (1), it is appropriate to define the global feature score $F_{score}$ as the weighted sum of different feature scores $f_s^i$, i.e.,

$$F_{score}(\alpha) = \sum_i^N w_i \cdot f_s^i \qquad \sum w_i = 1 \tag{3}$$

where $\alpha$ is the sample balance coefficient defined in (2), the weights $w_i \in [0, 1]$ are introduced to represent the importance of feature $i$ to the classification, and $N$ is the number of features used in the final classification. By construction, the definition of (3) links the number of synthetic samples and the quality of the features. It provides a unified performance metric which is essential for further investigations. To proceed, all features are first used to build a forest consisting of a number of decision trees. Each node in the trees is a condition on a single feature, which is designed to well separate the dataset into two subsets. The Gini index measures the global impurity of the subsets at the node, the higher the better [34]. A good feature should help to split a tree node with increased Gini index (Gini information gain). Thus the contribution of a feature (feature importance score) can be quantified as the normalized total information gain over tree nodes in the forest [35].

## 2.2. Size of the synthesized data and distinguishability of the features

After introducing the definitions of the sample balance coefficient $\alpha$ and of the feature score $f_s^i$ / $F_{score}(\alpha)$, it is useful to discuss the properties of the conventional data synthesis algorithms such as SMOTE or ADASYN, and to test their feasibility in the application of preterm birth diagnosis using the TPEHG database.

Although data synthesis algorithms tend to mimic the natural distribution of the sample in its feature space, the method used to synthesize samples of the minority class inevitably has an impact on the ability of features to discriminate different classes. According to the ADASYN algorithm [22], a synthetic sample $s_s$ is generated using a minority sampe $s_i$, and another randomly selected minority one $s_k$ among the $k$ nearest neighbors of $s_i$, as $s_s = s_i + \lambda(s_k - s_i)$, where $\lambda$ is a random number ($\lambda \in [0, 1]$). For each original minority sample $s_i$, this processed will be repeated for $n_{syn}(= G \cdot \frac{\Delta_i}{\sum_i^{n_+} \Delta_i})$ times, where $G$ is the total number of synthetic samples to be generated and $\Delta_i$ is the number of minority samples within the $k^{th}$ nearest neighbors of $s_i$. Manifestly, ADASYN attempts to synthesize more data from minority samples that are surrounded by more majorities [22]. As shown in Fig.1 (a), the synthetic samples are more likely to appear on the left since there are more majority data samples around these minority ones. Although they are intended to facilitate the classification by focusing on samples that are hard to learn from, they would at the same time result in originally separated datasets not being discriminated in feature space.

Contrary to ADASYN, SMOTE does not take into account the surrounding of the minority samples. For each minority sample $s_i$, its $k$ nearest neighbors are first determined and stored in an array. A sample $s_k$ in the array is randomly selected to synthesize an artificial sample as $s_s = s_i + \lambda(s_k - s_i)$, with a random number $\lambda \in [0, 1]$ [21]. As a result, the synthesized samples will concentrate in the region containing more minority samples (see Fig.1(b)), which implies that the original distribution pattern could be better maintained, without degrading the feature's contribution to the classification.

To see the effect of synthetic sampling on the contribution of features to classification, we apply the aforementioned two synthetic algorithms to the TPEHG database of commonly used features obtained directly from PhyioNet (https://www.physionet.org/content/tpehgdb/1.0.1/): 1) the root mean square value of the signal ($rms$); 2) the median ($F_{med}$) and peak ($F_{peak}$) frequency of the power spectrum; 3) the sample entropy of the signal ($E_{samp}$) extracted from each recorded EHG signals. The 4 electrodes configuration

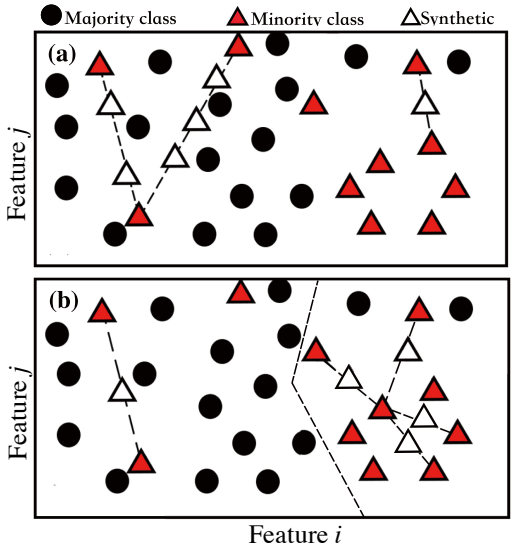


**Figure 1:** Schematic illustration of sample synthetic algorithms. (a) ADASYN tends to synthesize more artificial samples with the original minorities surrounded with more majority samples. It might cripple the original separability of samples. (b) SMOTE algorithm synthesize data point with a randomly chosen minority sample in its closest neighbors. It is easier to keep the original sample distribution in feature space.

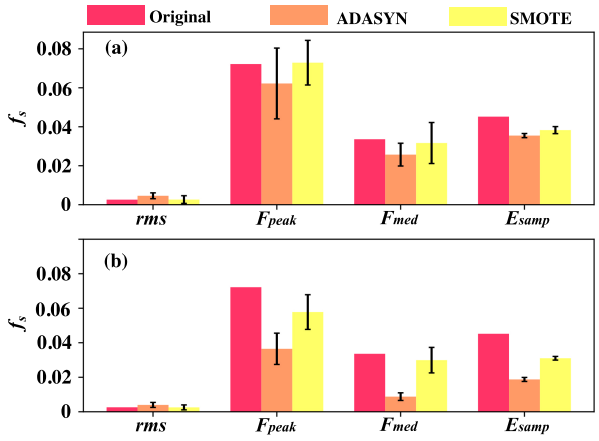


**Figure 2:** Effect of synthetic samples on features' contribution to class separation measured using feature score $f_s$. Both synthetic methods weaken features importance with the increase of synthetic samples. (a) Sample balance coefficient $\alpha = 0.3$, (b) $\alpha = 0.5$

and bipolar signal recording scheme mean that there are 3 EHG signals for each patients. By combining different digital filters, 12 versions of each above features can be extracted. Here, we used signals filtered within the frequency range from 0.3Hz to 4Hz. This is the frequency range of the fast wave. The two components fast, i.e., wave low (FWL) and fast wave high (FLW) are related to the excitability of the uterus and to the burst pattern of the EHG signal [36]. Applying a feature selection algorithm to each of the features extracted from different channels allows us to identify the optimal combination of channels on which we should extract the features. In the current study, S1, S2 are selected to compute $F_{peak}$ and $rms$ respectively, and S3 is used for the calculation of $F_{med}$ and $E_{samp}$. We use those with the gestation age less than 37 weeks as the preterm (minority class) samples.

Fig. 2 shows the variation of the contribution of the features to classification, represented by its $f_s^i$ after synthetic sampling. It can be clearly seen that the peak frequency gives the highest feature score $f_s^i$ among these four features. The effectiveness of this feature for classification has been confirmed by other authors [2, 37, 38]. It is also surprising that both techniques tend to deteriorate the ability of the features to separate the two classes. SMOTE turns out to be better, i.e., the feature scores are higher after applying SMOTE than after applying ADASYN. This is consistent with the previous analysis of data synthesis mechanics of SMOTE and ADASYN.

It is also worth noting that the feature classification ability is sensitive to the number of synthetic samples introduced. This is evident from Fig. 2, which shows the feature scores $f_s^i$ after adding synthetic data with different sample balance coefficients $\alpha$ (see Eq. (2)) $\alpha = 0.3$, panel (a) and $\alpha = 0.5$, panel (b). Adding synthetic samples degrades the ability of the algorithm to distinguish the different classes. However, synthetic samples are required to prevent the classification bias against the minority. Therefore, a trade-off should be found between the number of synthetic samples and the quality of the features to optimize the final performance of the classifiers trained with these data.

## 3. Determination of optimal sample balance coefficient

As discussed in Section 2, synthesizing artificial samples from the minority class is essential for machine-learning-based preterm early diagnosis, and an optimal balance between the synthetic data and feature quality must be found. Intuitively, increasing the number of minority samples by generating synthetic samples should increase the prediction accuracy for the minority class and reduce the bias towards the majority. On the other hand, the prediction precision for the majority class could be decreased if there are too many synthetic samples. Ideally, we would expect a learning system without bias when the sample balance coefficient reaches $\alpha = 1.0$. In real applications, however, due to the imbalance of the available original samples between classes, the optimum may differ from $\alpha = 1$. To proceed, we introduce two functions $C_+$ and $C_-$ describing the putative biases induced by the sampling on the minority and the majority

class, respectively,

$$\begin{aligned} C_+(\alpha) &= \frac{1}{1+\exp(-k(\alpha-1+\overline{\alpha}_0))} \\ C_-(\alpha) &= \frac{1}{1+\exp(k(\alpha-1+\overline{\alpha}_0))} \\ \overline{\alpha}_0 &= \alpha_0(1-\alpha_0) \end{aligned} \tag{4}$$

where $\alpha_0$ is the original sample balance coefficient before any synthetic sample is generated. Manifestly, the parameter $k$ controls how fast the prediction performances on term ($C_-$) and preterm ($C_+$) change with $\alpha$. However, it does not change the relation between $C_-$ and $C_+$ (the interception point). In the present work, we chose the parameter $k = 4$ to facilitate the identification of the optimal balance coefficient $\alpha^*$ defined in Eq( 5), as small $k$ will give a relative flat area of the $F^e_{score}$ near $\alpha^*$. Varying $k$ around the value selected in this work is expected to change only very moderately the quantitative results presented here.

Eq (4) implies that when $\alpha$ is smaller than 1, the majority is well described, and $C_-$ is close to 1. On the other hand, when $\alpha$ is large, the minority is accurately described ($C_+$ is close to 1), but the majority is affected ($C_-$ is reduced). For most of the classification purposes, a high precision on different classes is required (no bias towards any class). To achieve this, for learning problems from unbalanced data samples, one needs to resample the minority training class to match the number of majority, i.e., $\alpha = 1$, in order to eliminate potential bias towards majority. However, recall the synthesis procedure, additional training examples are generated from the original minority data sets and can be considered as a kind of "copy" of the original data. As a consequence, adding too many synthetic samples improves only the precision on the minority class, while degrading the accuracy on the majority classification.

To account for these effects, we introduce the constant $\overline{\alpha}_0$ in the definitions of $C_+$ and $C_-$ in Eq.(4). This equation expresses that the constant $\overline{\alpha}_0$ goes to 0 when there are enough original minority samples ($\alpha_0 \to 1$), so there is no need to synthesize training samples. When learning tasks from imbalanced dataset are encountered ($\alpha_0 < 1$), the optimal balance coefficient $\alpha$ deviates from the ideal value.

To account for the bias described above, we simply multiply the effective feature score, $F_{score}$, as defined in Eq. (3), by $C_+ \times C_-$, to determine an effective score, $F^e_{score}$. Combining this with the requirement for high predictive performance on preterm and normal term birth, we determine the optimal sample balance coefficient $\alpha^*$ as follows,

$$\begin{aligned} \alpha^* &= \text{argmax}_{\alpha\in[0,\text{inf}]} F^e_{score}(\alpha) \\ F^e_{score}(\alpha) &= F_{score}(\alpha)\cdot C_+(\alpha)\cdot C_-(\alpha) \end{aligned} \tag{5}$$

## 4. Empirical verification

The above analysis suggests a way to improve preterm birth diagnosis by determining the optimal sample balance coefficient without unduly affecting the term prediction accuracy. In this section, we provide empirical results to test the effectiveness of the proposed method. In particular, we propose a numerical way to determine the value of $\alpha^*$ numerically. The general procedure is summarized in Fig. 3,

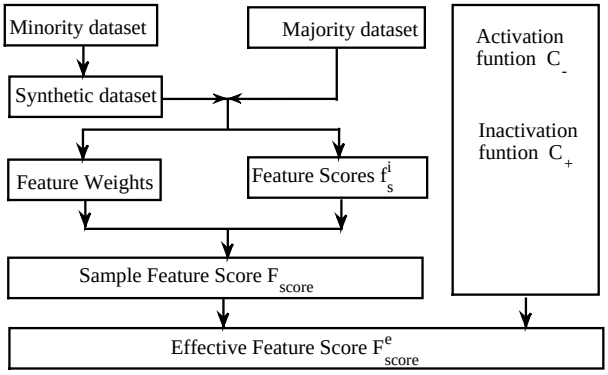


**Figure 3:** Flowchart for determining effective feature scores. After having synthesized artificial samples, calculate the weight $w_i$ and score value $f^i_s$ of each feature. Based on these two values, the weighted feature score $F_{score}$ is defined. Combining the activation/inactivation functions with $F_{score}$ gives effective feature score $F^e_{score}$.

### 4.1. Material and methods

Bearing in mind the purpose of predicting gestational status, simple, but powerful classifers are often used, such as Support Vector Classifier (SVC), Linear Discriminant Classifier (LDC), Decision Tree Classifier (DTC), Gradient Boosting Classifier (GBC) etc., are used to verify the proposed method. To ensure a good performance of the classifier, a grid search (GridSearchCV method in sklearn package) is first performed in parameter space, using accuracy as the criterion to determine their optimal values. It turns out that the SVC should be used with a Gaussian kernel and the hyperparameter $C$ is set to 20 to emphasize on the effect of misclassified data points during training. Following [38], the Holdout Cross-Validation is performed using 80% of the entire dataset for training the classifiers and the remaining 20% for testing, from which we determine the sensitivity (True Positive (preterm) Rate, TPR) and specificity (True Negative (term) Rate, TNR). Note that these two quantities alone cannot well represent our requirements for high performance in both positive and negative prediction. For this reason, we introduce $G_{mean}$ and the Overall Accuracy ($OA$) as the performance metrics as follows:

$$\begin{aligned} G_{mean} &= \sqrt{\text{TPR}\times\text{TNR}} \\ OA &= \frac{\text{TP+TN}}{\text{TP+FP +TN +FN}} \end{aligned} \tag{6}$$

where TP and TN are correctly predicted positive and negative test samples, FP and FN are incorrect predictions, respectively. It appears that classifiers provide the most accurate prediction for both the preterm and term classes, typ-

ically with the highest value of $G_{mean}$ and overall accuracy $OA$. Besides these quantities, the Area Under (Receiver Operator) Curve (AUC) [39] is also used to verify the proposed method.

## 4.2. Optimal synthetic preterm samples

Based on the intuitive analysis given in the previous sections, SMOTE has a stronger ability than ADASYN to maintain the importance of the features in classification. We first apply it to generate synthetic minority samples corresponding to a given sample balance coefficient $\alpha$. Fig. 4(a) shows the measured feature score $f_s^i$ with different $\alpha$. It is clearly difficult to determine how many synthetic samples should be generated, because the measured values of feature score $f_s^i$ are well separated and vary with $\alpha$.

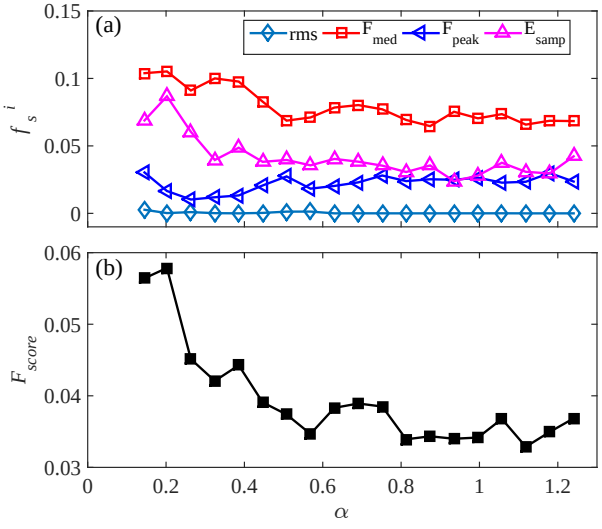


**Figure 4:** Measured feature scores $f_s^i$ (a) and $F_{score}$ (b) at different sample balance coefficient $\alpha$. The importances of each of the frequently used features $f_s^i$ show variation after applying synthetic algorithm SMOTE. However, the global feature score $F_{score}$ obtained from the importance $w_i$ and feature score $f_s^i$ shows continuous decrease with the increase of $\alpha$.

After synthesizing enough artificial samples, we use the features to build a forest, from which we obtain a classification accuracy value. The reduced classification accuracy by randomly permuting a node in the tree gives a reliable measure of the importance (weight) of the feature [40]. Calculating the importance (weight) of the features at each $\alpha$ allows us to investigate the combined effect of synthetic samples on the importance of the features $F_{score}$. As shown in Fig. 4 (b), $F_{score}$ decreases when the number of synthetic samples is increased, which illustrates the drawbacks of the synthetic sampling. This also implies that it is important to determine the optimal sampling adjustement coefficient $\alpha$.

Combining the activation and inactivation functions introduced earlier ( Fig. 5(a) ), the effective feature score $F_{score}^e$ shows a trend that helps us to easily find out the optimal sample balance coefficient $\alpha^*$. As shown in Fig. 5(b), $F_{score}^e$ initially increases with $\alpha$, manifesting the effect of the synthetic samples to eliminate the bias towards the majority (term) sampling. Due to the attenuation of features' scores and the bias towards term at large $\alpha$, $F_{score}^e$ reaches its peak at $\alpha \approx 0.7$, and decreases for higher values of $\alpha$. The position of the peak $F_{score}^e$ locates provides the optimal $\alpha^*$. Fig. 5 shows that with the optimal $\alpha^*$, the ability of the various features to distinguish between different classes has not been lost, while the bias towards the majority has been reduced.

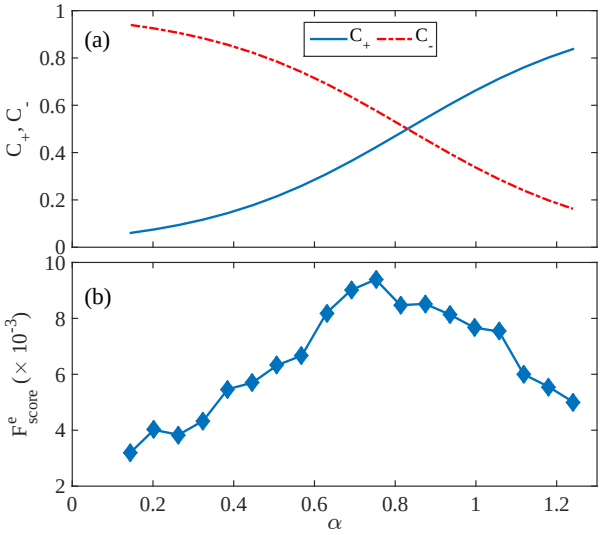


**Figure 5:** Variation of the effective feature score $F_{score}^e$ ( panel b) calculated from the activation and inactivation function (panel a) at different sample balance coefficient $\alpha$. $F_{score}^e$ shows a peak at $\alpha \approx 0.7$, which determines the optimal sample balance coefficient $\alpha^*$.

## 4.3. Validation

To verify the obtained optimal sample balance coefficient $\alpha^*$, the same features extracted from 80% of samples in the TPEHG database are used to train a SVC classifier, and the remaining 20% are then used for verification. Since the samples used for training and for testing are randomly selected, the computed quantities representing performances of the classifiers depend on the random choice. For this reason, we repeated the training-testing procedure for 100 times at each value of $\alpha$. To see the generalization ability of the classifier, we use the overall accuracy ($OA$) as an indicator to evaluate its performance in both training and testing processes. As shown in Fig. 6 (a), the measured $OA$ in both training and testing show a positive correlation with the sample balance coefficient $\alpha$, which is the expected benefits of introducing synthetic samples. More interestingly, the classifier performs similarly in both training and testing when the sample balance coefficient reaches a critical value close to $\alpha^*$ (see Fig. 6 (b) ). Although, a slight overestimation can be seen when slightly more or less synthetic samples are introduced, due to the higher correlation of training and testing

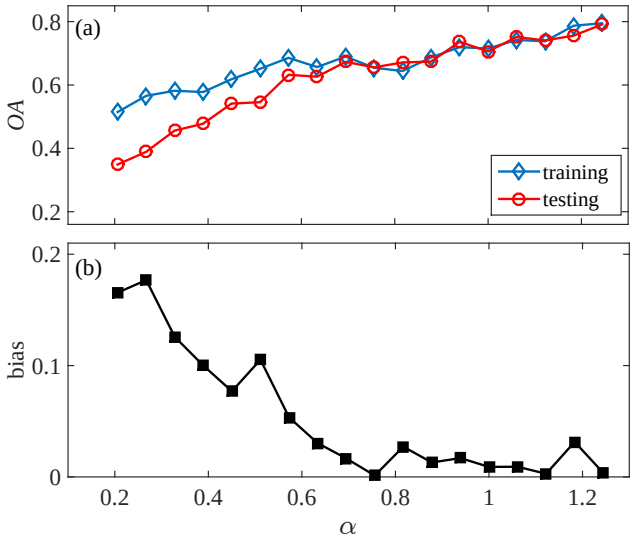


**Figure 6:** Classifier performances in both training and testing processes. (a) The measured overall accuracy ($OA$) shows positive correlation with sample balance coefficient $\alpha$. (b) Difference of classifier performance in training and testing.

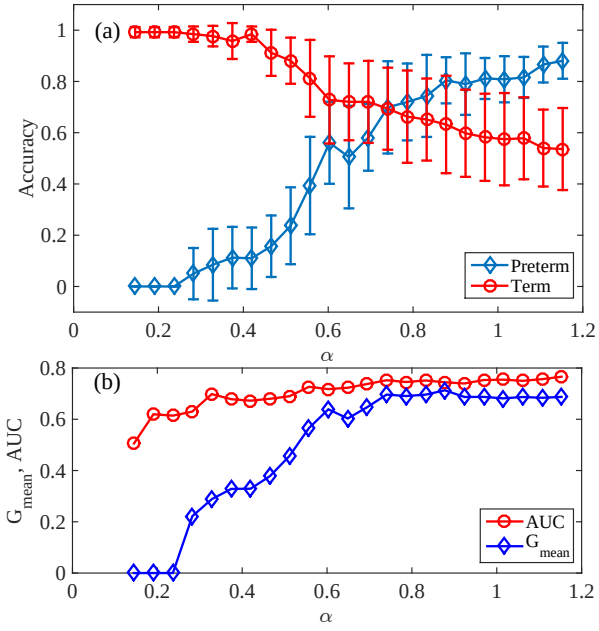


**Figure 7:** Prediction accuracy of SVC classifier on term (TNR) and preterm (TPR) at different sample balance coefficient $\alpha$. The SVC is trained with 80% of 262 term and $262 \cdot \alpha$ preterm samples synthesized by applying SMOTE algorithm. Prediction results are obtained with the rest of 20% samples. Results shown in panel (a) are averaged true positive rate ( preterm prediction precision) and true negative rate ( term prediction precision). The lower panel (b) shows the variation of $G_{mean}$ and AUC (area under curve) with respect to sample balance $\alpha$.

samples, no significant difference in classifier performance can be observed in training and testing processes. This gives us the confidence that the results in testing dataset are sufficient to validate the effectiveness of the proposed method.

In the following, we use TPR and TNR as metrics to quantify how accurately the proposed method could help to predict both preterm and term births. In addition, we use $G_{mean}$ and $OA$ defined in Eq. 6 to demonstrate the overall performance. Fig. 7 shows the variation of these quantities as a function of the number of synthetic samples added to the dataset, i.e., the increase of $\alpha$. As expected, the prediction accuracy for the minority increases, while that of majority decreases. In particular, when minority and majority samples are balanced ($\alpha = 1.0$), the trained classifier loses its ability on term birth prediction (see the average TNR $\sim 0.5$ and high variance). This would imply, in practical terms, error in diagnosis in half of the the cases, and unnecessary treatments. Another concern in the case of the ideal balanced dataset synthesizing strategy is the unrealistic prediction performance in testing process. This is confirmed by the reducing variance in TPR while an increasing variance in TNR shown in Fig.7(a), which is a direct consequence of higher correlation between training and testing samples. This problem can be solved by determining an optimal sample balance coefficient. It is worth noting that the two curves intersect at the optimal $\alpha \approx 0.7$ , which corresponds to the previously determined optimal sample balance coefficient $\alpha^*$. At this point the trained classifier eliminates most of the bias toward the majority (term) and increases the accuracy in predicting the minority (preterm birth) group without sacrificing too much accuracy in term prediction. This is confirmed by the accompanied variation of $G_{mean}$ and AUC, see Fig. 7(b), where both of these two quantities peak at $\alpha \approx \alpha^*$.

**Table 1**
Comparison of classifier performance under optimal sampling

| Classifier | $\alpha = 1.0$ | | | $\alpha = \alpha^*$ | | |
|---|---|---|---|---|---|---|
| | OA | $G_{mean}$ | AUC | OA | $G_{mean}$ | AUC |
| LDC | 0.62 | 0.62 | 0.66 | **0.68** | **0.65** | **0.71** |
| SVC | 0.73 | 0.72 | 0.74 | **0.75** | **0.74** | **0.75** |
| DTC | 0.75 | 0.74 | 0.83 | **0.82** | **0.82** | **0.86** |
| GBC | 0.78 | 0.78 | 0.83 | **0.85** | **0.84** | **0.91** |

The receiver operator characteristic curve (ROC) and the associated AUC values shown in Fig. 8 show the cut-off values for the true positive and false positive rates at different sample balance level (different $\alpha$). It can be seen that the SVC classifier performs better at the optimal sample balance coefficient $\alpha^*$. Compared to the case of ideal balance ($\alpha = 1.0$), training with the optimal amount of synthetic samples leads to an improved performance.

The advantages shown in SVC for determining the optimal sample balance coefficient also apply to other classifiers. Table 1 provides a comparison of frequently used parameters for evaluating the performance of the classifiers. It can be observed that with the previously determined optimal sample balance coefficient $\alpha^*$, all classifiers show an improvement in performance, especially for SVC based classifiers.

The effect of the optimal sample balance coefficient also

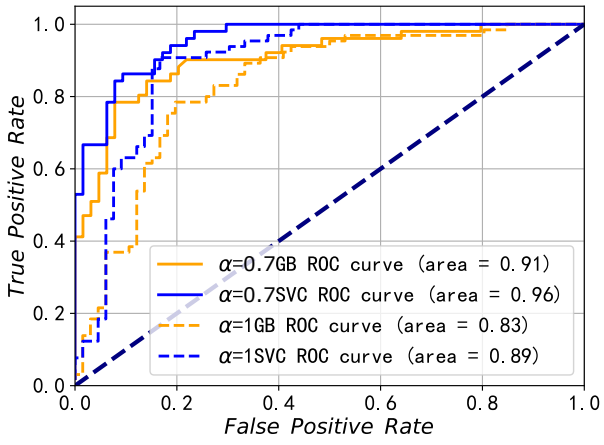


**Figure 8:** Receiver Operator Curve (ROC) for SVC ( blue ) and GBC (yellow) classifiers under different sample balance coefficient $\alpha$. In the case of optimal sample balance coefficient $\alpha^*$(=0.7, determined using the proposed method, solid line), classifiers show better performance.

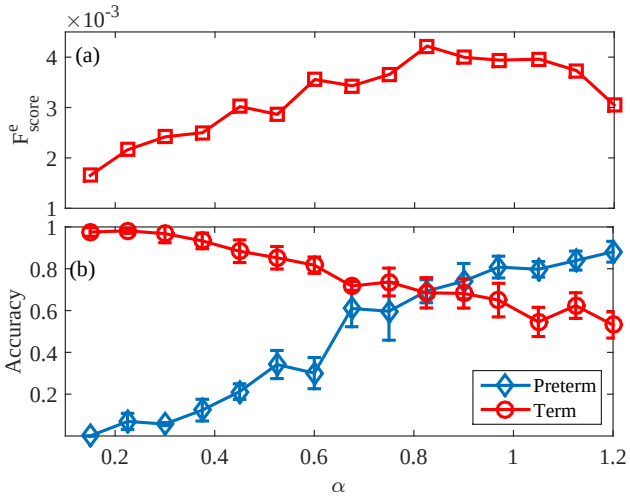


**Figure 9:** (a) Variation of effective feature score $F^e_{score}$ with respect to sample balance coefficient $\alpha$. Artificial samples are synthesized using ADASYN to give sample balance coefficient $\alpha$. $F^e_{score}$ shows a peak at $\alpha^* \approx 0.8$. (b) Prediction accuracy of SVC classifiers on term(TNR) and preterm (TPR) after been trained with different amount of synthetic samples using ADASYN method.

works with ADASYN. Although this method has less power in maintaining the ability of the features in classification, the combination of the proposed activation and inactivation functions does lead to an easy-to-identify optimal sample balance coefficient $\alpha_* = 0.8$. As shown in Fig. 9, the term and preterm prediction accuracy of a SVC classifier trained from these dataset gives the optimal performance at this $\alpha_*$. However, as the effective feature score $F^e_{score}$ obtained from ADASYN is less than what was obtained by using SMOTE as indicated in Fig.5, consistent with the expectation that the use of SMOTE method leads to better performance than that of ADASYN.

## 5. Conclusion & Discussion

Machine-learning based systems for automatic disease diagnosis offers a promising direction for modern healthcare. In these applications, the availability of healthcare data and the effectiveness of the features extracted from these samples play a decisive role. However, healthcare data are generally unbalanced, with most samples corresponding to healthy (majority), and a few corresponding to sick individuals (minority). When trained with unbalanced dataset, classifiers typically introduce biases towards the majority, making the automatic diagnosis system less useful. Synthetic sampling techniques are frequently used in the study of EHG signal based to predict preterm birth, with the aim of eliminating bias toward majority. However, they bring a side-effect of reduced ability of features in class separation and unrealistic high performance reported in literature[41].

In the traditional data synthesizing strategy, the number of minority samples must be equal to that of majority, no matter how small the minority class is. Excessive synthetic samples not only decreases classifier's ability in identifying majority sample, but may also lead to un-realistic classification performance (due to high correlation between samples). To go beyond this analysis of the advantages and disadvantages of introducing sythetic samples, we propose here a method for determining the optimal number of artificial samples that compromises the effect of synthetic samples in reducing feature effectiveness and eliminating bias towards majority. We expect that our work opens. To proceed, we measure the contributions of the features and their weights in class separation in the case of introducing different numbers of synthetic samples. Combining with the activation and inactivation functions introduced to describe the effect of sample abundance on classification precision, we obtain the optimal sample balance coefficient that compromises the effect of synthetic samples on eliminating bias and the side-effect of weakening feature importance. We apply the proposed method to predict preterm behavior using features extracted from public available database TPEHG. After applying synthetic algorithms, system performances are compared under different scenarios and the results highlight the importance of optimal sample balance coefficient proposed in the work.

It could be argued that it is more critical for an automatic diagnosis system to misidentify a true preterm patient than a normal patient, considering the serious consequences the preterm infants are facing. For this reason, increasing the number of synthetic samples should be of greater interest, which corresponds to the case shown in Fig.7 and 9. However, before drawing this conclusion, special attention should be paid. In the absence of a field test for an EHG-based preterm birth diagnosis system, its performance is typically verified using data samples randomly selected from the total sample set. The datasets used to check the perfor-

mance in preterm prediction are synthesized from the same minority class as those used for the training purpose. As the number of synthetic samples increases, the validation samples become closer and closer to the training samples, which lead to unrealistically high scores of the accuracy of preterm birth prediction, especially in real-world applications[42]. In the proposed method, we suppress this side-effect by introducing the activation/inactivation functions that account for the original size of minority samples. It is believed that validation results should be close to the actual applications.

As pointed by Vandewiele and co-authors in [43], introducing synthetic samples before partitioning the whole dataset into training and testing subsets can lead to un-realistic prediction performance. However, due to the very few available samples in the TPEHG, partitioning datasets before synthesizing artificial samples could give even worse results. For this reason, researchers all adopted the first scenario. The paper focuses on the two inevitable problems when using synthetic sampling technique to address data imbalance, particularly in the case of dataset with very few samples. Although the proposed method is believed to be general, no test have been done to other datasets or the TPEHG dataset with more features. Another limitation goes to the numerical way of determining the optimal balance coefficient $\alpha^*$, which would be of tedious computational work. All these questions require future work.

## 6. Acknowledgement

This work was supported by the National Natural Science Foundation of China under grant 61873238,

## References

[1] Christopher P Howson, Mary V Kinney, Lori Mcdougall, and Joy E Lawn. Born too soon: Preterm birth matters. *Reproductive Health*, 10(1):1–9, 2013.

[2] Miha Lucovnik, Ruben J Kuon, Linda R Chambliss, William L Maner, Shaoqing Shi, Leili Shi, James Balducci, and Robert E Garfield. Use of uterine electromyography to diagnose term and preterm labor. *Acta Obstetricia et Gynecologica Scandinavica*, 90(2):150–157, 2011.

[3] Wim J E P Lammers, H Mirghani, Betty Stephen, S Dhanasekaran, Athiq Wahab, M A H Al Sultan, and F Abazer. Patterns of electrical propagation in the intact pregnant guinea pig uterus. *American Journal of Physiology-regulatory Integrative and Comparative Physiology*, 294(3), 2008.

[4] H Leman, Catherine Marque, and Jean Gondry. Use of the electrohysterogram signal for characterization of contractions during pregnancy. *IEEE Transactions on Biomedical Engineering*, 46(10):1222–1229, 1999.

[5] Mahmoud Hassan, Sofiane Boudaoud, Jeremy Terrien, B Karlsson, and Catherine Marque. Combination of canonical correlation analysis and empirical mode decomposition applied to denoising the labor electrohysterogram. *IEEE Transactions on Biomedical Engineering*, 58(9):2441–2447, 2011.

[6] G Feležorž, G Kavsek, Ž Novakantolic, and F Jager. A comparison of various linear and non-linear signal processing techniques to separate uterine emg records of term and pre-term delivery groups. *Medical & Biological Engineering & Computing*, 46(9):911–922, 2008.

[7] Asgeir Alexandersson, Thora Steingrimsdottir, Jeremy Terrien, C Marque, and B Karlsson. The icelandic 16-electrode electrohysterogram database. *Scientific Data*, 2(1):150017–150017, 2015.

[8] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[9] Peng Ren, Shuxia Yao, Jingxuan Li, Pedro A Valdessosa, and Keith M Kendrick. Improved prediction of preterm delivery using empirical mode decomposition analysis of uterine electromyography signals. *PLOS ONE*, 10(7), 2015.

[10] Paul Fergus, Ibrahim Olatunji Idowu, Abir Jaafar Hussain, and Chelsea Dobbins. Advanced artificial neural network classification for detecting preterm births using ehg records. *Neurocomputing*, 188:42–49, 2016.

[11] U Rajendra Acharya, K Vidya Sudarshan, Soon Qing Rong, Zechariah Tan, Choo Min Lim, Joel E W Koh, Sujatha Nayak, and Sulatha V Bhandary. 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.

[12] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[13] Bing Wang, Lei Wang, Chun-Hou Zheng, and Yan Xiong. Imbalance data processing strategy for protein interaction sites prediction. *IEEE/ACM transactions on computational biology and bioinformatics*, 2019.

[14] Sophia Daskalaki, Ioannis Kopanas, and Nikolaos M Avouris. Evaluation of classifiers for an uneven class distribution problem. *Applied Artificial Intelligence*, 20(5):381–417, 2006.

[15] Gilles Vandewiele, Isabelle Dehaene, György Kovács, Lucas Sterckx, Olivier Janssens, Femke Ongenae, Femke De Backere, Filip De Turck, Kristien Roelens, Johan Decruyenaere, Sofie Van Hoecke, and Thomas Demeester. Overly optimistic prediction results on imbalanced data: Flaws and benefits of applying over-sampling, 2020.

[16] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[17] Y. Yan, R. Liu, Z. Ding, X. Du, J. Chen, and Y. Zhang. A parameter-free cleaning method for smote in imbalanced classification. *IEEE Access*, 7:23537–23548, 2019.

[18] Fadi Thabtah, Suhel Hammoud, Firuz Kamalov, and Amanda H Gonsalves. Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513:429–441, 2020.

[19] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

[20] Nitesh V Chawla. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 875–886. Springer, 2009.

[21] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, 2002.

[22] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *International Symposium on Neural Networks*, pages 1322–1328, 2008.

[23] C. H. Kok, C. Y. Ooi, M. Moghbel, N. Ismail, H. S. Choo, and M. Inoue. Classification of trojan nets based on scoap values using supervised learning. In *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, May 2019.

[24] Javier Andreuperez, Carmen C Y Poon, Robert Merrifield, Stephen T C Wong, and Guangzhong Yang. Big data for health. *IEEE Journal of Biomedical and Health Informatics*, 19(4):1193–1208, 2015.

[25] C Chiara Rabotti, M Massimo Mischi, L Beulen, S Guid Oei, and Jwm Jan Bergmans. Modeling and identification of the electrohysterographic volume conductor by high-density electrodes. *IEEE Transactions on Biomedical Engineering*, 57(3):519–527, 2010.

[26] M. Hassan, J. Terrien, C. Muszynski, A. Alexandersson, C. Marque, and B. Karlsson. Better pregnancy monitoring using nonlinear correlation analysis of external uterine electromyography. *IEEE Transac-*

*tions on Biomedical Engineering*, 60(4):1160–1166, April 2013.

[27] Marta Borowska, Ewelina Brzozowska, Pawe Ku, Edward Oczeretko, Romuald Mosdorf, and Piotr Laudaski. Identification of preterm birth based on rqa analysis of electrohysterograms. *Computer Methods and Programs in Biomedicine*, 153:227–236, 2018.

[28] Mehdi Shahrdad and Mehdi Chehel Amirani. Detection of preterm labor by partitioning and clustering the ehg signal. *Biomedical Signal Processing and Control*, 45:109–116, 2018.

[29] Rok Blagus and Lara Lusa. Class prediction for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 11(1):523–523, 2010.

[30] Douglas M Hawkins. The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1):1–12, 2004.

[31] Chumphol Bunkhumpornpat, Krung Sinapiromsaran, and Chidchanok Lursinsap. Dbsmote: Density-based synthetic minority oversampling technique. *Applied Intelligence*, 36(3):664–684, 2012.

[32] Samad Nejatian, Hamid Parvin, and Eshagh Faraji. Using subsampling and ensemble clustering techniques to improve performance of imbalanced classification. *Neurocomputing*, 276:55–66, 2018.

[33] QingJun Song, HaiYan Jiang, and Jing Liu. Feature selection based on fda and f-score for multi-class classification. *Expert Systems with Applications*, 81:22 – 27, 2017.

[34] M. T. Uddin and M. A. Uddiny. A guided random forest based feature selection approach for activity recognition. In *2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, pages 1–6, 2015.

[35] H. Deng and G. Rouger. Gene selection with guided regularized random forest. *Pattern Recognition*, 46(12):3483–3489, 2013.

[36] Derek Kweku Degbedzui and Mehmet Emin Yuksel. Accurate diagnosis of term–preterm births by spectral analysis of electrohysterography signals. *Computers in Biology and Medicine*, 119:103677, 2020.

[37] William L Maner, Robert E Garfield, Holger Maul, Gayle Olson, and George Saade. Predicting term and preterm delivery with transabdominal uterine electromyography. *Obstetrics & Gynecology*, 101(6):1254 – 1260, 2003.

[38] Paul Fergus, Pauline Cheung, Abir Jaafar Hussain, Dhiya Aljumeily, Chelsea Dobbins, and Shamaila Iram. Prediction of preterm deliveries from ehg signals using machine learning. *PLOS ONE*, 8(10), 2013.

[39] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[40] Alexandre Bureau, Josee Dupuis, Brooke Hayward, Kathleen Falls, and Paul Van Eerdewegh. Mapping complex traits using random forests. *BMC Genetics*, 4(1):1–5, 2003.

[41] György Kovács Lucas Sterckx Olivier Janssens Femke Ongenae Femke De Backere Filip De Turck Kristien Roelens Johan Decruyenaere Sofie Van Hoecke Thomas Demeester Gilles Vandewiele, Isabelle Dehaene. Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling. *Artificial Intelligence in Medicine*, 111:101987, 2021.

[42] Gilles Vandewiele, Isabelle Dehaene, Olivier Janssens, Femke Ongenae, Femke De Backere, Filip De Turck, Kristien Roelens, Sofie Van Hoecke, and Thomas Demeester. A critical look at studies applying over-sampling on the tpehgdb dataset. In David Riaño, Szymon Wilk, and Annette ten Teije, editors, *ARTIFICIAL INTELLIGENCE IN MEDICINE, AIME 2019*, volume 11526, pages 355–364. Springer, 2019.

[43] Gilles Vandewiele, Isabelle Dehaene, Olivier Janssens, Femke Ongenae, Femke De Backere, Filip De Turck, Kristien Roelens, Sofie Van Hoecke, and Thomas Demeester. A critical look at studies applying over-sampling on the tpehgdb dataset. 11526:355–364, 2019.