UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

DOCTORAL THESIS

Design and Assessment of a Computer-Assisted Artificial Intelligence System for Predicting Preterm Labor in Women Attending Regular Check-Ups. Emphasis in Imbalance Data Learning Technique

by
Félix Nieto del Amor

Supervised by
PhD. Yiyao Ye Lin
PhD. Gema Prats Boluda

València, Spain
October 2023

A mi familia.

Agradecimientos

En primer lugar, quiero agradecer de forma especial a mis directoras y compañeras, Yiyao y Gema, por todo su apoyo, orientación y dedicación.

Gracias también a mis amigos. A pesar de que muchos no tienen muy claro a qué me dedico, los momentos que compartimos sin duda han ayudado a pasar estos últimos tres años más llevaderos y agradables.

Gracias, finalmente, a mis padres y mi hermana, que siempre están ahí.

Abstract

Preterm delivery, defined as birth before 37 weeks of gestation, is a significant global concern with implications for the health of newborns and economic costs. It affects approximately 11% of all births, amounting to more than 15 million individuals worldwide. Current methods for predicting preterm labor lack precision, leading to overdiagnosis and limited practicality in clinical settings. Electrohysterography (EHG) has emerged as a promising alternative by providing relevant information about uterine electrophysiology. However, previous prediction systems based on EHG have not effectively translated into clinical practice, primarily due to biases in handling imbalanced data and the need for robust and generalizable prediction models.

This doctoral thesis aims to develop an artificial intelligence based preterm labor prediction system using EHG and obstetric data from women undergoing regular prenatal check-ups. This system entails extracting relevant features, optimizing the feature subspace, and evaluating strategies to address the imbalanced data challenge for robust prediction.

The study validates the effectiveness of temporal, spectral, and non-linear features in distinguishing between preterm and term labor cases. Novel entropy measures, namely dispersion and bubble entropy, outperform traditional entropy metrics in identifying preterm labor. Additionally, the study seeks to maximize complementary information while minimizing redundancy and noise features to optimize the feature subspace for accurate preterm delivery prediction by a genetic algorithm.

Furthermore, we have confirmed leakage information between train and test data set when generating synthetic samples before data partitioning giving rise to an overestimated generalization capability of the predictor system. These results emphasize the importance of using partitioning-resampling techniques for ensuring data independence between train and test samples. We propose to combine genetic algorithm and resampling method at the same iteration to deal with imbalanced data learning using partition-resampling pipeline, achieving an Area Under the ROC Curve of 94% and Average Precision of 84%. Moreover, the model demonstrates an F1-score and recall of approximately 80%, outperforming existing studies on partition-resampling pipeline.

This finding reveals the potential of an EHG-based preterm birth prediction system, enabling patient-oriented strategies for enhanced preterm labor prevention, maternal-fetal well-being, and optimal hospital resource management.

Overall, this doctoral thesis provides clinicians with valuable tools for decision-making in preterm labor maternal-fetal risk scenarios. It enables clinicians to design a patient-oriented strategies for enhanced preterm birth prevention and management. The proposed methodology holds promise for the development of an integrated preterm birth prediction system that can enhance pregnancy planning, optimize resource allocation, and ultimately improve the outcomes for both mother and baby.

Resumen

El parto prematuro, definido como el nacimiento antes de las 37 semanas de gestación, es una importante preocupación mundial con implicaciones para la salud de los recién nacidos y los costes económicos. Afecta aproximadamente al 11% de todos los nacimientos, lo que supone más de 15 millones de individuos en todo el mundo. Los métodos actuales para predecir el parto prematuro carecen de precisión, lo que conduce a un sobrediagnóstico y a una viabilidad limitada en entornos clínicos. La electrohisterografía (EHG) ha surgido como una alternativa prometedora al proporcionar información relevante sobre la electrofisiología uterina. Sin embargo, los sistemas de predicción anteriores basados en EHG no se han trasladado de forma efectiva a la práctica clínica, debido principalmente a los sesgos en el manejo de datos desbalanceados y a la necesidad de modelos de predicción robustos y generalizables.

Esta tesis doctoral pretende desarrollar un sistema de predicción del parto prematuro basado en inteligencia artificial utilizando EHG y datos obstétricos de mujeres sometidas a controles prenatales regulares. Este sistema implica la extracción de características relevantes, la optimización del subespacio de características y la evaluación de estrategias para abordar el reto de los datos desbalanceados para una predicción robusta.

El estudio valida la eficacia de las características temporales, espectrales y no lineales para distinguir entre casos de parto prematuro y a término. Las nuevas medidas de entropía, en concreto la dispersión y la entropía de burbuja, superan a las métricas de entropía tradicionales en la identificación del parto prematuro. Además, el estudio trata de maximizar la información complementaria al tiempo que minimiza la redundancia y las características de ruido para optimizar el subespacio de características para una predicción precisa del parto prematuro mediante un algoritmo genético.

Además, se confirmó la fuga de información entre el conjunto de datos de entrenamiento y el de prueba al generar muestras sintéticas antes de la partición de datos, lo que da lugar a una capacidad de generalización sobreestimada del sistema predictor. Estos resultados subrayan la importancia de particionar y después remuestrear para garantizar la independencia de los datos entre las muestras de entrenamiento y de prueba. Se propone combinar el algoritmo genético y el remuestreo en la misma iteración para hacer frente al desequilibrio en el aprendizaje de los datos mediante el enfoque de partición-remuestreo, logrando un área bajo la curva ROC del 94% y una precisión media del 84%. Además, el modelo demuestra un F1-score y una sensibilidad de aproximadamente el 80%, superando a los estudios existentes que consideran el enfoque de remuestreo después de particionar. Esto revela el potencial de un sistema de predicción de parto prematuro basado en EHG, permitiendo estrategias orientadas al paciente para mejorar la prevención del parto prematuro, el bienestar materno-fetal y la gestión óptima de los recursos hospitalarios.

En general, esta tesis doctoral proporciona a los clínicos herramientas valiosas para la toma de decisiones en escenarios de riesgo materno-fetal de parto prematuro. Permite a los clínicos diseñar estrategias orientadas al paciente para mejorar la prevención y el manejo del parto prematuro. La metodología propuesta es prometedora para el desarrollo de un sistema integrado de predicción del parto prematuro que pueda mejorar la planificación del embarazo, optimizar la asignación de recursos y reducir el riesgo de parto prematuro.

Resum

El part prematur, definit com el naixement abans de les 37 setmanes de gestació, és una important preocupació mundial amb implicacions per a la salut dels nounats i els costos econòmics. Afecta aproximadament a l'11% de tots els naixements, la qual cosa suposa més de 15 milions d'individus a tot el món. Els mètodes actuals per a predir el part prematur manquen de precisió, la qual cosa condueix a un sobrediagnòstic i a una viabilitat limitada en entorns clínics. La electrohisterografia (EHG) ha sorgit com una alternativa prometedora en proporcionar informació rellevant sobre l'electrofisiologia uterina. No obstant això, els sistemes de predicció anteriors basats en EHG no s'han traslladat de manera efectiva a la pràctica clínica, degut principalment als biaixos en el maneig de dades desequilibrades i a la necessitat de models de predicció robustos i generalitzables.

Aquesta tesi doctoral pretén desenvolupar un sistema de predicció del part prematur basat en intel·ligència artificial utilitzant EHG i dades obstètriques de dones sotmeses a controls prenatals regulars. Aquest sistema implica l'extracció de característiques rellevants, l'optimització del subespai de característiques i l'avaluació d'estratègies per a abordar el repte de les dades desequilibrades per a una predicció robusta.

L'estudi valguda l'eficàcia de les característiques temporals, espectrals i no lineals per a distingir entre casos de part prematur i a terme. Les noves mesures d'entropia, en concret la dispersió i l'entropia de bambolla, superen a les mètriques d'entropia tradicionals en la identificació del part prematur. A més, l'estudi tracta de maximitzar la informació complementària al mateix temps que minimitza la redundància i les característiques de soroll per a optimitzar el subespai de característiques per a una predicció precisa del part prematur mitjançant un algorisme genètic.

A més, hem confirmat la fugida d'informació entre el conjunt de dades d'entrenament i el de prova en generar mostres sintètiques abans de la partició de dades, la qual cosa dona lloc a una capacitat de generalització sobreestimada del sistema predictor. Aquests resultats subratllen la importància de particionar i després remostrejar per a garantir la independència de les dades entre les mostres d'entrenament i de prova. Proposem combinar l'algorisme genètic i el remostreig en la mateixa iteració per a fer front al desequilibri en l'aprenentatge de les dades mitjançant l'enfocament de partició-remostrege, aconseguint una àrea sota la corba ROC del 94% i una precisió mitjana del 84%. A més, el model demostra una puntuació F1 i una sensibilitat d'aproximadament el 80%, superant als estudis existents que consideren l'enfocament de remostreig després de particionar. Això revela el potencial d'un sistema de predicció de part prematur basat en EHG, permetent estratègies orientades al pacient per a millorar la prevenció del part prematur, el benestar matern-fetal i la gestió òptima dels recursos hospitalaris.

En general, aquesta tesi doctoral proporciona als clínics eines valuoses per a la presa de decisions en escenaris de risc matern-fetal de part prematur. Permet als clínics dissenyar estratègies orientades al pacient per a millorar la prevenció i el maneig del part prematur. La metodologia proposada és prometedora per al desenvolupament d'un sistema integrat de predicció del part prematur que puga millorar la planificació de l'embaràs, optimitzar l'assignació de recursos i millorar la qualitat de l'atenció.

Contents

List of Figures

List of Tables

List of Acronyms

Acc	Accuracy
AdaBoost	Adaptive Boosting
ADASYN	Adaptive Synthetic Sampling
AF	Feature Subset with all features
AI	Artificial Intelligence
ANN	Artificial Neural Network
AP	Average Precision
App	Peak to Peak Amplitude
AUC	Area Under the ROC Curve
BubbEn	Bubble Entropy
$\mathbf{BubbEn_{FWH}}$	Bubble Entropy in Fast Wave High Bandwidth (see Table 4.2)
$\mathbf{BubbEn_{WBW}}$	Bubble Entropy in Whole Bandwidth (see Table 4.2)
c	Number of Classes for Dispersion Entropy
Chrom	Chromosome
CI	Confidence Interval
CL	Cervical Length
CLF	Classifier
CNN	Convolutional Neural Network
CRP	C-Reactive Protein
D1	Decile 1 (0.2–1 Hz)
D2	Decile 2 (0.2–1 Hz)
D3	Decile 3 (0.2–1 Hz)
D4	Decile 4 (0.2–1 Hz)
D5	Decile 5 (0.2–1 Hz)
D6	Decile 6 (0.2–1 Hz)

D7	Decile 7 (0.2–1 Hz)
D8	Decile 8 (0.2–1 Hz)
D9	Decile 9 (0.2–1 Hz)
DF1	Dominant Frequency (0.2–1 Hz)
DF2	Dominant Frequency (0.34–1 Hz)
DispEn	Dispersion Entropy
DispEn$_{\mathbf{FWH}}$	Dispersion Entropy in Fast Wave High Bandwidth (see Table 4.2)
DispEn$_{\mathbf{WBW}}$	Dispersion Entropy in Whole Bandwidth (see Table 4.2)
DNN	Deep Neural Network
E1	Electrode 1 of the TPEHG DB and TPEHGT DS databases
E2	Electrode 2 of the TPEHG DB and TPEHGT DS databases
E3	Electrode 3 of the TPEHG DB and TPEHGT DS databases
E4	Electrode 4 of the TPEHG DB and TPEHGT DS databases
EEG	Electroencephalogram
EHG	Electrohysterography
EMD	Empirical Mode Decomposition
En$_{\mathbf{ALL}}$	Set of sample, fuzzy, spectral, dispersion, and bubble entropy
En$_{\mathbf{SFS}}$	Entropy of Selected Frequency Bands
fFN	Fetal Fibronectin
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
FS$_{\mathbf{TH}}$	Feature Subset with Training Hybrid (Under/Oversampling)
FS$_{\mathbf{TO}}$	Feature Subset with Training Oversampling
FS$_{\mathbf{TU}}$	Feature Subset with Training Undersampling
FS$_{\mathbf{VH}}$	Feature Subset with Validation Hybrid (Under/Oversampling)

FS$_{\mathbf{VO}}$	Feature Subset with Validation Oversampling
FS$_{\mathbf{VU}}$	Feature Subset with Validation Undersampling
FuzEn	Fuzzy Entropy
FuzEn$_{\mathbf{FWH}}$	Fuzzy Entropy in Fast Wave High Bandwidth (see Table 4.2)
FuzEn$_{\mathbf{WBW}}$	Fuzzy Entropy in Whole Bandwidth (see Table 4.2)
FW	Fast Wave Bandwidth(0.1–4Hz)
FWH	Fast Wave High Bandwidth(0.34–4Hz)
FWL	Fast Wave Low Bandwidth(0.1–0.34Hz)
GA	Genetic Algorithm
GANs	Generative Adversarial Networks
GBC	Gradient Boosting Classifier
GNB	Gaussian Naive Bayes
H/L Ratio	High (0.34–1 Hz)–to Low (0.2–0.34 Hz) Frequency Energy Ratio
IL-6	Interleukin-6
IMF	Intrinsic Mode Function
IoT	Internet of Things
IUPC	Intrauterine Pressure Catheter
KFD	Katz Fractal Dimension
KNN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
LEn$_{\mathbf{ALL}}$	Linear and all entropy features
Linear	Linear Features
LNL	Linear and Nonlinear Features
LNLEn$_{\mathbf{ALL}}$	Linear, nonlinear, and all entropy features
LR	Logistic Regression
LZBin	Lempel-Ziv Complexity (Binary)

LZMulti	Lempel-Ziv Complexity (Multivariate)
m	Embedding Dimension for Entropy Measures
MeanF	Mean Frequency
MRI	Magnetic Resonance Imaging
N	Population Size
NCFeat	Number of Features in the Current Subset
NCL	Neighborhood Cleaning Rule
NFeat	Number of Features in the Initial Set
NL	Nonlinear Features
NormEn	Normalized Entropy
NPV	Negative Predictive Value
OBST	Obstetrics Features
PCA	Principal Component Analysis
PPV	Positive Predictive Value
P-R	Partition-Resampling Approach
PSO	Particle Swarm Optimization
QDA	Quadratic Discriminant Analysis
QSVM	Quantum Support Vector Machine
r	Scaling Factor for Entropy Measures
RANOVA	Repeated Measures Analysis of Variance
RF	Random Forest
RH	Resampling Technique: Hybrid (SMOTE + Neighborhood Cleaning Rule)
RN	Resampling Technique: Not applicable
RNN	Recurrent Neural Network
RO	Resampling Technique: Oversampling
ROC	Receiver Operating Characteristic Curve

R-P	Resampling-Partition Approach
RU	Resampling Technique: Undersampling
RUS	Random Under-Sampling
RUSBoost	Random Under-Sampling with Boosting
S1	Channel 1 of the TPEHG DB and TPEHGT DS databases
S2	Channel 2 of the TPEHG DB and TPEHGT DS databases
S3	Channel 3 of the TPEHG DB and TPEHGT DS databases
SAE	Sparse Autoencoder
SampEn	Sample Entropy
SampEn$_{\mathbf{FWH}}$	Sample Entropy in Fast Wave High Bandwidth (see Table 4.2)
SampEn$_{\mathbf{WBW}}$	Sample Entropy in Whole Bandwidth (see Table 4.2)
SD1	Minor Axes of the Poincaré Ellipse
SD2	Major Axes of the Poincaré Ellipse
SDRR	$\sqrt{(\mathrm{SD1}^2 + \mathrm{SD2}^2)/2}$
SMOTE	Synthetic Minority Oversampling TEchnique
SpEn	Spectral Entropy
SpMR	Spectral Moment Ratio
SSAE	Stacked Sparse Autoencoder
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine
SW	Slow Wave Bandwidth(0.005–0.03Hz)
TimeRev	Time Reversibility
TN	True Negatives
TOCO	Tocodynamometer
TP	True Positives
TPEHG DB	Term-Preterm EHG Database

TPEHGT DS	Term-Preterm EHG DataSet with Tocogram
TPL	Threatened Preterm Labor
TPR	True Positive Rate
WBW	Whole Bandwidth(0.1–4Hz)
WOG	Week of Gestation at Recording Time

Chapter 1

Introduction

1.1 Preterm birth

1.1.1 Origin, prevalence and relevance

The World Health Organization defines preterm births as those that occur before the 37$^{\text{th}}$ week of gestation. Prematurity is the major determinant of neonatal morbidity and mortality and affects around 11% of all births and more than 15 million persons worldwide, while its incidence is increasing annually [1, 2]. Prematurity causes 1 million newborn deaths each year and is the leading cause of death in the first four weeks of life [3]. Thanks to the medical advances made in the last decades, survival of preterm newborns has improved considerably. More than 95% of the preterm infants that receive modern neonatal and pediatric care survive to adulthood (>18 years old) [4]. In case of survival, preterm newborns may suffer complications in the short term. The possibilities of survival increase with gestational age, leading to high differences in the survival rates between extremely preterm (<28 weeks), very preterm (≥28 and <32 weeks), moderately (≥32 and <34 weeks) preterm and late preterm infants (≥34 and <37 weeks) [5]. Extremely preterm deliveries result in several months of lost fetal development, leaving infants more susceptible to morbidities [6]. The proportion of surviving infants born without severe morbidity is ∼9% at 22 weeks of gestation, and ∼64%, ∼95% and 99% for extremely, very and moderately preterm infants, respectively [7, 8]. The risk of hospital admission to the neonatal unit is twice as high than in the term group, even in the late preterm group [9]. Morbidity is mainly due to respiratory distress syndrome, with a prevalence of 88% of the total of preterm births before 34 weeks of gestation, followed by retinopathy (45%), intraventricular hemorrhages (37.4%) and bronchopulmonary dysplasia (32%) [10]. One in every five babies who survive will have an intellectual disability, one in two will have cerebral palsy, and one in three will have eye damage [3]. Premature babies are also at risk of developing other long-term conditions, such as asthma, learning disabilities, attention deficit disorder, and emotional problems [11].

Medical care for preterm birth requires significant hospital resources and has a great impact on public health systems. The average cost of a preterm birth is 5 to 10 times higher than that of a term birth [12]. For an extremely preterm baby, the average cost in Canada per baby amounted to $67,467 in the first ten years of life [12]. Saving a baby weighing less than 750 grams costs more than $117,000, the highest cost in public health [13]. In the United States, preterm birth incurred an economic cost of at least $26.2 billion in 2005, including medical, educational and lost productivity expenses [5]. Furthermore, the average first-year medical costs for both in- and outpatient care were about 10 times greater for preterm ($32,325) than term infants ($3325) [5].

1.1.2 Current methods of predicting preterm delivery

The criteria for the diagnosis of preterm labor is somewhat imprecise since the underlying etymology and sequence of events preceding preterm labor are not fully understood. Symptoms such as pain caused by uterine contractions, pelvic pressure, increased uterine secretions and low back pain have been associated with preterm labor [14]. However, these symptoms can also be associated with normal pregnancies, making the diagnosis of preterm labor even more complex, which often results in overdiagnosis of as many as 40% of women with preterm labor symptoms [15]. Less than 10% of the women clinically diagnosed with threatened preterm labor give birth within 7 days of the onset of symptoms [16].

Several biomarkers have been used to predict preterm labor by measuring specific molecules or substances in maternal or fetal fluids or tissues associated with the biological processes that lead to preterm birth [17]. These biomarkers can be measured at different stages of pregnancy, although their levels may change over time, allowing clinicians to monitor women with a higher risk of preterm birth and adjust their supervision accordingly [18]. The main molecular biomarkers used for preterm birth prediction found in the literature are as follows:

- Fetal Fibronectin (fFN) is a glycoprotein that acts as a scaffold for cell adhesion and is present in the fetal membranes, cervix, and vaginal secretions. It acts as a glue between the fetal membranes and the uterine lining. The absence of fFN suggests that the patient is at low risk of preterm birth. The presence of fFN in cervical or vaginal secretions during pregnancy indicates that the fetal membranes have been disrupted, which is a common precursor of preterm labor. The predictive value of preterm birth is usually evaluated by different levels of fFN $\geq$10, $\geq$50, $\geq$200 and $\geq$500 ng/ml. A meta-analysis performed in 2021 [19] reviewed 15 studies which reported predictive values to address preterm labor in women with a singleton pregnancy and <34 weeks of gestation. The threshold of <10 ng/ml resulted in the highest average sensitivity of 0.78 (95% CI, 0.69–0.85) and the lowest specificity of 0.63 (0.52–0.73) with respect to the other cut-off values. In contrast, 500 ng/ml had the lowest sensitivity, 0.11 (95% CI, 0.07–0.18), but the highest specificity, 0.99 (95% CI, 0.97–1.00), i.e. a threshold of 500 ng/ml is highly effective in identifying women at low risk of preterm labor [19].

- Cervical Length (CL) is also a well-known preterm labor biomarker. The cervix is the lower part of the uterus that opens into the vagina. During pregnancy, the cervix undergoes changes in preparation for delivery, including the shortening of the cervix, which occurs gradually. If the cervix shortens too early, it may be a sign of preterm labor. Cervical length is usually measured by transvaginal ultrasound. Most women (75%) with a shortened cervix do not deliver preterm [20]. Women with a cervical length $<$ 25 mm before 28 weeks of gestation and contractions have twice the incidence of preterm birth than those with a cervical length $<$ 25 mm but no contractions [21]. The incidence of premature births increases in women who have already delivered prematurely, from ∼3% to ∼20% [22]. Taipale et al. studied 3694 women with no prior preterm birth and a gestational age <37 weeks and reported considering a CL $<$ 25 mm: sensitivity = 6%, specificity = 100%, Positive Predictive Value (PPV) = 39% and Negative Predictive Value (NPV) = 99% [23]. As in women with a prior preterm birth and <34 weeks of gestation, the following has been reported (CL $<$ 25mm): sensitivity = 76%, specificity = 68%, PPV = 20% and NPV = 96% [24]. A recent study combined the measurement of cervical length and fetal fibronectin levels to predict preterm delivery and resulted in an Area Under the ROC Curve (AUC) of 67%, with a maximum positive predictive value of only 14% and a negative predictive value of 96.1% [25]. Another recent work used a Convolutional Neural Network (CNN) model to predict preterm births from 354 two-dimensional transvaginal cervical ultrasound images, in which 319 and 35 images were for term and preterm groups, respectively [26]. In the classification task, they reported a sensitivity of 67.7 ±4.2%, precision of 68.3 ±8.7% and AUC of 72.3 ±13.4%. The deep learning algorithm focuses on different parts of the transvaginal cervical ultrasound images to classify them into preterm and term classes. The lower segment of the cervix, close to the ectocervix and the heterogeneity of the density of tissues around the cervix were the most important features in identifying preterm births, while the term cases were identified by mainly focusing on the top of the largest homogeneous region in the middle part of the anterior cervical lip [26].

- C-Reactive Protein (CRP) is an acute-phase protein produced by the liver in response to inflammation. CRP measurement is quick, noninvasive and risk-free, and can be an effective diagnostic test for evaluating and categorizing the risk of preterm labor and predicting the morbidity of both mother and fetus [27]. Significant differences were obtained between the mean of the CRP levels in preterm and term deliveries, CRP levels being lower in the latter [28]. Another study was performed on 59 preterm and 17 term women, who did not present complications during pregnancy. Considering the C-reactive protein >3.6 reported a sensitivity of 41.3 (30.1–53.3) %, specificity of 89.3 (80.1–95.3), positive predictive value of 79.5 (63.5–90.7) % and AUC of 0.683, indicating a significant relationship of CRP with preterm labor [27].

- Interleukin-6 (IL-6) is a pro-inflammatory cytokine involved in the initiation of labor. Elevated levels of IL-6 have been found in the amniotic fluid of preterm labor women. Lockwood et al. evaluated a total of 161 patients seen at 3- to 4-week intervals between 24 and 36 weeks. In premature deliveries, IL-6 concentrations in the cervix and vaginal canal increased ∼4-fold compared to women who delivered at term [29]. Another study predicted preterm labor in women with threatened preterm labor using IL-6 and achieved an AUC of 87.6%, sensitivity of 73.2% and specificity of 85.7% for predicting preterm labor. It should be noted that the database used was seriously skewed in the number of samples and acquisition protocol between the groups. The preterm group was made up of 82 patients with symptoms of threatened preterm labor and the control group included only 21 outpatients seen in routine obstetric visits [30].

- Progesterone plays a critical role in maintaining pregnancy quiescence. Low progesterone levels have been associated with an increased risk of preterm labor [31]. Progesterone supplementation may be recommended for women at high risk of preterm labor. Although it is a widely used biomarker [31], a recent extensive meta-analysis reviewed the efficacy of vaginal progesterone in preventing recurrent preterm birth in women with a singleton gestation and a history of spontaneous preterm birth. They determined that there is no strong evidence to support its use in reducing preterm labor because of the methodological limitations in most if the publications, including small samples [31].

- Uterine dynamic monitoring is a standard protocol in obstetric care during labor and delivery as it offers valuable information on uterine dynamics. The gold standard for this process is by inserting an Intrauterine Pressure Catheter (IUPC) into the uterine cavity for accurate and reliable measurements [32, 33]. IUPC provides essential information such as the contraction duration, frequency and resting tone. However, it involves certain risks such as pain, bleeding, and infection. As its use also requires rupturing the membrane, which can increase the risk of infection and premature rupture, it cannot be used to prevent preterm labor and is mainly used in high-risk pregnancies for monitoring uterine activity during labor [32, 33]. In clinical practice, Tocodynamometer (TOCO) is the most commonly used technique for uterine dynamic monitoring, as it is safe and non-invasive method using a device on the mother's abdomen to measure uterine pressure on the abdominal wall. However, the measurements depend largely on the subjective opinion of the specialist and require constant transducer repositioning. TOCO monitoring is seriously affected by maternal obesity, uterine fibroids, and other factors [34]. To provide a more comprehensive assessment of fetal well-being during labor and delivery, TOCO monitoring is often used in conjunction with other methods, such as electronic fetal heart rate monitoring.

This combination can help detect abnormalities in uterine activity and fetal distress, and determine the optimal timing of delivery [32, 33, 34].

To sum up, although various techniques are available for predicting preterm labor, none can accurately detect all preterm labor cases. The advantage of these techniques mainly lies in their ability to identify patients who are not at risk of preterm labor, known as a negative predictive value [35], thus reducing unnecessary interventions.

1.2 Electrohysterography for predicting preterm births

Electrohysterography (EHG), or recording the bioelectric potential generated by billions of uterine myometrial cells on the abdominal wall, has emerged as a promising alternative that provides relevant information on the uterine electrophysiological state for use in predicting preterm labor.

From the physiological point of view, like other biological cells, uterine smooth muscle cells exhibit negative resting potential with small and slow spontaneous fluctuations. When resting, the potential fluctuations reach a threshold, isolated or action burst potentials are induced with a peak-to-peak action potential amplitude ranging 33-69 mV [36, 37]. As action potentials propagate on the surface of a myometrial cell, the depolarization causes voltage dependent Ca^{2+} channels to open. When this occurs, Ca^{2+} enters the muscle cells, traveling down its electrochemical gradient to activate the myofilaments and generating a contraction by increasing the size and/or number of actual portals for Ca^{2+} entry [36, 37]. This cyclic depolarization and repolarization of the muscle cell membrane therefore results in a contraction and relaxation sequence of the myometrium. Although a single electrical spike can initiate a contraction, multiple coordinated spikes are needed for forceful and continuous contractions [36, 37]. Studies in isolated myometrial tissue have shown the temporal association between electrical activity and contractions [38, 39, 40, 41, 42]. In all the species investigated, each contraction is accompanied by a burst of action potentials, starting slightly earlier than the corresponding contraction (see Figure 1.1), and stopping before the uterus has completely relaxed. The frequency, amplitude and duration of the contractions are mainly determined by the occurrence frequency of the uterine electrical bursts, the total number of cells simultaneously active during the burst and the duration of the uterine electrical bursts, respectively [43, 44].

Uterine electrical activity is low and uncoordinated in early gestation. As pregnancy progresses, uterine myometrial cell conductance increases due to more gap-junctions, leading to coordinated, intense contractions previous to labor [37, 43]. Figure 1.2 shows an example of a 10 min EHG record at 31st Week of Gestation at Recording Time (WOG) for women delivered at term (lower trace) and prematurely (upper trace). Term labor women did not record any contractions, since the record was obtained far from delivery. There were two uterine contractions (EHG-bursts with

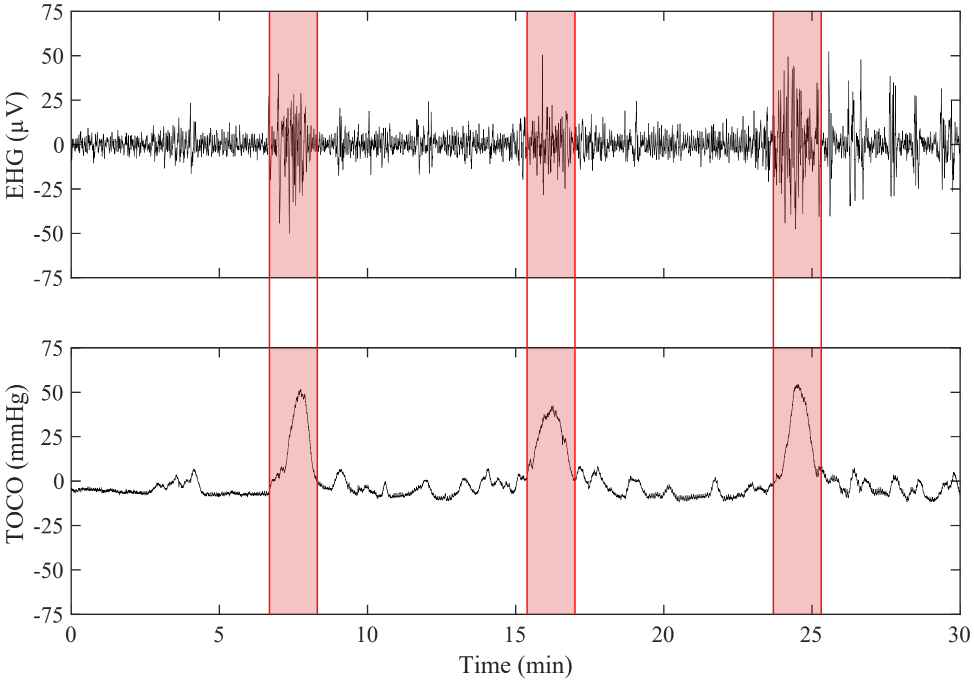


Figure 1.1: EHG (top) and TOCO (bottom) records of a woman at 26 weeks who delivered prematurely [45]. Red boxes are manual annotations of three contraction intervals in both records.

increased amplitude and frequency with respect to basal activity when the uterus is at rest) in preterm labor women.

EHG is composed of two components, the slow wave (SW, 0.005–0.03 Hz) and fast wave (FW, 0.1–4 Hz). The fast wave component can be further divided into two subcomponents: the fast wave low associated with signal propagation (FWL, 0.1–0.34 Hz) and fast wave high related to cell excitation (FWH, 0.34–4 Hz) [39, 46]. Traditionally, the EHG signal analysis is mainly focused on temporal, spectral, and non-linear parameters to describe the FW component [18, 47, 48] because the physiological significance of slow waves on surface recordings is questionable due to its overlapping with skin stretching and baseline oscillations [43]. Also, EHG energy, mainly distributed below 1 Hz, FWH, is usually analyzed in the 0.34–1 Hz range to minimize respiratory and cardiac interference [18].

Many studies have shown that the amplitude of the EHG signal increases as pregnancy progresses due to the increase in the number of uterine cells involved in the contractions [39]. Furthermore, as labor approaches, the spectral content of the EHG signal is shifted to higher frequencies, which are associated with an increase in cellular

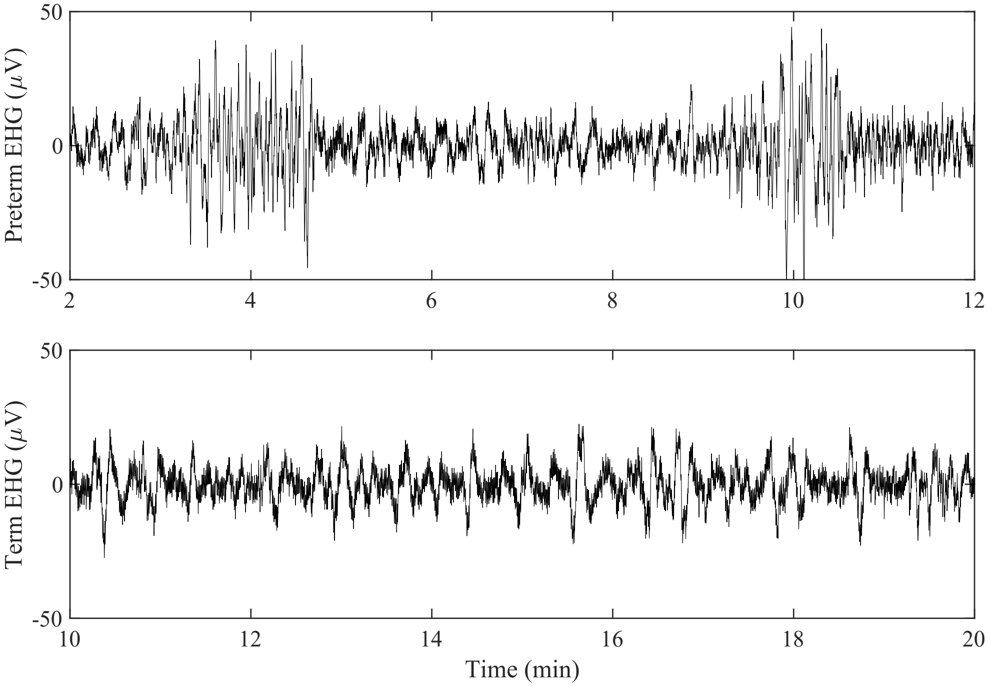


Figure 1.2: Example of EHG signals recorded at 31 weeks of gestation from woman who delivered prematurely at 33 weeks of gestation (top), and woman who delivered at term in the 37th week of gestation (bottom).

excitability [39, 46]. Labor proximity is also associated with higher signal predictability and regularity, reducing the chaos and complexity of the EHG signal [46, 47].

It should be noted that EHG recordings do not only contain uterine electrical activities of muscle cells, but also corrupted segments with multi-source interference (see Figure 1.3) such as motion artifacts or breathing. In surface myoelectric recordings, the former are unpredictable and vary in waveform depending on multiple factors. They can cause sudden changes in the bioelectric potential amplitude, which is also associated with alteration of the power spectral density, by distributing energy in the high-frequency range [49, 50, 51]. Breathing mainly affects the Fast Wave Low component bandwidth of the EHG signal [52]. Other factors that can corrupt EHG signals include the fetal or maternal electrocardiogram (1.38–1.5 Hz), electromyography noise (∼30 Hz) and electromagnetic noise from external devices (∼60 Hz) [52]. As the two latter are easily removed from the EHG by a low-pass filter at 4 Hz, a preprocessing step to discard both motion artifacts and respiratory interference is crucial to extract robust features from EHG records.

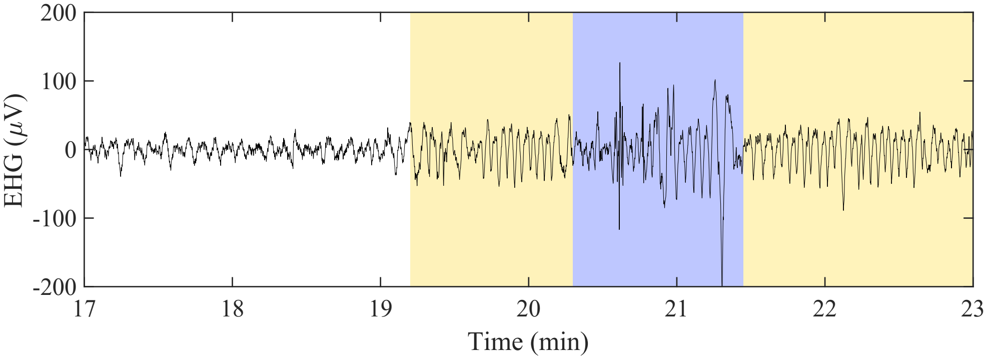


Figure 1.3: EHG signal with respiratory (yellow) and motion (blue) artifacts.

1.3 Machine learning

1.3.1 The application of machine learning in EHG signals to predict the preterm labor

Machine learning, a subfield of Artificial Intelligence (AI), has revolutionized healthcare by analyzing large amounts of medical data. Its main objective is to develop prediction models by automatically detecting the data structure for decision support. Machine learning allows computers to interpret complex patterns in the data, extract meaningful insights and make accurate classifications or predictions [53, 54]. Thanks to machine learning, medical professionals can gain deeper insights into patient data, make more accurate diagnoses, develop better treatment plans, and improve patient healthcare and their quality of life. It can also improve diagnosis, treatment, and patient outcomes through image and speech recognition, precision medicine, disease diagnosis, personalized treatment recommendations, and predictive analytics. Machine learning has been widely used in multiple medical applications, such as detection of retinopathy, bone age, identification of skin cancer or to predict liver disease and preterm labor from EHG records [54, 55].

The commonly used workflow for predicting preterm labor with EHG signals involves several steps (see Figure 1.4) [18, 56]:

1) Data acquisition: EHG signals are recorded from the surface of the mother's abdomen using electrodes connected to biomedical devices.

2) Signal preprocessing: As previously noted, EHG recordings encompass not only uterine electrical activity but also motion artifacts and respiratory interferences. Signal preprocessing typically involves the removal of noise and artifacts and the application of filters to enhance signal quality.

3) Feature extraction: temporal, spectral, non-linear and synchronization features,

when multichannel Electroencephalogram (EEG) is available, can be extracted from the preprocessed EHG signals. These features capture relevant information on the uterine electrophysiological state and serve as input features for the machine learning model [18, 56].

4) Dataset preparation: The extracted features are then used to create a dataset, in which each set of features obtained from the subject's EHG recordings is assigned a label indicating whether the delivery occurred prematurely or at term.

5) Dataset balancing: As mentioned above, the prevalence of preterm labor is about 10% in the general population, giving rise to an imbalanced dataset for preterm labor prediction systems [57]. This can lead to a bias towards the majority class and result in lower predictive performance [57]. To address this issue, researchers have employed various techniques to balance the dataset, such as oversampling the minority class, undersampling the majority class, and generating synthetic samples using techniques like the Synthetic Minority Oversampling TEchnique (SMOTE) [57]. These techniques allow machine learning algorithms to be trained on a more balanced dataset, leading to better performance and more accurate predictions for preterm labor from EHG signals [56]. However, it is important to note that these techniques should be used with caution and evaluated carefully, as they may also introduce other biases or tradeoffs in the model's performance [57].

6) Machine learning model training: To predict preterm delivery, a machine learning model, such as a logistic regression, linear discriminant analysis, k-nearest neighbors, neural network, or support vector machine, is trained using input features extracted from the EHG signals.

7) Model evaluation: The performance of the trained model is evaluated by various metrics such as accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUC).

8) Model optimization: If the model's performance is unsatisfactory, the workflow may need to be optimized by adjusting various parameters such as feature extraction, feature selection, dataset balancing techniques, or model hyperparameters (internal configurable parameters of each classification algorithm).

9) Model deployment: Once the model's performance is deemed satisfactory, it can be deployed to predict preterm labor using new, unseen EHG signals in real-world clinical settings.

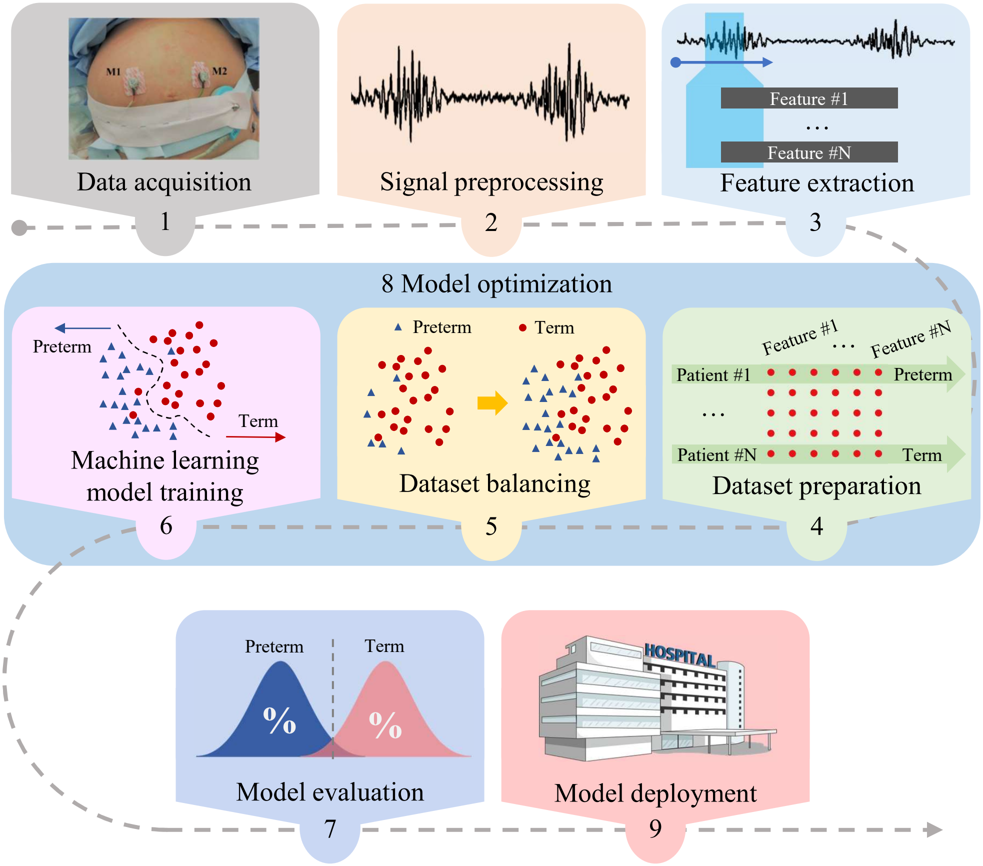


Figure 1.4: Workflow for predicting preterm labor with EHG.

1.3.2 Dimensionality reduction methods

As mentioned above, multiple input features are normally used to perform target classification tasks. The features can generally be divided into three groups with respect to the degree of importance required to discriminate the target classes. Relevant features are the attributes with the ability to partially describe some properties of the target classes, while their role cannot be assumed by others. An irrelevant attribute lacks any description of the target classes' properties. Redundant features are those attributes whose roles can be taken on by another attribute. In this regard, the different input features may contain complementary, redundant or noise information [53]. Complementary features can capture different aspects of the data and provide a more complete representation of the underlying patterns as a result, combining features that provide relevant and complementary information and can enhance the overall performance of a classification model. In contrast, the inclusion of noise information may even worsen classification performance [53].

When the number of features becomes too large, it leads to the "curse of dimensionality" [58]. If too many input features are used, the model may become too complex, giving rise to a high computational cost. In addition, the high dimensional input feature data may suffer from model "overfitting", leading to poor predictive system performance. Dimensionality reduction is the conversion of high-dimensional data into a reduced-dimensional representation. Ideally, the reduced representation should be that of the intrinsic dimensionality of the data, which is the minimum number of parameters necessary to interpret the characteristics of the data under consideration [59]. Principal Component Analysis (PCA) is one of the most widely feature reduction methods used. It projects the initial feature space onto a lower dimensionality subspace, which preserves the "essence" of the original data [53]. Although the data may appear large, there may only be a small number of degrees of variability, corresponding to latent factors [53]. The resulting principal components are essentially linear combinations of the original data that capture most of the variance in the data. However, although this method helps to solve the model complexity problem, it is not very useful in determining complementarity and redundancy between features or in discarding noisy information [60].

In contrast, feature selection methods can not only reduce data dimensionality, but also take the complementary data and data redundancy. These methods can obtain the subset of features that best fits the target classification problem, in this case the prediction of preterm delivery. Feature selection aims to identify the relevant features by their relations with the corresponding class labels, i.e. the preterm and term labor classes, and discards irrelevant and redundant features [61]. Feature selection methods are classified into filters, embedded, and wrapper methods, according to their relationship to the machine learning classification method [62]:

- Filter methods are independent of any learning method, as they focus on the general data features. These methods basically consist of computing a statistic such as the Pearson correlation coefficient or chi-squared test, which evaluates the discriminatory capacity between classes for each feature and selects the best. Some limitations include a lack of feature interaction, no interaction with the classification model, dependence on data distribution, and the computational expense of high-dimensional data [63].

- Embedded methods: In these methods, the feature selection process is integrated in the machine learning algorithm itself. These methods train classifiers while simultaneously selecting the most relevant features for optimal performance. The latter is also in fact one of its major limitations, since embedded methods can make it difficult to compare different models and their feature selection approaches, as feature selection is part of the model training process. This can pose a challenge in choosing the best model for a given task. They are also prone to model overfitting, find it difficult to compare different models, and are computationally expensive for large datasets [62, 63].

- Wrapper methods employ a search algorithm to find the subset of features that optimizes the classifier algorithm used. It has also been proven to outperform the filter method for predicting pregnancy and labor contractions [64]. Several wrapper method alternatives are based on how they find the optimum feature subset [62]. Some, such as Recursive Feature Elimination or Sequential Backward Selection, consist of iteratively remove the least important features from the initial whole feature set [62]. Other methods, like Sequential Forward Selection or Forward Selection, start with an empty feature set and iteratively add one feature at each iteration, based on the performance of a machine learning model [62]. Alternatively, methods like Particle Swarm Optimization and Genetic Algorithm evaluate the random feature subset and combine the best to iteratively converge in the optimum feature subset [62]. The literature highlights feature selection by Genetic Algorithm as one of the best choices, as it is faster and achieves a better optimum feature subset than the alternatives [61, 65, 66, 67, 68]. The Genetic Algorithm [69] optimization technique is a heuristic, population-based, algorithmic search method that mimics the human natural evolutionary process. The operations in a genetic algorithm are iterative procedures that manipulate a population of chromosomes (feature subset solution candidates) to produce a new population through genetic functions such as crossover and mutation (similar to Charles Darwin's evolutionary principle of reproduction, genetic recombination and survival of the fittest) [69]. In other words, it initially evaluates random feature subsets by a fitness function, which depends on the classifier, and then, combines the best feature subset (mutation and crossover functions) to generate new feature subsets. In the next iteration, the genetic algorithm uses these new feature subsets together with new random feature subsets. The mutation operator introduces changes in the feature subset (chromosome) by adding or removing some random feature and so creates a new feature subset (children). Crossover involves combining two feature subsets (chromosomes) to create a new feature subset (children). The procedure finishes when the termination condition is met, which may not improve the best optimal subset of features over certain successive iterations [69]. The workflow of feature selection used to predict preterm labor is shown in Figure 1.5. These algorithms have been shown to escape from local minima to reach global minima in complex functions. With this technique, the problems of noise, redundancy and complementarity can be solved, since only those features that provide relevant information for the prediction of preterm birth will be selected. It also solves the problem of complexity due to overfitting of the model, since only one subset of features from the initial set is introduced.

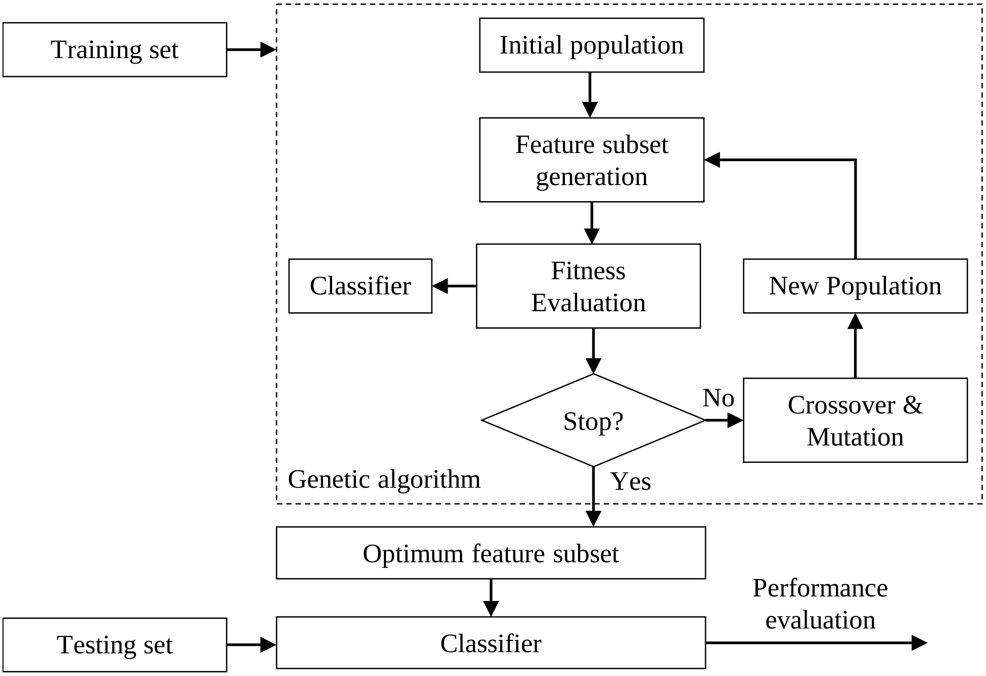


Figure 1.5: Flowchart of feature selection using genetic algorithm.

1.3.3 Resampling Methods for Imbalanced Data Learning

Only around 10% of births occur before 37 weeks of gestation, giving rise to a highly imbalanced data problem, which can significantly impact the performance of classification systems. Classic machine learning algorithms may struggle with imbalanced datasets because they are typically designed to optimize overall accuracy. In an imbalanced dataset, where the majority class greatly outnumbers the minority class, a model that predicts only the majority class can achieve a highly accurate score but would not be good at predicting the minority class, which is often the target in imbalanced datasets [70]. In a medical diagnosis scenario, the cost of a false negative (i.e. failing to diagnose a preterm delivery) may be much higher than the cost of a false positive (i.e. wrongly predicting a preterm delivery).

Self-composition of the database is another issue, i.e. the size and number of samples of each class in the database. It is important to understand why imbalanced databases lead to poorer classification results than balanced ones. The imbalance of a binary database (e.g. preterm vs. term) is determined by comparing the number of observations in both groups. Research suggests that the problem is not just the imbalance of the database, but also the amount of the total sample available [71]. For example, dealing with a total database of 300 samples with a minority class sample is more challenging than a database of 3000 samples with a minority class of 500. There is also a lack of information due to the limited data available in the minority class, which may lead to an insufficient characterization of all the possible variations in the minority class [71, 72].

Several techniques have been described in the literature to address imbalanced data learning, including resampling techniques, ensemble classifier methods, cost-sensitive learning and synthetic data generation [73].

- Resampling techniques obtain a balanced sample set by increasing or decreasing the number of samples of the minority or majority group. Oversampling techniques obtain new synthetic samples from a set of features of the "real" data. These techniques include multiple selection of a sample from the minority class, synthetic generation of samples from the minority class (SMOTE) [74], variants of SMOTE, and Adaptive Synthetic Sampling (ADASYN) generation of samples from the minority class [75], which adds a bias to make them more realistic. Undersampling techniques reduce the number of samples in the majority class to reduce the imbalance. These approaches attempt to maintain the data distribution and improve the visibility of the minority class to help the classifier generate a better decision surface. Techniques such as Tomek Link [76] and the Neighborhood Cleaning Rule [77] are examples of this. Hybrid methods combine oversampling and undersampling techniques to increase or reduce the minority and majority class observations [73].

- Cost-sensitive learning techniques are applied by directly modifying the cost function of a common classifier, thus increasing the cost of misclassifying a certain class. These techniques are useful in applications where an error in the classification of one of the classes has greater weight than the other. For example, in medical diagnosis a false positive is limited to increasing the number of tests to be performed on the patient, while a false negative can be fatal for the patient. Class weighting is usually grouped into cost-sensitive methods and consists of assigning weights to each class to balance the dataset during training [73].

- Ensemble classifier methods use multiple single classifiers to improve the final classification performance. Two fundamental concepts underlie these methods: bagging and boosting. Bagging, which stands for Bootstrap Aggregating, is a method that combines classifiers in which the training input data set has been varied or assigns weights to the observations. Boosting improves the performance of a weak learning algorithm that only performs slightly better than random by iteratively adjusting the weight of each training sample and combining multiple weak learners to create a strong and more reliable algorithm [73].

- Synthetic data generation is used to balance the dataset, for example by using Generative Adversarial Networks (GANs) [78]. Unlike resampling methods, which attempt to achieve balanced data, defined as a set of features, GANs are a type of generative model that can learn to generate synthetic time series data similar to the real training data without relying on explicit modification of the original dataset. GANs cannot be directly grouped into resampling methods for class imbalance, although they can be used to address class imbalance indirectly by generating synthetic data for the minority class, thereby increasing its representation in the training set and improving the model's performance on that class. This approach is called the Synthetic Minority Over-sampling Technique with GANs (SMOTE-GAN) [79].

1.3.4 Performance Metrics for Predicting Preterm Labor

Evaluation metrics are essential in assessing the performance and accuracy of any predictive model, including those used for predicting preterm labor from EHG signals. Since the classifier algorithm's output are usually continuous data ranging from 0 to 1, it is advisable to obtain first the threshold-independent metrics such as AUC and precision-recall to assess the prediction model's performance [53]. The descriptions of the mathematical formulations for various model metrics in which the optimal threshold should be determined are as follows:

- AUC: The area under the receiver operating characteristic curve, which is a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) at various classification thresholds. AUC measures the model's ability to distinguish between preterm and term deliveries across all possible thresholds.

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR(FPR)} \times \mathrm{dFPR} \tag{1.1}$$

  where TPR and FPR are true positive rate and false positive rate, respectively.

- Average Precision (AP): The area under the precision-recall curve, which is a plot of precision against recall (sensitivity) at various classification thresholds. Average precision measures the model's ability to correctly identify positive cases (preterm deliveries) while minimizing false positives.

$$\mathrm{AP} = \sum_n (\mathrm{R}_n - \mathrm{R}_{n-1}) \times \mathrm{P}_n \tag{1.2}$$

  where P_n and R_n are the precision and recall at the nth threshold, respectively.

Unlike ROC curves, which are insensitive to data imbalance, precision-recall curves indicate that poor performance can be masked by ROC curves [80], so that the average precision complements AUC by providing a more comprehensive assessment of the model's performance in scenarios with imbalanced class distribution. Figure 1.6 shows an example of the ROC curve and precision recall curve, in which the dashed line denotes the random classifier with no capability to predict the target classes. The performance of the current model is depicted by the solid line. The ROC curve reaches its maximum performance as it approaches the upper left corner, while the precision-recall curve reaches its peak at the upper right corner. The closer it is to these corners and the farther away from the dashed line the better the model performance.

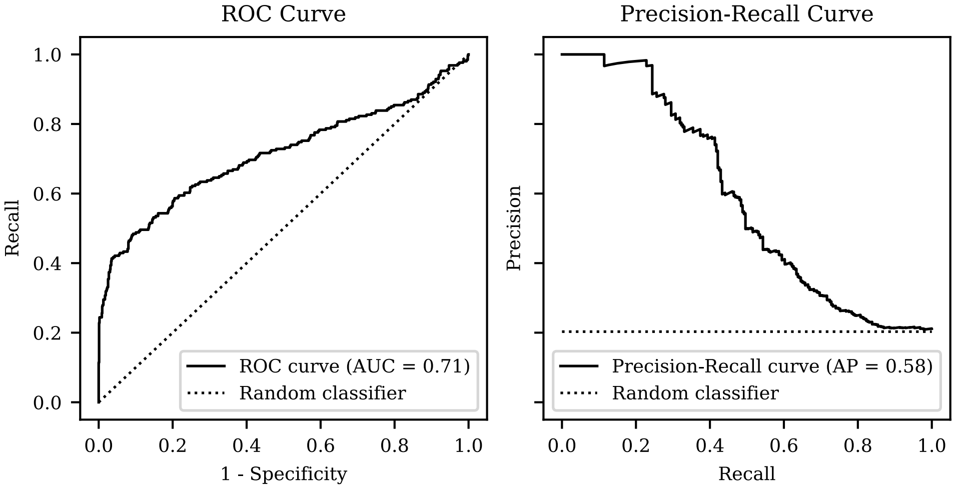


Figure 1.6: Example of an ROC curve (left) and precision-recall curve (right). The dotted lines in the graphs represent the ROC baseline and precision-recall curves (random classifier).

Once the machine learning model has been fully satisfied, a threshold is usually specified to classify the input data into the preterm or term groups. Threshold-dependent metrics are more user-friendly and comprehensible in evaluating performance. After thresholding the classification algorithm's output, we can determine the number of true positives, true negatives, false positives and false negatives (see definition below) taking into account the target outcome and the algorithm's prediction. In the context of predicting preterm labor, TP, TN, FP, and FN are used to describe the results of a binary classification model that predicts whether a delivery will be preterm or term [53].

- True positives (TP) are cases in which the model correctly predicts a preterm delivery.
- True negatives (TN) are cases in which the model correctly predicts a term delivery.
- False positives (FP) are cases in which the model predicts a preterm delivery, but the actual delivery is term.

- False negatives (FN) are cases in which the model predicts a term delivery, but the actual delivery is preterm.

In other words, TP and TN represent correct predictions, while FP and FN are incorrect predictions, which should be minimized. Various metrics can then be computed to evaluate the predictive model's performance. Some of the most frequently used include: accuracy, sensitivity, precision, specificity, negative predictive value (NPV), AUC, and average precision [53, 80, 81]. These metrics help to assess the model's ability to correctly identify positive and negative cases, minimize false positives and distinguish between preterm and term deliveries. The metrics often used to assess the prediction models' performance include accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1-score, and G-mean [53].

- Accuracy: The proportion of correct predictions (both true positives and true negatives) out of all predictions made by the model. In the context of predicting preterm labor, accuracy measures how well the model correctly identifies both preterm and term deliveries.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1.3}$$

- Sensitivity: Also referred to as recall, quantifies the proportion of actual positive cases (preterm deliveries) correctly identified by the model as positive and is the proportion of actual positive cases (preterm deliveries) that are correctly identified by the model as positive. In other words, sensitivity measures how well the model correctly identifies preterm deliveries out of all the actual preterm deliveries.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1.4}$$

- Specificity: The proportion of actual negative cases (term deliveries) that are correctly identified by the model as negative. In other words, specificity measures how well the model correctly identifies term deliveries out of all actual term deliveries.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{1.5}$$

- Precision: It is also referred to as positive predictive value (PPV) and quantifies the proportion of predicted positive cases (preterm deliveries) that are actually positive. Precision measures how often the model is correct when it predicts preterm delivery.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1.6}$$

- Negative predictive value (NPV): The proportion of predicted negative cases (term deliveries) that are actually negative. NPV measures how often the model is correct when it predicts term delivery.

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{1.7}$$

- F1-score, the harmonic mean of precision and sensitivity, combines precision and sensitivity into a single value and ranging from 0 to 1.,Awhere a score of 1 indicates perfect precision and sensitivity. In the context of preterm labor prediction, the F1 score is useful for measuring the model's ability to correctly identify both preterm and term deliveries while minimizing false positives.

$$\text{F1-score} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \tag{1.8}$$

- G-mean calculates the geometric mean of sensitivity and specificity, used to evaluate the model's ability to identify all positive cases while minimizing false positives. In the context of predicting preterm labor, G-mean is useful in assessing the model's performance in correctly identifying preterm deliveries while minimizing false positives. It should be noted that both F1-score and G-mean are generally used to assess the prediction models' performance in imbalance data learning [80, 82, 83, 84].

$$\text{G-mean} = \sqrt{\text{recall} \times \text{specificity}} \tag{1.9}$$

1.4 State of the art of preterm delivery prediction using EHG signals

Many different approaches have been proposed to predict preterm labor from EHG signals. The goal of the present review is to provide an overview of the current state of this field of research and covers the different approaches proposed, including both feature extraction, dimensionality reduction and machine learning, putting special emphasis on the approach to mitigate imbalance data problem. The review also includes the strengths and limitations of these approaches. The information given in Table 1.1 comprises the studies conducted since 2018, providing details such as the number of preterm and term records in the dataset, the dimensional reduction method employed, the resampling strategy used (whether the dataset was resampled before or after partitioning, i.e. data resampling-splitting or data splitting-resampling pipeline), and the classifier used.

The first studies include those that did not use any strategy to deal with imbalanced data and highlight their limitations and then focus on the different strategies used to mitigate the imbalanced data learning problem.

1.4.1 Studies with no strategy to deal with imbalanced data learning

Due to the nature of the study, few studies attempted to predict preterm labor by EHG without any strategy to deal with imbalance preterm and term data. Selvaraju et al. extracted temporal and spectral features from the TPEHG DB to train a random forest classifier, reporting an accuracy of 75.2%, F1-score of 34.6%, and AUC of 67.7% in testing [85]. Goldsztejn & Nehorai used EHG records from the TPEHG DB and TPEHGT DS databases on women with over 28 weeks of gestation. They fed a Recurrent Neural Network classifier (RNN) with a time-frequency domain short-time Fourier transform of EHG signals and clinical data, reporting a sensitivity of 77.3%, with a specificity of 70%, AUC 80%, and PPV 30.9% [86].

Mohammadi Far et al. achieved an excellent performance using only three features obtained by the Empirical Mode Decomposition (EMD) to split the EHG signal into the first two intrinsic mode functions (IMF) amplitudes, plus a support vector machine classifier, with a sensitivity of 99.5%, specificity of 99.7%, and AUC of 99.9% [87, 88]. In this regard, the first two IMFs, which represent the highest frequency component of EHG signals, are probably outside the main bandwidth, in which the EHG signal distributes its energy (0.2–1 Hz). Several studies reported that peak-to-peak amplitude is an unreliable metric for predicting preterm labor in cross-sectional studies, as it is influenced by various factors such as body mass index, age and skin preparation, among others [18, 46, 89]. Other studies [57, 87, 90] showed that similar features obtained from EMD can distinguish between preterm and term births, but involve sample entropy being the best performing feature (AUC of 69.0 ±4.2%)[57]. Further studies are thus still needed to validate the methodology proposed by Mohammadi Far et al. Moreover, Shahrdad & Amirani also achieved high performance by characterizing EHG signals from TPEHG DB with linear predictive coding and fitting an RUSBoost classifier (random undersampling + AdaBoost), obtaining a sensitivity of 90%, specificity of 89%, and AUC of 97% [91]. However, it is unclear whether the reported performance was obtained for training, validation and test data due to a lack of information.

1.4.2 Resampling techniques for imbalanced data learning

The most common approach to address the imbalanced data problem in predicting preterm delivery from EHG signals involves oversampling the minority class (preterm) using techniques such as SMOTE or ADASYN. Most recent studies oversample the whole database (preterm and term) before splitting into training and testing partition [57]. For example, the TPEHG DB [46]. contains 38 preterm and 262 term labor observations, resulting in an imbalanced dataset. The dataset can be balanced to 262 preterm and 262 term labor cases by oversampling. The balanced samples are then split into balanced training and testing partitions to resolve the imbalance problem.

Following this scheme, in a study by Fergus et al., various machine learning techniques were evaluated by different sets of EHG signal parameters. The minority class was oversampled using SMOTE, and the resulting decision tree classifier achieved high scores of 90% sensitivity, 83% specificity, and 89% AUC on the TPEHGT DS database [92]. Ahmed et al. also oversampled the minority class with ADASYN and achieved an accuracy of 99% and an AUC of 95.4% using a support vector classifier on the same database [93]. Peng et al. attempted to predict preterm births from EHG recordings obtained after the 26th week of gestation from the TPEHG DB. They extracted linear, spectral, and non-linear features from the signals and resampled them using ADASYN. The subsequent classification by a random forest (RF) algorithm achieved 88% sensitivity, 97% specificity and 88% AUC [94]. Xu et al. used the TPEHG DB to predict preterm labor by extracting linear, spectral, and non-linear features and balancing the dataset by SMOTE. They used a Gradient Boosting Classifier (GBC) and reported an overall accuracy of 85%, Gmean of 84%, and AUC of 91% [95]. Khan et al. also used the TPEHG DB and extracted a similar set of features using ADASYN to balance the dataset and the support vector machine to fit the model, obtaining an accuracy of 95.5%, sensitivity of 93.5%, and specificity of 97.1% [96]. Mas-Cabo et al. used a similar approach to predict preterm labor from the TPEHG DB. They addressed the issue of imbalanced data using SMOTE and fitted an Artificial Neural Network (ANN) to solve the classification task. The results reported a sensitivity of 84.4 ±2.5%, specificity of 89.2 ±2.1%, and AUC of 91.1 ±2.6% [97].

Despite the promising results, recent research shows that the previous studies may have overestimated the performance of preterm labor prediction system due to a methodological bias, i.e. oversampling techniques before data splitting (data resampling-splitting pipeline) can lead to correlation data structures between the training and test data, giving rise to unrealistic performance estimates [57] (see Figure 1.7). Indeed, they repeated the experiment carried out by Fergus et al. and Ahmed et al., using the oversampling technique after data splitting (data splitting-resampling pipeline), obtaining AUC values of 60.75% and 56.04%, respectively [57], which were much lower than those obtained by the data resampling-splitting pipeline.

The influence of resampling before or after data splitting on model performance was also reported by other research groups [98, 99]. Xu et al. developed a support vector machine model using EHG signals from the TPEHG DB and found that data resampling-splitting pipeline led to a sensitivity of 99.5 ±1.2%, specificity of 99.2 ±0.2% and AUC of 94.3 ±2.1%. However, when oversampling was performed after data splitting, the model achieved a sensitivity of 89.1 ±2.6%, specificity of 92.5 ±1.2% and AUC of 96.8 ±6% [98]. Note that in this work feature selection was performed prior to data partitioning, potentially leading to information leakage for prediction models. Cross-validation was performed by k-fold rather than stratified k-fold, which does not guarantee the proportion of the groups, giving rise to imbalanced partitions [98]. Lou et al. compared the performance of preterm delivery prediction models using different resampling schemes. They fitted a Gaussian Naïve Bayes (GNB) classifier using the same database. Using the data resampling-splitting

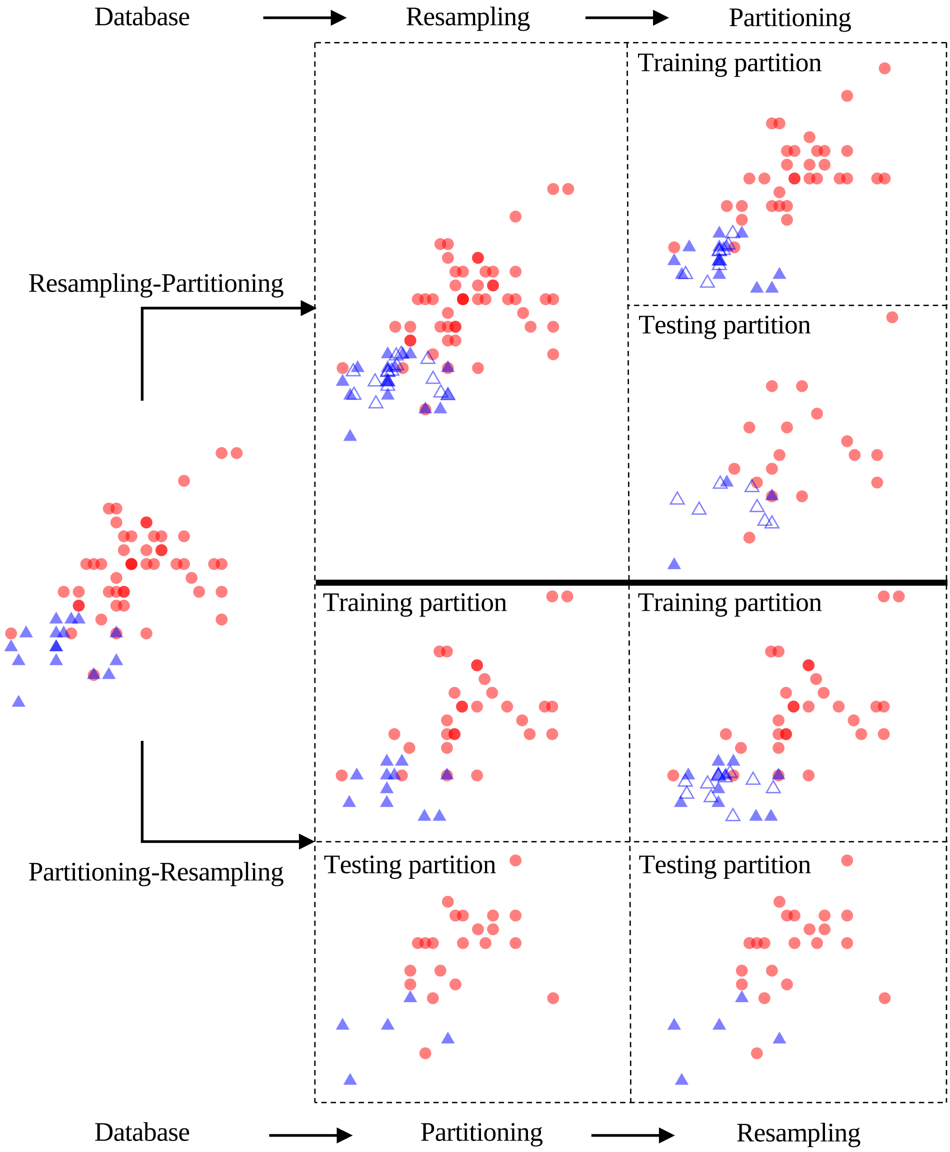


Figure 1.7: Comparison of the effect of applying oversampling before and after data separation in a two-dimensional artificial classification problem.

pipeline, the classifier achieved a sensitivity of 98 ±1%, specificity of 91 ±4% and AUC of 98 ±1%. In contrast, when partitioning preceded oversampling, sensitivity was 84 ±10%, specificity 66 ±6% and AUC 84 ±7% [99].

1.4.3 Matching the composition of the preterm and term database

Other studies attempted to balance the database during the study design stage and matched "artificially" preterm and term women, which did not preserve the original distribution between classes (∼15% preterm labor records). Chen et al. used the Icelandic database to extract 150 EHG signal segments for both preterm and term classes and used components of wavelet transformation and sample entropy as input features for the stacked sparse autoencoder classifier (SSAE). Their model achieved a sensitivity of 92%, specificity of 88% and AUC of 90% [100]. Mischi et al. studied different features to distinguish between preterm and term births using their own database of EHG recordings from women with symptoms of threatened preterm labor at the time of recording (34 preterm and 20 term registers). They found that modified approximate entropy was able to predict preterm labor with a sensitivity of 68 ±23%, specificity of 82 ±18% and AUC of 75.4% [101]. Cheng et al. used an extended version of Mischi et al.'s database with 44 preterm and 30 term recordings and characterized the EHG signals with temporal, spectral, non-linear and graph features. They used the support vector machine to fit the classification model and achieved a sensitivity of 93.2%, specificity of 86.7% and AUC of 84.2% [102]. Romero-Morales et al. selected 17 preterm and 17 term EHG registers from the TPEHG DB and TPEHGT DS, extracted temporal, spectral and non-linear features and used the Quadratic Support Vector Machine to classify them. They reported an accuracy of 88.52 ±1.47%, sensitivity of 83.83 ±3.07% and specificity of 93.22 ±1.31% [103]. Finally, Chen & Xu extracted 20 entropy metrics from 450 EHG signal segments of 51.2s from 13 term and 13 preterm women in the TPEHGT DS, i.e. each patient was represented by 450×20 features, and designed the classifier using a Sparse Autoencoder (SAE) followed by a Deep Neural Network (DNN), achieving a sensitivity of 98.02%, specificity of 97.74% and AUC of 97.89% [104]. Despite the promising results, future studies are still needed to further corroborate the method, due to the possible overfitting problem when attempting a high dimensional feature space, especially when the sample size is extremely low.

Table 1.1: Features taken from studies that used EHG parameters for preterm labor diagnosis.

Ref.	Dataset	No. of preterm / term records		Input data	Dimensionality reduction method	Resampling strategy (R-P / P-R)	Classifier	Performance (%)
[91]	TPEHG DB	38	262	LPC + Clinical data	None	RUS (P-R)	RUSBoost	AUC = 97
[101]	Self-database (58 patients, contractions)	34 TPL	20 TPL	Modified ApEn	None	None	Statistical class	AUC = 75.4
[100]	Icelandic database	150 pregnancy	150 labor	Components of wavelet transformation + sample entropy	None	None	SSAE	AUC = 90
[96]	TPEHG DB	38	262	Temporal, spectral and non-linear features	None	ADASYN (R-P)	SVM	Acc = 95.5
[88]	TPEHG DB	38	262	One feature from EMD: RMS	None	None	SVM	Acc = 99.6
[97]	TPEHG DB	38	262	Temporal, spectral and non-linear features	PCA	SMOTE (R-P)	ANN	AUC = 91.1 ±2.6
[94]	TPEHG DB (Only early)	19	143	Temporal, spectral and non-linear features	None	ADASYN (R-P)	RF	AUC = 88
[104]	TPEHGT DS	450 samples of 51.2s	450 samples of 51.2s	20 entropy features	None	None	SAE + DNN	AUC = 97.89
[85]	TPEHG DB	38	262	Temporal and spectral features	None	ADASYN (P-R)	RF	AUC = 67.7
[95]	TPEHG DB	38	262	Temporal, spectral and non-linear features	None	SMOTE (R-P)	GBC	AUC = 91

[57]	TPEHG DB	38	262	Temporal, spectral and non-linear features	None	SMOTE-Tomek (R-P)	QDA	AUC = 65.33
[86]	TPEHG DB + TPEHGT DS (>28^{th} WOG)	28	123	STFT + clinical data	None	None	RNN	AUC = 80
[98]	TPEHG DB	38	262	Network Theory Features	Filter Feature Selection	R-P: SMOTE P-R: SMOTE training and testing partition	SVM	R-P: AUC = 94.3±2.1 P-R: AUC = 96.8 ±6.0
[99]	TPEHG DB	38	262	STFT	PCA	R-P: SMOTE P-R: SMOTE training and testing partition	GNB	R-P: AUC = 98 ±1 P-R: AUC = 84 ±7
[87]	TPEHG DB	38	262	Three features from EMD	None	None	SVM	AUC = 99.9
[102]	Self-database (58 nonlabor / 16 labor)	44 TPL	30 TPL	Temporal, spectral, non-linear and graph features	Sequential Forward Selection (SFS)	None	SVM	AUC = 84.2
[103]	TPEHG DS + TPEHGT DB (27-33^{rd} WOG)	17	17	Temporal, spectral and non-linear features	Feature selection (F-test, chi-test, linear regression, and sequential selection)	None	QSVM	AUC = 89 ±2

Note: The abbreviations used in the table are as follows: Acc - Accuracy, ADASYN - Adaptive Synthetic Sampling, ANN - Artificial Neural Network, DNN - Deep Neural Network, EMD - Empirical Mode Decomposition, GBC - Gradient Boosting Classifier, GNB - Gaussian Naive Bayes, LPC - Linear Predictive Coding, PCA - Principal Component Analysis, P-R - Partition-Resampling Approach, QSVM - Quantum Support Vector Machine, RNN - Recurrent Neural Network, RF - Random Forest, RMS - Root Mean Square, R-P - Resampling-Partition Approach, SVM - Support Vector Machine, STFT - Short-Time Fourier Transform, TPEHG DB - Term-Preterm EHG Database, TPEHGT DS - Term-Preterm EHG DataSet with Tocogram, TPL - Threatened Preterm Labor, QDA - Quadratic Discriminant Analysis, RUS - Random Undersampling, RUSBoost - Random Undersampling Boosting, SMOTE - Synthetic Minority Over-sampling Technique, WOG - Week of Gestation.

1.5 References

[1] S. Chawanpaiboon *et al.*, Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis, *The Lancet Global Health*, vol. 7, no. 1, e37–e46, 2019, ISSN: 2214109X. DOI: 10.1016/S2214-109X(18)30451-0.

[2] J. A. Martin and M. J. K. Osterman, Describing the Increase in Preterm Births in the United States, 2014-2016 Key findings Data from the National Vital Statistics System, *NCHS Data Brief*, vol. 312, no. 312, pp. 1–29, 2018.

[3] C. P. Howson, M. V. Kinney, L. McDougall, and J. E. Lawn, Born Too Soon: Preterm birth matters, *Reproductive Health*, vol. 10, no. SUPPL. 1, pp. 1–9, 2013, ISSN: 17424755. DOI: 10.1186/1742-4755-10-S1-S1.

[4] C. Crump, J. Sundquist, M. A. Winkleby, and K. Sundquist, Gestational age at birth and mortality from infancy into mid-adulthood: a national cohort study, *The Lancet Child and Adolescent Health*, vol. 3, no. 6, pp. 408–417, 2019, ISSN: 23524642. DOI: 10.1016/S2352-4642(19)30108-7.

[5] C. Leung, Born too soon, *Neuroendocrinology Letters*, vol. 25, no. SUPPL. 1, J. L. CP Howson, MV Kinney, Ed., pp. 133–136, 2004, ISSN: 0172780X. DOI: 10.2307/3965140.

[6] R. M. Patel, Short and Long-Term Outcomes for Extremely Preterm Infants, *American Journal of Perinatology*, vol. 33, no. 03, pp. 318–328, 2016, ISSN: 1472-3263. DOI: 10.1055/s-0035-1571202.Short.

[7] R. H. Paul *et al.*, Trends in Care Practices, Morbidity, and Mortality of Extremely Preterm Neonates, 1993–2012, vol. 34, no. 5, pp. 737–748, 2021. DOI: 10.1001/jama.2015.10244.Trends.

[8] P. Y. Ancel *et al.*, Survival and Morbidity of Preterm Children Born at 22 Through 34Weeks' Gestation in France in 2011 Results of the EPIPAGE-2 Cohort Study, *JAMA Pediatrics*, vol. 169, no. 3, pp. 230–238, 2015, ISSN: 21686211. DOI: 10.1001/jamapediatrics.2014.3351.

[9] P. Rojas Feria, A. Pavón Delgado, M. Rosso González, and A. Losada Martínez, Complicaciones a corto plazo de los recién nacidos pretérmino tardíos, *Anales de Pediatria*, vol. 75, no. 3, pp. 169–174, 2011, ISSN: 16954033. DOI: 10.1016/j.anpedi.2011.04.001.

[10] F. Wu *et al.*, Short-term outcomes of extremely preterm infants at discharge: A multicenter study from Guangdong province during 2008-2017, *BMC Pediatrics*, vol. 19, no. 1, pp. 1–11, 2019, ISSN: 14712431. DOI: 10.1186/s12887-019-1736-8.

[11] C. S. H. Aarnoudse-Moens, N. Weisglas-Kuperus, J. B. Van Goudoever, and J. Oosterlaan, Meta-analysis of neurobehavioral outcomes in very preterm and/or very low birth weight children, *Pediatrics*, vol. 124, no. 2, pp. 717–728, 2009, ISSN: 00314005. DOI: 10.1542/peds.2008-2816.

[12] S. Petrou, H. H. Yiu, and J. Kwon, Economic consequences of preterm birth: A systematic review of the recent literature (2009-2017), *Archives of Disease in Childhood*, vol. 104, no. 5, pp. 456–465, 2019, ISSN: 14682044. DOI: 10.1136/archdischild-2018-315778.

[13] N. X. Thanh, J. Toye, A. Savu, M. Kumar, and P. Kaul, Health Service Use and Costs Associated with Low Birth Weight - A Population Level Analysis, *Journal of Pediatrics*, vol. 167, no. 3, pp. 551–556, 2015, ISSN: 10976833. DOI: 10.1016/j.jpeds.2015.06.007.

[14] M. Katz, K. Goodyear, and R. K. Creasy, Early signs and symptoms of preterm labor. *American journal of obstetrics and gynecology*, vol. 162, no. 5, pp. 1150–3, 1990, ISSN: 0002-9378. DOI: 10.1016/0002-9378(90)90004-q.

[15] J. D. Iams, F. F. Johnson, and M. Parker, A prospective evaluation of the signs and symptoms of preterm labor, *Obstetrics and Gynecology*, vol. 84, no. 2, pp. 227–230, 1994, ISSN: 1873233X.

[16] I. B. Fuchs, W. Henrich, K. Osthues, and J. W. Dudenhausen, Sonographic cervical length in singleton pregnancies with intact membranes presenting with threatened preterm labor, *Ultrasound in Obstetrics and Gynecology*, vol. 24, no. 5, pp. 554–557, 2004, ISSN: 09607692. DOI: 10.1002/uog.1714.

[17] R. Romero, J. Espinoza, L. F. Gonçalves, J. P. Kusanovic, L. A. Friel, and J. K. Nien, Inflammation in preterm and term labour and delivery, *Seminars in Fetal and Neonatal Medicine*, vol. 11, no. 5, pp. 317–326, 2006, ISSN: 1744165X. DOI: 10.1016/j.siny.2006.05.001.

[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, 02TR01, 2018, ISSN: 13616579. DOI: 10.1088/1361-6579/aaad56.

[19] J. Chen, G. Gong, W. Zheng, J. Xu, X. Luo, and Y. Zhang, Diagnostic accuracy of quantitative fetal fibronectin to predict spontaneous preterm birth: A meta-analysis, *International Journal of Gynecology and Obstetrics*, vol. 153, no. 2, pp. 220–227, 2021, ISSN: 18793479. DOI: 10.1002/ijgo.13550.

[20] S. O'Hara, M. Zelesco, and Z. Sun, Cervical length for predicting preterm birth and a comparison of ultrasonic measurement techniques, *Australasian Journal of Ultrasound in Medicine*, vol. 16, no. 3, pp. 124–134, 2013, ISSN: 1836-6864. DOI: 10.1002/j.2205-0140.2013.tb00100.x.

[21] M. T. Mella and V. Berghella, Prediction of Preterm Birth: Cervical Sonography, *Seminars in Perinatology*, vol. 33, no. 5, pp. 317–324, 2009, ISSN: 01460005. DOI: 10.1053/j.semperi.2009.06.007.

[22] M. Son and E. S. Miller, Predicting preterm birth: Cervical length and fetal fibronectin, *Seminars in Perinatology*, vol. 41, no. 8, pp. 445–451, 2017, ISSN: 1558075X. DOI: 10.1053/j.semperi.2017.08.002.

[23] P. Taipale and V. Hiilesmaa, Sonographic measurement of uterine cervix at 18-22 weeks' gestation and the risk of preterm delivery, *Obstetrics and Gynecology*, vol. 92, no. 6, pp. 902–907, 1998, ISSN: 00297844. DOI: 10.1016/S0029-7844(98)00346-9.

[24] E. R. Guzman *et al.*, A comparison of sonographic cervical parameters in predicting spontaneous preterm birth in high-risk singleton gestations, *Ultrasound in Obstetrics and Gynecology*, vol. 18, no. 3, pp. 204–210, 2001, ISSN: 09607692. DOI: 10.1046/j.0960-7692.2001.00526.x.

[25] M. Sean Esplin *et al.*, Predictive accuracy of serial transvaginal cervical lengths and quantitative vaginal fetal fibronectin levels for spontaneous preterm birth among nulliparous women, *Obstetrical and Gynecological Survey*, vol. 72, no. 7, pp. 397–399, 2017, ISSN: 15339866. DOI: 10.1097/OGX.0000000000000455.

[26] T. Włodarczyk *et al.*, Spontaneous Preterm Birth Prediction Using Convolutional Neural Networks, *Lecture Notes in Computer Science*, vol. 12437 LNCS, pp. 274–283, 2020, ISSN: 16113349. DOI: 10.1007/978-3-030-60334-2_27. arXiv: 2008.07000.

[27] Z. Shahshahan and O. Rasouli, The use of maternal C-reactive protein in the predicting of preterm labor and tocolytic therapy in preterm labor women, *Advanced Biomedical Research*, vol. 3, no. 1, p. 154, 2014, ISSN: 2277-9175. DOI: 10.4103/2277-9175.137864.

[28] R. Nikbakht, E. K. Moghadam, and Z. Nasirkhani, Maternal serum levels of C-reactive protein at early pregnancy to predict fetal growth restriction and preterm delivery: A prospective cohort study, *International Journal of Reproductive BioMedicine*, vol. 18, no. 3, pp. 157–164, 2020, ISSN: 24763772. DOI: 10.18502/ijrm.v18i3.6710.

[29] C. J. Lockwood, A. Ghidini, R. Wein, R. Lapinski, D. Casal, and R. L. Berkowitz, Increased interleukin-6 concentrations in cervical secretions are associated with preterm delivery, *American Journal of Obstetrics and Gynecology*, vol. 171, no. 4, pp. 1097–1102, 1994, ISSN: 00029378. DOI: 10.1016/0002-9378(94)90043-4.

[30] N. Ö. Turhan, A. Karabulut, and B. Adam, Maternal serum interleukin 6 levels in preterm labor: Prediction of admission-to-delivery interval, *Journal of Perinatal Medicine*, vol. 28, no. 2, pp. 133–139, 2000, ISSN: 03005577. DOI: 10.1515/JPM.2000.018.

[31] A. Conde-Agudelo and R. Romero, Does vaginal progesterone prevent recurrent preterm birth in women with a singleton gestation and a history of spontaneous preterm birth? Evidence from a systematic review and meta-analysis, *American Journal of Obstetrics and Gynecology*, vol. 227, no. 3, 440–461.e2, 2022, ISSN: 10976868. DOI: 10.1016/j.ajog.2022.04.023.

[32] T. Y. Euliano, M. T. Nguyen, S. Darmanjian, J. D. Busowski, N. Euliano, and A. R. Gregg, Monitoring Uterine Activity during Labor: Clinician Interpretation of Electrohysterography versus Intrauterine Pressure Catheter and Tocodynamometry, *American Journal of Perinatology*, vol. 33, no. 9, pp. 831–838, 2016, ISSN: 10988785. DOI: 10.1055/s-0036-1572425.

[33] R. E. Garfield, K. Chwalisz, L. Shi, G. Olson, and G. R. Saade, Instrumentation for the diagnosis of term and preterm labour, *Journal of Perinatal Medicine*, vol. 26, no. 6, pp. 413–436, 1998, ISSN: 0300-5577. DOI: 10.1515/jpme.1998.26.6.413.

[34] R. E. Garfield *et al.*, Methods and devices for the management of term and preterm labor, *Annals of the New York Academy of Sciences*, vol. 943, pp. 203–224, 2001, ISSN: 00778923. DOI: 10.1111/j.1749-6632.2001.tb03803.x.

[35] R. E. Garfield and W. L. Maner, Physiology and electrical activity of uterine contractions, *Seminars in Cell and Developmental Biology*, vol. 18, no. 3, pp. 289–295, 2007, ISSN: 10849521. DOI: 10.1016/j.semcdb.2007.05.004.

[36] B. M. Sanborn, Hormones and calcium: Mechanisms controlling uterine smooth muscle contractile activity, *Experimental Physiology*, vol. 86, no. 2, pp. 223–237, 2001, ISSN: 09580670. DOI: 10.1113/eph8602179.

[37] S. Wray and C. Prendergast, *The Myometrium: From Excitation to Contractions and Labour*. 2019, vol. 1124, pp. 233–263, ISBN: 9789811358951. DOI: 10.1007/978-981-13-5895-1_10.

[38] N. Demianczuk, M. E. Towell, and R. E. Garfield, Myometrial electrophysiologic activity and gap junctions in the pregnant rabbit, *American Journal of Obstetrics and Gynecology*, vol. 149, no. 5, pp. 485–491, 1984, ISSN: 00029378. DOI: 10.1016/0002-9378(84)90021-8.

[39] 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, ISSN: 00029378. DOI: 10.1016/0002-9378(93)90456-S.

[40] J. Gondry, C. Marque, J. Duchene, and D. Cabrol, Electrohysterography during pregnancy: Preliminary report, *Biomedical Instrumentation and Technology*, vol. 27, no. 4, pp. 318–324, 1993, ISSN: 08998205.

[41] D. Schlembach, W. L. Maner, R. E. Garfield, and H. Maul, Monitoring the progress of pregnancy and labor using electromyography, *European Journal of Obstetrics and Gynecology and Reproductive Biology*, vol. 144, no. SUPPL 1, pp. 2–8, 2009, ISSN: 18727654. DOI: 10.1016/j.ejogrb.2009.02.016.

[42] G. M. Wolfs and M. van Leeuwen, Electromyographic Observations on the Human Uterus during Labour, *Acta Obstetricia et Gynecologica Scandinavica*, vol. 58, no. 90 S, pp. 1–61, 1979, ISSN: 16000412. DOI: 10.3109/00016347909156375.

[43] R. E. Garfield, L. Murphy, K. Gray, and B. Towe, Review and Study of Uterine Bioelectrical Waveforms and Vector Analysis to Identify Electrical and Mechanosensitive Transduction Control Mechanisms During Labor in Pregnant Patients, *Reproductive Sciences*, vol. 28, no. 3, pp. 838–856, 2021, ISSN: 19337205. DOI: 10.1007/s43032-020-00358-5.

[44] J. M. MARSHALL, Regulation of activity in uterine smooth muscle. *Physiological reviews. Supplement*, vol. 5, pp. 213–27, 1962, ISSN: 0554-1395.

[45] F. Jager, S. Libenšek, and K. Geršak, Characterization and automatic classification of preterm and term uterine records, *PLoS ONE*, vol. 13, no. 8, O. Uthman, Ed., e0202125, 2018, ISSN: 19326203. DOI: 10.1371/journal.pone.0202125.

[46] 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 pre-term delivery groups, *Medical and Biological Engineering and Computing*, vol. 46, no. 9, pp. 911–922, 2008, ISSN: 01400118. DOI: 10.1007/s11517-008-0350-y.

[47] F. Nieto-del-amor, R. Beskhani, Y. Ye-lin, J. Garcia-casado, and A. Diaz-martinez, Assessment of Dispersion and Bubble Entropy Measures for Enhancing Preterm Birth Prediction Based on Electrohysterographic Signals, *Sensors*, vol. 21, no. 18, 2021. DOI: 10.3390/s21186071.

[48] F. Nieto-del-Amor *et al.*, Combination of Feature Selection and Resampling Methods to Predict Preterm Birth Based on Electrohysterographic Signals from Imbalance Data, *Sensors*, vol. 22, no. 14, p. 5098, 2022, ISSN: 1424-8220. DOI: 10.3390/s22145098.

[49] J. Liang, J. Y. Cheung, and J. D. Chen, Detection and deletion of motion artifacts in electrogastrogram using feature analysis and neural networks, *Annals of Biomedical Engineering*, vol. 25, no. 5, pp. 850–857, 1997, ISSN: 00906964. DOI: 10.1007/BF02684169.

[50] M. A. Verhagen, L. J. Van Schelven, M. Samsom, and A. J. Smout, Pitfalls in the analysis of electrogastrographic recordings, *Gastroenterology*, vol. 117, no. 2, pp. 453–460, 1999, ISSN: 00165085. DOI: 10.1053/gast.1999.0029900453.

[51] Y. Ye-Lin, J. Garcia-Casado, G. Prats-Boluda, J. Alberola-Rubio, and A. Perales, Automatic identification of motion artifacts in EHG recording for robust analysis of uterine contractions, *Computational and Mathematical Methods in Medicine*, vol. 2014, 2014, ISSN: 1748670X. DOI: 10.1155/2014/470786.

[52] J. Xu, Z. Chen, H. Lou, G. Shen, and A. Pumir, Review on EHG signal analysis and its application in preterm diagnosis, *Biomedical Signal Processing and Control*, vol. 71, pp. 1–18, 2022, ISSN: 17468108. DOI: 10.1016/j.bspc.2021.103231.

[53] K. P. Murphy, *Machine Learning: A Probabilistic Perspective.* 2012, pp. 1–100, ISBN: 978026208029.

[54] A. Rajkomar, J. Dean, and I. Kohane, Machine Learning in Medicine, *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019, ISSN: 0028-4793. DOI: 10.1056/nejmra1814259.

[55] R. Aggarwal *et al.*, Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis, *npj Digital Medicine*, vol. 4, no. 1, 2021, ISSN: 23986352. DOI: 10.1038/s41746-021-00438-z.

[56] T. Włodarczyk *et al.*, Machine learning methods for preterm birth prediction: A review, *Electronics (Switzerland)*, vol. 10, no. 5, pp. 1–24, 2021, ISSN: 20799292. DOI: 10.3390/electronics10050586.

[57] G. Vandewiele *et al.*, Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling, *Artificial Intelligence in Medicine*, vol. 111, p. 101 987, 2021, ISSN: 18732860. DOI: 10.1016/j.artmed.2020.101987. arXiv: 2001.06296.

[58] L. O. Jimenez and D. A. Landgrebe, Supervised classification in high-dimensional space: Geometrical, statistical, and asymptotical properties of multivariate data, *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 28, no. 1, pp. 39–54, 1998, ISSN: 10946977. DOI: 10.1109/5326.661089.

[59] L. Van Der Maaten, E. Postma, and J. Van den Herik, Dimensionality reduction: a comparative review, *J Mach Learn Res*, vol. 10, pp. 66–71, 2009.

[60] G. Doquire and M. Verleysen, A comparison of multivariate mutual information estimators for feature selection, in ICPRAM 2012 - Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods, vol. 1, 2012, pp. 176–185, ISBN: 9789898425980. DOI: 10.5220/0003726101760185.

[61] S. B. Kotsiantis, Feature selection for machine learning classification problems: A recent overview, *Artificial Intelligence Review*, vol. 42, no. 1, p. 157, 2014, ISSN: 02692821. DOI: 10.1007/s10462-011-9230-1.

[62] B. Remeseiro and V. Bolon-Canedo, A review of feature selection methods in medical applications, *Computers in Biology and Medicine*, vol. 112, no. July, p. 103 375, 2019, ISSN: 18790534. DOI: 10.1016/j.compbiomed.2019.103375.

[63] S. Biswas, M. Bordoloi, and B. Purkayastha, Review on Feature Selection and Classification using Neuro-Fuzzy Approaches, *International Journal of Applied Evolutionary Computation*, vol. 7, no. 4, pp. 28–44, 2016, ISSN: 1942-3594. DOI: 10.4018/ijaec.2016100102.

[64] D. Alamedine, M. Khalil, and C. Marque, Comparison of Feature selection for Monopolar and Bipolar EHG signal, in Journees Recherche en Imagerie et Technologies pour la Santé (RITS 2015), 2015, pp. 100–101.

[65] P. Asvestas *et al.*, Use of genetic algorithm for the selection of EEG features, *Journal of Physics: Conference Series*, vol. 633, no. 1, 2015, ISSN: 17426596. DOI: 10.1088/1742-6596/633/1/012123.

[66] W. Bouaguel, A New Approach for Wrapper Feature Selection Using Genetic Algorithm for Big Data, in Intelligent and Evolutionary Systems, 2016, pp. 75–83. DOI: 10.1007/978-3-319-27000-5_6.

[67] D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. 1989, vol. 27, ISBN: 9780201157673.

[68] M. W. Huang, C. H. Chiu, C. F. Tsai, and W. C. Lin, On combining feature selection and over-sampling techniques for breast cancer prediction, *Applied Sciences (Switzerland)*, vol. 11, no. 14, 2021, ISSN: 20763417. DOI: 10.3390/app11146574.

[69] O. Babatunde, L. Armstrong, J. Leng, and D. Diepeveen, A Genetic Algorithm-Based Feature Selection, *International Journal of Electronics Communication and Computer Engineering*, vol. 5, no. 4, pp. 899–905, 2014, ISSN: 2278-4209.

[70] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, Learning from class-imbalanced data: Review of methods and applications, *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017, ISSN: 09574174. DOI: 10.1016/j.eswa.2016.12.035.

[71] M. Denil and T. Trappenberg, Overlap versus imbalance, in Canadian conference on artificial intelligence, A. Farzindar and V. Kešelj, Eds., vol. 6085 LNAI, 2010, pp. 220–231, ISBN: 3642130585. DOI: 10.1007/978-3-642-13059-5_22.

[72] T. Jo and N. Japkowicz, Class imbalances versus small disjuncts, *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 40–49, 2004. DOI: https://doi.org/10.1145/1007730.1007737.

[73] A. Fernández, S. García, M. Galar, and R. C. Prati, *Learning from Imbalanced Data Sets*. 2019, ISBN: 9783319980737.

[74] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, ISSN: 10769757. DOI: 10.1613/jair.953. arXiv: 1106.1813.

[75] H. He, Y. Bai, E. A. Garcia, and S. Li, ADASYN: Adaptive synthetic sampling approach for imbalanced learning, in Proceedings of the International Joint Conference on Neural Networks, 2008, pp. 1322–1328, ISBN: 9781424418213. DOI: 10.1109/IJCNN.2008.4633969.

[76] I. Tomek, Tomek Link: Two Modifications of CNN, *IEEE Trans. Systems, Man and Cybernetics*, pp. 769–772, 1976.

[77] J. Laurikkala, Improving Identification of Difficult Small Classes by Balancing Class Distribution, in Conference on artificial intelligence in medicine in Europe, June 2001, vol. 2101, 2001, pp. 63–66, ISBN: 3540422943. DOI: 10.1007/3-540-48229-6_9.

[78] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, Generative Adversarial Networks: An Overview, no. January, pp. 53–65, 2018.

[79] A. Sharma, P. K. Singh, and R. Chandra, SMOTified-GAN for Class Imbalanced Pattern Classification Problems, *IEEE Access*, vol. 10, pp. 30 655–30 665, 2022, ISSN: 21693536. DOI: 10.1109/ACCESS.2022.3158977. arXiv: 2108.03235.

[80] L. A. Jeni, J. F. Cohn, and F. De La Torre, Facing imbalanced data - Recommendations for the use of performance metrics, in 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, 2013, pp. 245–251, ISBN: 9780769550480. DOI: 10.1109/ACII.2013.47.

[81] H. R. Sofaer, J. A. Hoeting, and C. S. Jarnevich, The area under the precision-recall curve as a performance metric for rare binary events, *Methods in Ecology and Evolution*, vol. 10, no. 4, pp. 565–577, 2019, ISSN: 2041210X. DOI: 10.1111/2041-210X.13140.

[82] T. Saito and M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, *PLoS ONE*, vol. 10, no. 3, pp. 1–21, 2015, ISSN: 19326203. DOI: 10.1371/journal.pone.0118432.

[83] S. Vluymans, Learning from imbalanced data, *Studies in Computational Intelligence*, vol. 807, no. 9, pp. 81–110, 2019, ISSN: 1860949X. DOI: 10.1007/978-3-030-04663-7_4.

[84] Y. Yuan, W. Su, and M. Zhu, Threshold-Free Measures for Assessing the Performance of Medical Screening Tests, *Frontiers in Public Health*, vol. 3, no. April, 2015, ISSN: 2296-2565. DOI: 10.3389/fpubh.2015.00057.

[85] V. Selvaraju, P. A. Karthick, and R. Swaminathan, Analysis of frequency bands of uterine electromyography signals for the detection of preterm birth, *Public Health and Informatics: Proceedings of MIE 2021*, vol. 0, pp. 283–287, 2021. DOI: 10.3233/SHTI210165.

[86] U. Goldsztejn and A. Nehorai, Predicting preterm births from electrohysterogram recordings via deep learning, pp. 1–18, 2022.

[87] S. Mohammadi Far, M. Beiramvand, M. Shahbakhti, and P. Augustyniak, Prediction of Preterm Delivery from Unbalanced EHG Database, *Sensors (Basel, Switzerland)*, vol. 22, no. 4, pp. 1–14, 2022, ISSN: 14248220. DOI: 10.3390/s22041507.

[88] M. Shahbakhti, M. Beiramvand, M. R. Bavi, and S. Mohammadi Far, A New Efficient Algorithm for Prediction of Preterm Labor, *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 4669–4672, 2019, ISSN: 1557170X. DOI: 10.1109/EMBC.2019.8857837.

[89] J. Mas-Cabo *et al.*, Robust Characterization of the Uterine Myoelectrical Activity in Different Obstetric Scenarios, *Entropy*, vol. 22, no. 7, p. 743, 2020, ISSN: 10994300. DOI: 10.3390/e22070743.

[90] 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, e0132116, 2015, ISSN: 19326203. DOI: 10.1371/journal.pone.0132116.

[91] M. Shahrdad and M. C. Amirani, Detection of preterm labor by partitioning and clustering the EHG signal, *Biomedical Signal Processing and Control*, vol. 45, pp. 109–116, 2018, ISSN: 17468108. DOI: 10.1016/j.bspc.2018.05.044.

[92] 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, e77154, 2013, ISSN: 19326203. DOI: 10.1371/journal.pone.0077154.

[93] 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, ISSN: 10994300. DOI: 10.3390/e19010002.

[94] J. Peng *et al.*, Evaluation of electrohysterogram measured from different gestational weeks for recognizing preterm delivery: a preliminary study using random Forest, *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 352–362, 2020, ISSN: 02085216. DOI: 10.1016/j.bbe.2019.12.003.

[95] J. Xu, Z. Chen, J. Zhang, Y. Lu, X. Yang, and A. Pumir, Realistic preterm prediction based on optimized synthetic sampling of EHG signal, *Computers in Biology and Medicine*, vol. 136, no. April, p. 104 644, 2021, ISSN: 18790534. DOI: 10.1016/j.compbiomed.2021.104644.

[96] M. U. Khan, S. Aziz, S. Ibraheem, A. Butt, and H. Shahid, Characterization of Term and Preterm Deliveries using Electrohysterograms Signatures, *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2019*, pp. 899–905, 2019. DOI: 10.1109/IEMCON.2019.8936292.

[97] J. Mas-Cabo, G. Prats-Boluda, J. Garcia-Casado, J. Alberola-Rubio, A. Perales, and Y. Ye-Lin, Design and Assessment of a Robust and Generalizable ANN-Based Classifier for the Prediction of Premature Birth by means of Multichannel Electrohysterographic Records, *Journal of Sensors*, vol. 2019, pp. 1–13, 2019, ISSN: 16877268. DOI: 10.1155/2019/5373810.

[98] J. Xu *et al.*, Network Theory Based EHG Signal Analysis and its Application in Preterm Prediction, *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 7, pp. 2876–2887, 2022, ISSN: 21682208. DOI: 10.1109/JBHI.2022.3140427.

[99] H. Lou, H. Liu, Z. Chen, Z. Zhen, B. Dong, and J. Xu, Bio-process inspired characterization of pregnancy evolution using entropy and its application in preterm birth detection, *Biomedical Signal Processing and Control*, vol. 75, no. January, p. 103 587, 2022, ISSN: 17468108. DOI: 10.1016/j.bspc.2022.103587.

[100] L. Chen, Y. Hao, and X. Hu, Detection of preterm birth in electrohysterogram signals based on wavelet transform and stacked sparse autoencoder, *PLoS ONE*, vol. 14, no. 4, pp. 1–16, 2019, ISSN: 19326203. DOI: 10.1371/journal.pone.0214712.

[101] M. Mischi *et al.*, Dedicated Entropy Measures for Early Assessment of Pregnancy Progression From Single-Channel Electrohysterography, *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 4, pp. 875–884, 2018, ISSN: 0018-9294. DOI: 10.1109/TBME.2017.2723933.

[102] A. Cheng *et al.*, Novel Multichannel Entropy Features and Machine Learning for Early Assessment of Pregnancy Progression Using Electrohysterography, *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 12, pp. 3728–3738, 2022, ISSN: 15582531. DOI: 10.1109/TBME.2022.3176668.

[103] H. Romero-Morales, J. N. Muñoz-Montes de Oca, R. Mora-Martínez, Y. Mina-Paz, and J. J. Reyes-Lagos, Enhancing classification of preterm-term birth using continuous wavelet transform and entropy-based methods of electrohysterogram signals, *Frontiers in Endocrinology*, vol. 13, no. January, pp. 1–11, 2023, ISSN: 16642392. DOI: 10.3389/fendo.2022.1035615.

[104] L. Chen and H. Xu, Deep neural network for semi-automatic classification of term and preterm uterine recordings, *Artificial Intelligence in Medicine*, vol. 105, no. April 2019, p. 101 861, 2020, ISSN: 18732860. DOI: 10.1016/j.artmed.2020.101861.

Chapter 2

Scope of the Study

Preterm deliveries, i.e. those occurring before 37 weeks of gestation, are a worldwide problem with a prevalence of more than 10% of all deliveries [1]. Being born too early increases the risk of neurodevelopmental disorders and respiratory and gastrointestinal complications [2]. Preterm births also have a great economic impact, with an average cost of between 5 to 10 times higher than that of a term birth [3].

Predicting preterm delivery has traditionally been based on measuring uterine dynamics by TOCO, cervical length [4] and/or the use of biochemical markers [5, 6]. However, none of these techniques have been shown to precisely predict preterm labor due to a lack of sensitivity, giving rise to overdiagnosis in up to 40% of women with preterm labor symptoms [7]. Electrohysterography (EHG) has emerged as a promising alternative that provides relevant information on the uterine electrophysiological state for predicting preterm labor. Numerous studies have proposed temporal, spectral or nonlinear parameters for EHG signal characterization throughout pregnancy. However, no specific study has focused on the analysis of complementary, redundant and irrelevant information in the different parameters, so that including irrelevant or redundant information could even generate noise in the model and worsen its performance [8].

Several preterm delivery prediction systems have been developed that use temporal, spectral and non-linear EHG parameters as input features, obtaining promising results with an AUC higher than 90-95%, although with no significant impact on clinical practice. This is partially due to the methodology bias used to mitigate the imbalanced data problem (preterm 12% vs. 88% at term) when designing preterm labor prediction systems. As mentioned above, SMOTE oversampling-splitting is the most common approach to deal with imbalanced data. Applying the SMOTE technique to the total database before data partition could generate a data structure correlation between training, validation and test data and overestimate the model's predictive ability [9]. It still remains unclear whether other strategies for dealing with the data-imbalance problem, such as oversampling and hybrid resampling techniques applied after data partition combined with classification methods and/or ensemble classifiers would improve model performance and/or reliability. [10, 11, 12].

2.1 Hypothesis

In this study, we tested the following hypotheses:

- EHG recordings on the abdominal surface contain relevant information on the electrophysiological state of the uterus, which is linked to the labor time horizon. This information is not available in the purely mechanical and indirect information obtained in traditional TOCO recordings.
- Robust and generalizable preterm birth prediction systems can be achieved through applying artificial intelligence approaches using EHGs from single gestation women who undergo regular check-ups. This would help clinicians to improve pregnancy planning and management by optimizing both maternal-fetal well-being and hospital resources.
- Different strategies of dealing with imbalanced data learning could lead to biased performance of the machine-learning prediction system with inconclusive results and a low generalization ability.

2.2 General objective

This project aimed to provide artificial intelligence computer-assisted tools, with special emphasis on imbalanced data learning, to predict preterm labor risks to help clinicians in their decisions on the management and planning of pregnancy and childbirth in a preterm labor maternal-fetal risk scenario. The project was aimed to give clinical professionals the appropriate tools to predict preterm labor and facilitate the optimal and earlier selection of possible treatments and management of hospital resources to improve maternal-fetal well-being and reduce costs.

2.3 Specific objectives

To achieve the general objective, the following specific objectives were pursued:

- To extract relevant features from EHG signals to discriminate the preterm versus term labor in women undergoing regular prenatal check-ups.
- To determine the complementary, redundant and noisy information of EHG features to optimize the feature subspace for predicting preterm delivery.
- To assess different imbalanced data learning strategies to achieve a robust and generalizable preterm birth prediction system.

2.4 References

[1] C. Leung, Born too soon, *Neuroendocrinology Letters*, vol. 25, no. SUPPL. 1, J. L. CP Howson, MV Kinney, Ed., pp. 133–136, 2004, ISSN: 0172780X. DOI: 10.2307/3965140.

[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, ISSN: 01406736. DOI: 10.1016/S0140-6736(08)60074-4.

[3] S. Petrou, H. H. Yiu, and J. Kwon, Economic consequences of preterm birth: A systematic review of the recent literature (2009-2017), *Archives of Disease in Childhood*, vol. 104, no. 5, pp. 456–465, 2019, ISSN: 14682044. DOI: 10.1136/archdischild-2018-315778.

[4] R. E. Garfield and W. L. Maner, Physiology and electrical activity of uterine contractions, *Seminars in Cell and Developmental Biology*, vol. 18, no. 3, pp. 289–295, 2007, ISSN: 10849521. DOI: 10.1016/j.semcdb.2007.05.004.

[5] V. Berghella, E. Hayes, J. Visintine, and J. K. Baxter, Fetal fibronectin testing for reducing the risk of preterm birth, *Cochrane Database of Systematic Reviews*, no. 4, 2008, ISSN: 1469493X. DOI: 10.1002/14651858.CD006843.pub2.

[6] M. Pandey, M. Chauhan, and S. Awasthi, Interplay of cytokines in preterm birth, *Indian Journal of Medical Research*, vol. 146, no. September, pp. 316–327, 2017, ISSN: 09715916. DOI: 10.4103/ijmr.IJMR_1624_14.

[7] J. D. Iams, F. F. Johnson, and M. Parker, A prospective evaluation of the signs and symptoms of preterm labor, *Obstetrics and Gynecology*, vol. 84, no. 2, pp. 227–230, 1994, ISSN: 1873233X.

[8] G. Doquire and M. Verleysen, A comparison of multivariate mutual information estimators for feature selection, in ICPRAM 2012 - Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods, vol. 1, 2012, pp. 176–185, ISBN: 9789898425980. DOI: 10.5220/0003726101760185.

[9] G. Vandewiele *et al.*, Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling, *Artificial Intelligence in Medicine*, vol. 111, p. 101 987, 2021, ISSN: 18732860. DOI: 10.1016/j.artmed.2020.101987. arXiv: 2001.06296.

[10] A. Gosain and S. Sardana, Handling class imbalance problem using oversampling techniques: A review, *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*, vol. 2017-Janua, pp. 79–85, 2017. DOI: 10.1109/ICACCI.2017.8125820.

[11] S. A. P., K. Subramaniam, and N. V. Iqbal, A review of significant researches on prediction of preterm birth using uterine electromyogram signal, *Future Generation Computer Systems*, vol. 98, pp. 135–143, 2019, ISSN: 0167739X. DOI: 10.1016/j.future.2018.10.033.

[12] T. Włodarczyk *et al.*, Machine learning methods for preterm birth prediction: A review, *Electronics (Switzerland)*, vol. 10, no. 5, pp. 1–24, 2021, ISSN: 20799292. DOI: 10.3390/electronics10050586.

Chapter 3

Optimized Feature Subset Selection Using Genetic Algorithm for Preterm Labor Prediction Based on Electrohysterography

Félix Nieto-del-Amor[1], Gema Prats-Boluda[1], Jose Luis Martinez-De-Juan[1], Alba Diaz-Martinez[1], Rogelio Monfort-Ortiz[2], Vicente Jose Diago-Almela[2] and Yiyao Ye-Lin[1,*]

[1] Centro de Investigación e Innovación en Bioingeniería (CI2B), Universitat Politècnica de València (UPV), Camino de la Vera s/n Ed. 8B, 46022 Valencia, Spain; feniede@ci2b.upv.es; gprats@ci2b.upv.es; jlmartinez@ci2b.upv.es; adiaz@ci2b.upv.es

[2] Servicio de Obstetricia, H.U.P. La Fe, 46026 Valencia, Spain; monfort_isaort@gva.es; diago_vicalm@gva.es

[*] Correspondence: yiye@ci2b.upv.es

Abstract

Electrohysterography (EHG) has emerged as an alternative technique to predict preterm labor, which still remains a challenge for the scientific-technical community. Based on EHG parameters, complex classification algorithms involving non-linear transformation of the input features, which clinicians found difficult to interpret, were generally used to predict preterm labor. We proposed to use genetic algorithm to identify the optimum feature subset to predict preterm labor using simple classification algorithms. A total of 203 parameters from 326 multichannel EHG recordings and obstetric data were used as input features. We designed and validated 3 base classifiers based on k-nearest neighbors, linear discriminant analysis and logistic regression, achieving F1-score of 84.63 ±2.76%, 89.34 ±3.5% and 86.87 ±4.53%, respectively, for incoming new data. The results reveal that temporal, spectral and non-linear EHG parameters computed in different bandwidths from multichannel recordings provide complementary information on preterm labor prediction. We also developed an ensemble classifier that not only outperformed base classifiers but also reduced their variability, achieving an F1-score of 92.04 ±2.97%, which is comparable with those obtained using complex classifiers. Our results suggest the feasibility of developing a preterm labor prediction system with high generalization capacity using simple easy-to-interpret classification algorithms to assist in transferring the EHG technique to clinical practice.

Keywords: Preterm labor; Electrohysterography; Myoelectric activity; Genetic algorithm; Ensemble learning

3.1 Introduction

Premature delivery is defined as one that occurs before 37 weeks of gestation. Over 9–12% of children are born prematurely every year, this being the leading cause of new-born deaths and the second-leading cause of death after pneumonia in children under the age of 5 [1]. In the case of survivors, it is associated with 20% mental retardation, 50% cerebral palsy and 33% eye injuries [2]. Preterm births are also associated with long-term morbidity consequences such as learning disabilities, attention deficit disorder, emotional problems, respiratory distress and intraventricular hemorrhage [1]. The costs derived from premature pregnancy are significant for national healthcare systems. In the United States, the economic cost in 2005 (combined medical, educational and lost productivity) associated with preterm birth amounted to at least $26.2 billion [1]. The average first-year medical costs, including both inpatient and outpatient care, were about 10 times greater for preterm ($32,325) than for term infants ($3325) [1].

Obstetricians usually assess uterine dynamics by tocodynamometer (TOCO) and cervix status using cervical length and bishop scores to determine the risk of preterm labor [3]. Biochemical markers such as fetal fibronectine and interleukin-6 have also been shown to be useful to identify patients that are not at risk of preterm labor, thus obtaining a high negative predictive value [4, 5]. However, all these techniques fail to detect women who will deliver prematurely, with positive predictive values lower than 0.50. Electrohysterography (EHG), which consists of the recording of uterine myoelectrical activity on the abdominal surface, has emerged as a powerful tool to predict preterm labor due to its high sensitivity in identifying the real preterm labor patients [6]. In addition to identifying uterine contractions, which is the only useful information that can be derived from TOCO, relevant information on the uterine electrophysiological state can also be obtained from the EHG. Temporal, spectral and non-linear parameters are used to characterize the electrophysiological changes throughout pregnancy [7]. In this context, the EHG signal amplitude associated with the number of uterine cells involved in one contraction has been shown to increase as pregnancy progresses. As labor approaches, a shift of spectral content towards higher frequencies has also been reported, suggesting increased cell excitability [6]. Different entropy measurements such as sample entropy (SampEn), fuzzy entropy (FuzEn) and spectral entropy (SpEn) have shown that signal predictability increases as labor approaches [8], although some controversial results have been reported [7]. Likewise, signal complexity seems to decrease, which was shown by analyzing the Lempel-Ziv evolution in function of time-to-delivery [9]. Poincaré plot-derived parameters [10] have also been proposed to characterize the EHG signal, and it has been observed that signal randomness decreases as labor approaches [6].

Many efforts have focused on developing prediction models for forecasting preterm labor based on EHG features and achieved classifier accuracy of more than 95% [7]. Despite the promising results of these prediction systems, they have had no significant impact on clinical practice [11]. This is due to various factors. Firstly, most of these systems use neural networks or support vector machine, multilayer perceptron, or similar algorithms, which involved non-linear transformations of the input EHG features into high dimension space, in which data from the target classes offer better linear separability [12]. This could give rise to good prediction performance even when the input features apparently do not contain information to differentiate the target classes. Obstetricians often consider this type of classification algorithm as a black box or a mathematician's gadget due to its being difficult to interpret [13], and so find it difficult to trust the predictions of these complex classifiers. By contrast, obstetricians are familiar with linear discriminant analysis (LDA), logistic regression (LR) and k-nearest neighbors (KNN) [12], which are simple and easy to interpret [13]. In addition, since the nonlinear transformation of the data is avoided, the definition of input EHG features used to obtain the prediction model also contributes to a better understanding of the uterine electrophysiological mechanism associated with labor. It is therefore fundamental to develop preterm labor prediction systems using simple and easily interpretable algorithms to improve the transferability of the EHG technique to clinical practice by gaining obstetricians confidence in prediction model outcomes [14]. Secondly, due to using reduced sample sizes, many previous studies have used crossvalidation methods to design and validate the classifiers, without determining the real generalization capacity for incoming data "never" seen by the classifiers [7].

The aim of this work was therefore to develop easily interpreted prediction systems based on EHG features for forecasting preterm labor in women at regular check-ups and to determine its generalization capacity for incoming data "never" seen by the classifiers, to facilitate the transfer of this technique to clinical practice. We also attempted to identify those EHG features that presented relevant and complementary information on labor prediction.

3.2 Materials and Methods

3.2.1 Database

Two public EHG databases available in Physionet were used for the study: The "Term-Preterm EHG Database" (TPEHG DB) [14] and the "The Term-Preterm EHG Dataset with tocogram" (TPEHGT DS) [15]. Both databases were obtained by the Department of Obstetrics and Gynecology at the Ljubljana University Medical Center.

To agree with the data available for each patient, only EHG recordings from both databases were used to predict preterm labor, i.e. the tocogram signals included in Term-Preterm EHG DataSet with Tocogram (TPEHGT DS) were ignored. They comprised a total of 326 EHG signals from 275 term labor (labor > 37 weeks) and 51 preterm labor were recorded during routine checkups of pregnant women between 22 and 37 weeks of gestation. The protocol used to obtain the EHG recordings consisted of placing four disposable electrodes on the woman's abdomen at an interelectrode distance of 7 cm. Three bipolar channels (S1, S2 and S3) were obtained from the monopolar EHG recordings, as shown in Figure 3.1. Each signal was digitized at 20 samples per second per channel with a 16-bit resolution over a range of ±2.5 millivolts [14]. A demographic description of both databases was provided in Table 3.1.

Table 3.1: Demographic description of the both databases (TPEHG DB and TPEHGT DS).

Group	N	Maternal Age (Years)	Parity	Abortions	Maternal Weight (kg)	WOG (Weeks)	Birth (Weeks)
Term	275	29.33 ±4.34	0.40 ±0.74	0.23 ±0.61	68.55 ±10.55	26.95 ±4.19	39.21 ±1.12
Preterm	51	29.08 ±5.26	0.41 ±0.64	0.26 ±0.60	66.82 ±11.24	27.59 ±3.72	33.92 ±2.21

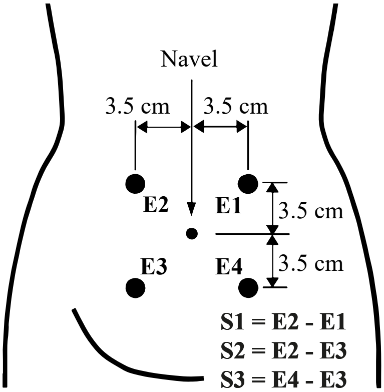


Figure 3.1: Recording protocol of EHG signals. Modified from [15].

3.2.2 EHG Signal Characterization

Physiologically, various types of uterine contractions such as Alvarez waves, Braxton Hicks contractions, preterm contractions, and with less frequency the so-called "long duration low-frequency band waves" [16] may be present in EHG recordings acquired from pregnant woman in the third trimester of gestation not close to delivery. The amplitude of the EHG bursts associated with these uterine contractions is expected to be very low, giving rise to subtle changes from basal activity. It is therefore very difficult to accurately identify the onset and offset of uterine contractions in these recordings and this could generate some uncertainty in the results derived from them. Previous results have revealed that whole window analysis can also be used for characterizing EHG signals [8, 14, 17] and that it even outperforms EHG-burst analysis for predicting imminent labor in women with threatened preterm labor [18]. Whole window analysis has the additional advantage of not requiring the uterine contractions to be identified in the EHG recordings, and only non-physiological segments should be excluded (such as artifacted segments and those with respiratory interference), thus facilitating its implementation in real time. In this work we therefore performed whole EHG window analysis to characterize the EHG signals rather than EHG-burst analysis. Two experts identified physiological segments in the EHG recordings by a double-blind process. The EHG characteristics of the recordings were analyzed in 120 s windows with a 50% overlap, the window length being a trade-off between computational cost and preserving the representative segment of the recordings [18]. We then computed the median value of all the analyzed windows in the recording to obtain a single representative value of each EHG parameter per recording.

A total of 66 temporal, spectral and non-linear parameters were worked out per recording channel and session, see Table 3.2. Firstly, the EHG signal is known to contain Fast Wave Low (FWL) and Fast Wave High (FWH) components, which are associated with signal propagability and cell excitability, respectively, their energy being distributed in 0.2–0.34 Hz and 0.34–4 Hz, respectively [6]. Due to the relatively lower signal-to-noise interference above 1 Hz [6, 14], we considered both 0.34–4 Hz and 0.34–1 Hz to characterize the FWH component. Therefore, we calculated the peak-to-peak amplitude (App) to describe the signal amplitude of different EHG components in four bandwidths: whole EHG bandwidth 0.1–4 Hz; FWH bandwidth 0.34–4 Hz; 0.2–0.34 Hz and 0.34–1 Hz, in which the energy of the FWL and FWH components is mainly distributed respectively. Due to the increasing formation of gap junctions as pregnancy progresses, signal amplitude, which is associated with the number of uterine cells involved in one contraction, was shown to increase as labor approaches [6].

As labor gets nearer, the EHG signal spectral content shifted to higher frequencies, suggesting increased cell excitability [6]. Different spectral parameters have been proposed to quantify the signal spectral content distribution: mean frequency (App), dominant frequency (DF) computed in the range 0.2–1 Hz and in 0.34–1 Hz, power spectrum deciles (D1, . . . , D9), normalized energy (NormEn) (0.2–0.34 Hz, 0.34–0.6 Hz 0.6–1 Hz) [19], high-to-low frequency energy ratio (H/L Ratio) and spectral moment ratio (SpMR), as in [15]. We also included Teager energy, which contains information not only on signal amplitude, but also on the frequency content [8]. Due to the increased cell excitability, different spectral parameters are expected to increase as pregnancy progresses, however there is no agreement in the literature about the spectral parameters that can best characterize the EHG signal, and above all those that provide information to complement temporal and non-linear parameters.

Due to the non-linear nature of the biological process dynamics, non-linear parameters have been widely used to characterize EHG signals. A previous study showed that the bandwidth in which the non-linear parameters are computed is a key factor in obtaining a robust and physically interpretable indicator for characterizing EHG signals [14, 17]. We therefore computed several non-linear parameters in the same four bandwidths as App to determine whether there was any redundant or complementary information between non-linear parameters computed in different bandwidths and to other linear features. SampEn, FuzEn and SpEn were used to measure time series regularity and predictability in both the time and frequency domains [8, 14]. A lower value of entropy metrics is associated with more self-similarity in the regular and predictable time series. We also computed the Lempel-Ziv index (binary (LZBin) and multistate n = 6 (LZMulti)), which evaluates time series complexity by measuring the "diversity" of the patterns embedded in a time series [8, 18]. Time reversibility (TimeRev) estimates the similarity in forward (natural) and reverse time and can be considered as a measurement of the degree of signal nonlinearity [8]. Uterine myoelectric activity has also been shown to possess fractal properties and is another way of measuring signal self-similarity [20, 21]. We also computed Katz's fractal dimension (KFD) [22] since it is less sensitive to noise than Higuchi's method [23]. It was defined as the ratio between the curve length, which corresponds the sum of the Euclidean distances between successive points of the time series, and the maximum distance between the first point and any sample of the time series [22].

Since the "present" EHG signal amplitude may significantly influence the "following" values, we represented the Poincaré plot of consecutive EHG signal amplitudes (EHG[n] vs. EHG[n-1]) to estimate the short (SD1) and long-term (SD2) variation of the dispersion along the minor and major axes of the ellipse, respectively [10]. We then obtained the SDRR, defined as square root of the variance of the whole time series $\sqrt{(\mathrm{SD1}^2 + \mathrm{SD2}^2)/2}$ and SD1/SD2 ratio, which measures signal randomness. This latter has been shown to decrease significantly in women with threatened preterm labor who delivered in less than 7 days, in comparison to those who delivered in more than 7 days [8]. Table 3.2 summarizes the EHG parameters and obstetric data used to design the preterm labor prediction model (3 channels × 66 EHG parameters per channel + 5 obstetric data = 203 features).

3.2.3 Classifier Design and Evaluation

Since the preterm birth rate is about 12% in women who have regular check-ups, this means that the two target classes are highly imbalanced. It is well known that the conventional classification algorithms are often biased towards the majority class for imbalanced data, obtaining a higher misclassification rate for the minority class instances [24]. In this case, low sensitivity can be expected for true preterm labor. In this work, we used the synthetic minority oversampling technique (SMOTE, k = 5) [24], which has been widely used to mitigate imbalanced class problems, to obtain balanced preterm labor and term labor data [7].

Table 3.2: Input features for predicting preterm labor that include both temporal, spectral, and non-linear.

EHG Temporal Parameters	**EHG Spectral Parameters**	**EHG Non-Linear Parameters**	**Obstetric data**
App	MeanF DF NormEn H/L Ratio [D1–D9] SpMR Teager Energy	SampEn FuzEn SpEn LZBin LZMulti (n = 6) SD1 SD2 SDRR SD1/SD2 TimeRev KFD	Maternal age Parity Abortions Weight Week of gestation (WOG)

We then used the conventional holdout method (30 partitions) to design and validate the classifiers. For each partition we randomly split the whole balanced database into 3 datasets with the same proportion between the classes: training (1/3), validation (1/3) and testing (1/3) for designing, validating and testing the classifier, respectively.

As mentioned above, a total of 203 EHG features derived from the 3-channels EHG recording and obstetric data was used to design the prediction system. Since they may contain mutual and redundant information or noise, which could lead to loss of prediction performance, it is fundamental to reduce the dimension of the data. The relevance of the features for predicting preterm labor can be evaluated either individually (unidimensional approaches) or multidimensionally. Unidimensional approaches are simple and fast and therefore appealing. Nevertheless, if we only consider a unidimensional approach, the outcomes suggest eliminating those with non-significant statistical differences. The individual discrimination power must not be the only consideration since possible correlations and dependencies between the features are not considered. Many authors [12, 25, 26, 27, 28] claim that the redundant information shared between the characteristics leads to discarding some of them with a feature selection algorithm. The same occurs with noisy features that add an artifact to the classification. Although complementarity between features is critical to achieve good performance, multidimensional search techniques such as mutual information estimation may be helpful to evaluate possible correlations and dependencies between features. Nevertheless, estimating the mutual information (especially through probability density function estimations) between high-dimensional variables is a hard task in practice due to the limited number of available data points for real-world problems [29]. In this work, we measured the capability of each individual feature to identify a premature delivery with a statistical test. The Wilcoxon Rank-Sum Test was performed to compare the features' ability to distinguish term and preterm deliveries from EHG recordings in routine check-ups. This is a non-parametric statistical hypothesis test used to compare two related samples to assess whether their population mean ranks differ ($\alpha < 0.05$) [30].

Lower p-values mean higher discriminatory capacity between target classes for the individual feature. We performed multidimensional analysis by using a wrapper method for feature selection which has been widely used in the literature [25, 27, 31] and has been proven to provide better results than filter methods based on the intrinsic information embedded in the features: the genetic algorithm [31]. The initial feature set was assessed by a genetic algorithm to fit with logistic regression (LR), linear discriminant analysis (LDA) and k-nearest neighbor (KNN) classifiers. If a feature was selected by one or several classifiers, it meant that it complemented others in predicting preterm delivery. The genetic algorithm is a random search strategy based on procedures of natural evolution and is widely used for selecting the optimal feature subset to design computer-aided systems for pattern recognition in different biomedical applications [32, 33]. In this work, population size (N) and genome length were fixed to the number of model features, N = 203 [26]. The crossover function implemented was the arithmetic crossover with a probability of 0.8. Typically, it was assumed between 0.6 and 1, increasing the randomness of the children generation the lower the value it takes [34]. We used mutation uniform for the mutation function, with a probability of 0.01, since the convergence to a lower minimum is better with low values (<0.1) [31, 34]. The tournament function with a size of 2 and an elite count of 2 was used to select the next generation population [26]. The termination condition of the genetic algorithm was defined as a differential tolerance of 10^{-6} for the fitness function between 150 consecutive generation's best fitness value.

Figure 3.2 summarizes the procedure carried out to fit the model and obtain the optimized feature subset. Firstly, an initial population was randomly established, then the balanced data set was masked with the chromosomes, obtaining the different feature subsets. The mask (selected features), which corresponded to an i-chromosome, was set to the balanced data set obtaining the i-subset, with $1 \leq \mathrm{i} \leq \mathrm{N}$. Subsequently, for each feature subset, the training dataset was used to fit the prediction model for each partition. The average F1-score over 30 partitions in validation subsets was then worked out to assess the model's performance. When all the chromosomes in the population were evaluated, a new one was created, crossing over and mutating the last population and keeping the chromosomes that outperformed the fitness function (3.1). It became an iterative procedure until the termination condition was reached, giving rise to the optimized feature subset which corresponded to the best chromosome that optimized the fitness function. Finally, we determined the model's generalization capacity for the testing dataset which could be considered as the incoming new data "never" seen by the model.

$$\text{Fitness function} = \max\{\overline{\text{F1-score}} \cdot (\text{NFeat} - \text{NCFeat})\} \tag{3.1}$$

where $\overline{\text{F1-score}}$ is the average F1-score of validation dataset for 30 partitions, NFeat is the number of features of the initial set and NCFeat is the number of features of the current subset.

For the classification methods, we compared different methods (LDA, LR and KNN) that can easily be interpreted by obstetricians. Due to the fact that the optimization cost function of the genetic algorithm is the weighted classifier performance (3.1) [26], different optimized EHG feature subsets may be obtained for each classification method.

To compare the different prediction model's performance the following metrics were used: accuracy, F1-score, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and area under curve (AUC) [12]. The Friedman non-parametric test was conducted to determine the statistical difference in different metrics [35]. The Nemenyi post-hoc test [35] was then used for pair-wise evaluation of the classifiers, checking the similarity of their performance.

Since individual classifiers may present a high bias or variance, giving rise to weak base models, to overcome these problems we also evaluated the utility of ensemble classifier to achieve better performance. The three previously obtained base models based on LDA, LR and KNN were used, with a commonly used majority voting strategy in the meta-level to obtain a more robust meta-classifier preterm labor predictor. We also determined the model's generalization capacity for testing incoming data "unseen" by the model for the ensemble classifier. Base and ensemble classifiers performance variability were compared by computing the coefficient of variation of the classifier metrics for both the validation and testing dataset.

3.3 Results

Figure 3.3 shows the three optimized feature subsets for predicting preterm labor using LDA, LR and KNN as classification methods. It also shows the outcomes of individual correlations of the Wilcoxon Rank-Sum test for each individual feature to discriminate between preterm and term labor classes. Firstly, we found that the EHG features computed from the 3 channels contained complementary information: 22, 25 and 23 features computed from S1, S2 and S3 respectively were included in at least one of the three optimized feature subsets. Peak-to-peak amplitude computed in the 0.2–0.34 Hz bandwidth, where the main energy of the FWL component is concentrated, seems to provide relevant information for predicting preterm labor, having been included in the "best chromosomes" of the LR and LDA classifiers. In general, there is important mutual information between different spectral EHG parameters. In this regard, deciles D1–D3, D8 and D9 were not included in the best chromosome of any of the three base classifiers. The DF in 0.1–1 Hz and 0.34–1 Hz, NormEn in 0.1–0.34 Hz and 0.34–0.6 Hz, D5, D6 and SpMR seem to provide most relevant information on cell excitability. As for the non-linear parameters, the information extracted from different bandwidths was not necessarily redundant, but rather complementary. The non-linear parameters, extracted from 0.34–4 Hz, where FWH components distributed their energy, seem to contain the most relevant information for predicting preterm labor. In contrast, the non-linear parameters computed from

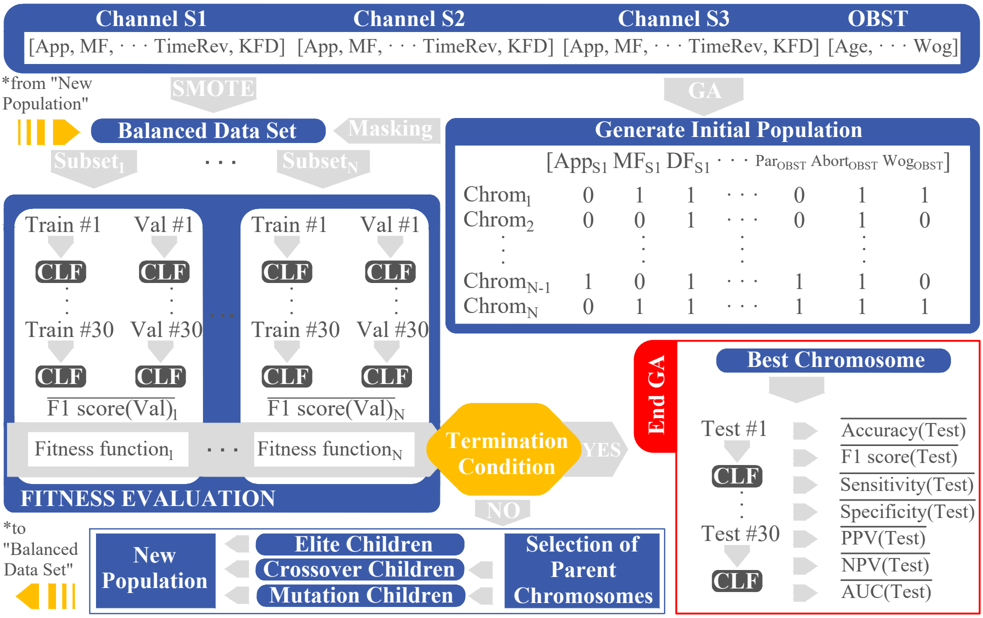
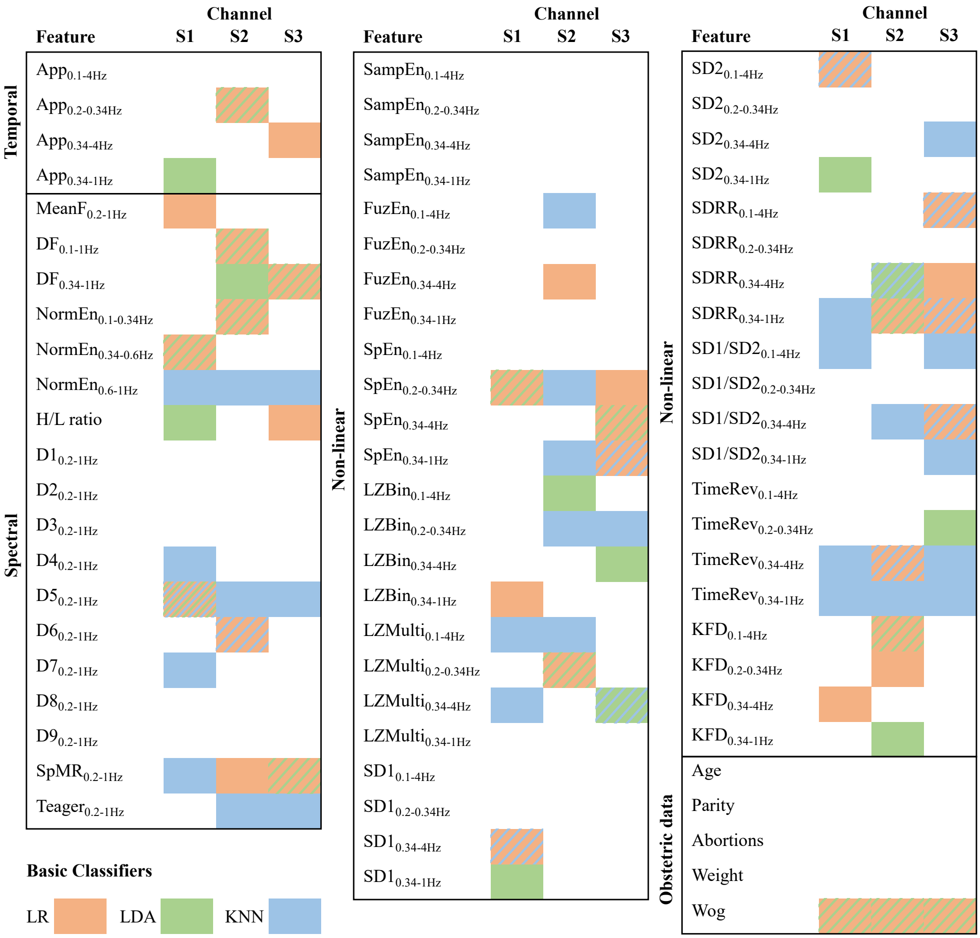


Figure 3.2: Genetic algorithm for selecting optimized feature subset to predict preterm labor based on EHG. Considering: genetic algorithm (GA), classifier (CLF), training partition (Train), validation partition (Val), testing partition (Test), chromosome (Chrom), population size (N), obstetrics features (OBST).

0.2–0.34 Hz contain less complementary information for differentiating preterm and term records. In this respect, the parameters derived from the Poincare plot, SampEn and FuzEn, computed in this bandwidth were not included in any of the three base classifiers. In comparison to LZBin, LZMulti seems to provide more complementary information to other EHG features and therefore more present in the best chromosome of the different algorithms (see Figure 3.3). The entropy metrics (SampEn, FuzEn and SpEn) contained redundant information, SpEn being the most relevant one forming part of the optimum feature subset of several classifiers. Both TimeRev and KFD also offered relevant and complementary information to other features. As for the common features shared between the different prediction models, only decile 5, which is equivalent to median frequency, was common for the best chromosome of the three base classifiers. It can be seen that a subset of 11 and 8 features was also shared by LR and LDA, and by LR and KNN classifiers, respectively. As for obstetrics data, only WOG was shown to be relevant for predicting preterm labor for both LDA and LR classifiers. As for individual statistical test, a total of 43 features with statistically significant differences were obtained. Analysing the results by features selected by classifiers, a total of 29, 22 and 40 characteristics for LR, LDA and KNN were chosen respectively. Only 17 of these features that obtained a p-value < 0.05. 21 features were selected by at least 2 classifiers and 8 of these obtained a p-value lower than 0.05. Not all those features which obtained a p-value < 0.05 (decile 1 to 3 or sample entropy) were included in the optimized feature subset. Various features which provided individual statistical significance between target classes were used in each optimized feature subset. Likewise, the optimized feature subsets also included some features that did not obtain individual statistical significance between target classes but provided complementary information to other relevant features, for example peak to peak amplitude, Teager, SD ratio or time reversibility.

Table 3.3 shows the average performance for both base and ensemble classifiers for training, validation and testing dataset and Figure 3.4 shows the results of the Nemenyi post-hoc test of F1-score between different classifiers for both validation and testing datasets. In general, the base classifiers performance for training dataset was better than the validation dataset, as expected. The classifier metrics of the testing dataset was similar to or slightly inferior than the validation dataset. Nevertheless, regardless of the classification method, the average F1-score was over 85% and 80% for validation and testing dataset, respectively. The LDA and LR classifiers showed similar performance and obtained no significant difference. Both LDA and LR classifiers obtained better performance than KNN, although a significant difference was only achieved for the LDA classifier. The ensemble classifier obtained a significantly better performance than the different base ones, achieving an average F1-score of around 92% for both the validation and testing datasets. The receiver operating characteristic (ROC) curves of the different classifiers for validation and testing partitions are shown in Figure 3.5.

As for classifier performance variability between partitions, training data usually obtained a lower value than the validation and testing datasets with similar results. In Figure 3.6 are depicted the different classifier metrics' variability for both validation and testing dataset are depicted. In general, base and ensemble classifiers presented low performance variability (<10%). Accuracy, F1-score and AUC metrics usually achieved the lowest variability, while sensitivity, specificity and PPV generally had the highest. It can also be seen that the ensemble classifier can considerably reduce the different metrics' variability, obtaining similar or lower values than that of the minimum variability achieved by the base classifiers.



		Channel		
	Feature	S1	S2	S3
Temporal	$\text{App}_{0.1\text{-}4\text{Hz}}$			
	$\text{App}_{0.2\text{-}0.34\text{Hz}}$		🟧🟩	
	$\text{App}_{0.34\text{-}4\text{Hz}}$			🟧
	$\text{App}_{0.34\text{-}1\text{Hz}}$	🟩		
Spectral	$\text{MeanF}_{0.2\text{-}1\text{Hz}}$	🟧		
	$\text{DF}_{0.1\text{-}1\text{Hz}}$		🟧🟩	
	$\text{DF}_{0.34\text{-}1\text{Hz}}$		🟩	🟧🟩
	$\text{NormEn}_{0.1\text{-}0.34\text{Hz}}$		🟧🟩	
	$\text{NormEn}_{0.34\text{-}0.6\text{Hz}}$	🟧🟩		
	$\text{NormEn}_{0.6\text{-}1\text{Hz}}$	🟦	🟦	🟦
	H/L ratio	🟩		🟧
	$\text{D1}_{0.2\text{-}1\text{Hz}}$			
	$\text{D2}_{0.2\text{-}1\text{Hz}}$			
	$\text{D3}_{0.2\text{-}1\text{Hz}}$			
	$\text{D4}_{0.2\text{-}1\text{Hz}}$	🟦		
	$\text{D5}_{0.2\text{-}1\text{Hz}}$	🟧🟩🟦	🟦	🟦
	$\text{D6}_{0.2\text{-}1\text{Hz}}$		🟧🟦	
	$\text{D7}_{0.2\text{-}1\text{Hz}}$	🟦		
	$\text{D8}_{0.2\text{-}1\text{Hz}}$			
	$\text{D9}_{0.2\text{-}1\text{Hz}}$			
	$\text{SpMR}_{0.2\text{-}1\text{Hz}}$	🟦	🟧	🟧🟩
	$\text{Teager}_{0.2\text{-}1\text{Hz}}$		🟦	🟦

Basic Classifiers

LR 🟧 LDA 🟩 KNN 🟦

		Channel		
	Feature	S1	S2	S3
Non-linear	$\text{SampEn}_{0.1\text{-}4\text{Hz}}$			
	$\text{SampEn}_{0.2\text{-}0.34\text{Hz}}$			
	$\text{SampEn}_{0.34\text{-}4\text{Hz}}$			
	$\text{SampEn}_{0.34\text{-}1\text{Hz}}$			
	$\text{FuzEn}_{0.1\text{-}4\text{Hz}}$		🟦	
	$\text{FuzEn}_{0.2\text{-}0.34\text{Hz}}$			
	$\text{FuzEn}_{0.34\text{-}4\text{Hz}}$		🟧	
	$\text{FuzEn}_{0.34\text{-}1\text{Hz}}$			
	$\text{SpEn}_{0.1\text{-}4\text{Hz}}$			
	$\text{SpEn}_{0.2\text{-}0.34\text{Hz}}$	🟧🟩	🟦	🟧
	$\text{SpEn}_{0.34\text{-}4\text{Hz}}$			🟧🟩
	$\text{SpEn}_{0.34\text{-}1\text{Hz}}$		🟦	🟧🟦
	$\text{LZBin}_{0.1\text{-}4\text{Hz}}$		🟩	
	$\text{LZBin}_{0.2\text{-}0.34\text{Hz}}$		🟦	🟦
	$\text{LZBin}_{0.34\text{-}4\text{Hz}}$			🟩
	$\text{LZBin}_{0.34\text{-}1\text{Hz}}$	🟧		
	$\text{LZMulti}_{0.1\text{-}4\text{Hz}}$	🟦	🟦	
	$\text{LZMulti}_{0.2\text{-}0.34\text{Hz}}$		🟧🟩	
	$\text{LZMulti}_{0.34\text{-}4\text{Hz}}$	🟦		🟩🟦
	$\text{LZMulti}_{0.34\text{-}1\text{Hz}}$			
	$\text{SD1}_{0.1\text{-}4\text{Hz}}$			
	$\text{SD1}_{0.2\text{-}0.34\text{Hz}}$			
	$\text{SD1}_{0.34\text{-}4\text{Hz}}$	🟧🟦		
	$\text{SD1}_{0.34\text{-}1\text{Hz}}$	🟩		

		Channel		
	Feature	S1	S2	S3
Non-linear	$\text{SD2}_{0.1\text{-}4\text{Hz}}$	🟧🟦		
	$\text{SD2}_{0.2\text{-}0.34\text{Hz}}$			
	$\text{SD2}_{0.34\text{-}4\text{Hz}}$			🟦
	$\text{SD2}_{0.34\text{-}1\text{Hz}}$	🟩		
	$\text{SDRR}_{0.1\text{-}4\text{Hz}}$			🟧🟦
	$\text{SDRR}_{0.2\text{-}0.34\text{Hz}}$			
	$\text{SDRR}_{0.34\text{-}4\text{Hz}}$		🟩🟦	🟧
	$\text{SDRR}_{0.34\text{-}1\text{Hz}}$	🟦	🟧🟩	🟧🟦
	$\text{SD1/SD2}_{0.1\text{-}4\text{Hz}}$	🟦		🟦
	$\text{SD1/SD2}_{0.2\text{-}0.34\text{Hz}}$			
	$\text{SD1/SD2}_{0.34\text{-}4\text{Hz}}$		🟦	🟧🟦
	$\text{SD1/SD2}_{0.34\text{-}1\text{Hz}}$			🟦
	$\text{TimeRev}_{0.1\text{-}4\text{Hz}}$			
	$\text{TimeRev}_{0.2\text{-}0.34\text{Hz}}$			🟩
	$\text{TimeRev}_{0.34\text{-}4\text{Hz}}$	🟦	🟧🟦	🟦
	$\text{TimeRev}_{0.34\text{-}1\text{Hz}}$	🟦	🟦	🟦
	$\text{KFD}_{0.1\text{-}4\text{Hz}}$		🟧🟩	
	$\text{KFD}_{0.2\text{-}0.34\text{Hz}}$		🟧	
	$\text{KFD}_{0.34\text{-}4\text{Hz}}$	🟧		
	$\text{KFD}_{0.34\text{-}1\text{Hz}}$		🟩	
Obstetric data	Age			
	Parity			
	Abortions			
	Weight			
	Wog	🟧🟩	🟧🟩	🟧🟩

Figure 3.3: Optimized feature subset (best chromosomes) achieved using genetic algorithm for each base classifiers implemented with LR (orange), LDA (green), and KNN (blue). A single frame color indicates that a feature is part of the classifier's best chromosome identified with that color; two and three frame colors indicate that the feature belongs to two and three classifiers' best chromosome, identified with the corresponding colours. *, ** and *** mean a p-value of the Wilcoxon Rank-Sum test lower than 0.05, 0.005 and 0.0005, respectively, computed by each feature between preterm and term labor classes.

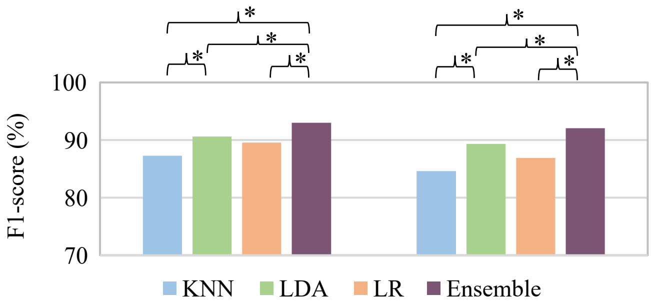


Figure 3.4: Nemenyi post-hoc test of F1-score between the different classifiers for both validation (left) and testing (right) dataset. * means significant statistical difference (p-value ≤ 0.05) between classifier performance.

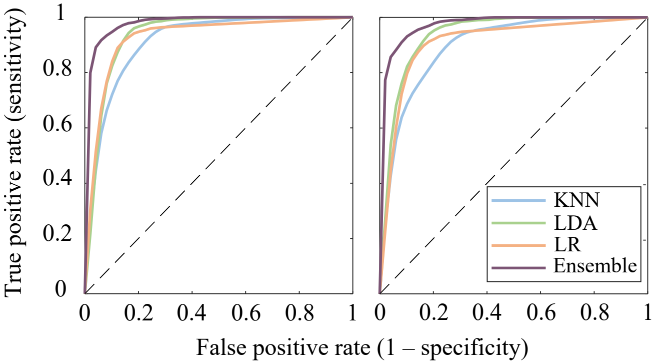


Figure 3.5: ROC curves of the different base and the ensemble classifiers using optimized feature subset for both validation (left) and testing dataset (right).

Table 3.3: Average performance of 30 partitions for base classifiers (LDA, LR and KNN) and for the classifiers' ensemble.

		KNN	**LDA**	**LR**	**Ensemble**
Accuracy (%)	Train	92.86 ±2.41	96.01 ±1.50	100 ±0.00	99.62 ±0.59
	Validation	85.82 ±3.82	90.03 ±2.53	89.78 ±3.29	92.61 ±2.48
	Test	82.92 ±2.90	88.77 ±3.89	87.42 ±40.00	91.64 ±3.2
F1-score (%)	Train	93.26 ±2.20	96.14 ±1.43	100.00 ±0.00	99.63 ±0.58
	Validation	87.26 ±3.21	90.61 ±2.23	89.55 ±3.49	92.97 ±2.27
	Test	84.63 ±2.76	89.34 ±3.50	86.87 ±4.53	92.04 ±2.97
Sensitivity (%)	Train	98.43 ±1.86	98.99 ±1.29	100.00 ±0.00	99.87 ±0.48
	Validation	96.48 ±2.88	95.79 ±3.04	87.86 ±5.07	97.36 ±2.25
	Test	94.21 ±5.00	93.58 ±3.63	84.03 ±6.73	96.23 ±3.17
Specificity (%)	Train	87.30 ±4.05	93.02 ±2.81	100.00 ±0.00	99.37 ±1.14
	Validation	75.16 ±7.40	84.28 ±5.48	91.7 ±4.22	87.86 ±4.34
	Test	71.64 ±4.58	83.96 ±6.59	90.82 ±3.67	87.04 ±5.47
PPV (%)	Train	88.67 ±3.28	93.48 ±2.49	100.00 ±0.00	99.39 ±1.11
	Validation	79.83 ±5.19	86.12 ±3.95	91.51 ±4.08	89.04 ±3.49
	Test	76.95 ±2.76	85.63 ±5.02	90.2 ±3.58	88.33 ±4.44
NPV (%)	Train	98.25 ±2.09	98.95 ±1.33	100.00 ±0.00	99.88 ±0.47
	Validation	95.64 ±3.32	95.41 ±3.18	88.50 ±4.28	97.13 ±2.4
	Test	92.92 ±5.42	92.99 ±3.78	85.33 ±5.28	95.94 ±3.35
AUC (%)	Train	98.49 ±0.83	99.30 ±0.56	100.00 ±0.00	100.00 ±0.00
	Validation	92.16 ±2.37	94.72 ±2.10	93.03 ±2.74	98.63 ±0.85
	Test	90.20 ±2.41	94.72 ±2.54	91.44 ±2.63	98.13 ±1.26

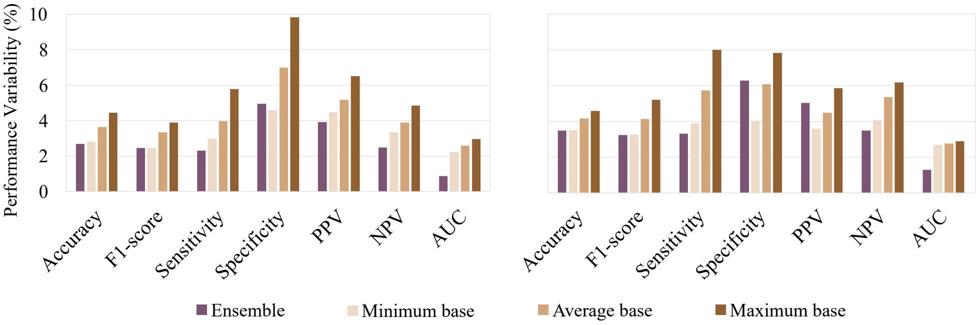


Figure 3.6: Base and ensemble classifiers performance variability for both validation (left) and testing (right) dataset. As for base classifiers, minimum, average and maximum variability were computed.

3.4 Discussion

Our aim in this work was to develop a preterm labor prediction system based on EHG using simple and easily interpretable classification algorithms in order to promote the transfer of the EHG technique to clinical practice. In this context, the features' quality, i.e. their capability to provide useful and complementary information to others, it is critical to achieve satisfactory classification performance. Although temporal, spectral and non-linear EHG parameters have been shown to provide relevant information for predicting preterm labor [7], the redundancy and complementarity of different EHG features were still unclear. The classical dimension reduction methods, such as principal component analysis (PCA), does not guarantee the extraction of complementary information or noise reduction to optimize classifier accuracy [36].

Unlike the filter methods for feature selection that reduce the number of features using the intrinsic properties of the data, regardless of the learning algorithm to be used, we proposed to use wrapper methods, which generally lead to better classification performance [27] to determine the optimized feature subset. Of the different search strategies, we preferred to use the random search strategy, which is a tradeoff between classification performance and search complexity for moderate and/or large numbers of features [27]. In this respect, both particle swarm optimization (PSO) and the genetic algorithm could be used to optimize data information in feature space. Benalcazar et al. proposed the use of PSO and the neural network to predict labor induction success [37]. Alamedine showed that PSO generally outperformed sequential forward selection and Jeffrey divergence distance for predicting labor and pregnancy contractions when using LDA, QDA and KNN as classification methods [25]. The genetic algorithm has also been shown to obtain better performance than the filter method to predict pregnancy and labor contractions using KNN [38], and also outperformed both forward and backward selection for predicting central nervous system embryonal tumor outcomes, based on gene expression [27]. This is due to the ability of the genetic algorithm to escape local optima and discover the global optimum in even a very rugged and complex fitness function [31]. In practice, the genetic algorithm may not always lead to a theoretically perfect solution to a problem, but always delivers at least a very good solution [31]. We here used the genetic algorithm to perform feature selection as it had been shown to theoretically outperform PSO in obtaining highest number of best minimum fitness and did so faster [39]. We believe that in our context, using the PSO strategy would give rise to similar or slightly worse results.

Using the genetic algorithm for feature selection, we proved the feasibility of developing a preterm labor prediction system with high generalization capacity using simple and easy-to-interpret algorithms, achieving an F1-score of the individual base classifier of over 80% for incoming data previously unseen by the model. Since these simple classifier's success depends mainly on the information embedded in the feature, we believe that this will help to gain obstetricians confidence of preterm labor prediction model's outcome, bringing thus the EHG technique closer to clinical practice.

The prediction model proposed here has the inherent advantage over PCA that it does not require all the parameters to be computed once the model is trained, and is therefore easier to implement on a portable device due to its lower computational cost. In this respect, the time necessary to compute the optimized feature subsets of each window was 19.03 seconds using just one core of an Intel Core I7 8550U laptop, which can be considerably less than the maximum time limit of 60 s (step size between analysis windows). In addition, the total time consumption required for each base classifier implemented by LR, LDA and KNN to obtain its outcome from the input features applying a majority voting strategy to generate the ensemble classifier's outcome was 0.094 s, indicating that the prediction outcome can be obtained immediately after the recording. To characterize the EHG signal, the median value of the temporal, spectral and non-linear parameters was worked out in windows of 120 s as the representative value of the whole recording. Previous studies have shown that the 10-90$^{\text{th}}$ percentiles of individual EHG parameters can better discriminate between preterm and term records [17]. The prediction model's performance when using these percentiles of EHG features (results not shown for the sake of brevity) was similar to that obtained in this work with no significant differences.

We also reported for the first time three EHG feature subsets (best chromosomes) that contain the maximum complementary information, thus optimizing the prediction model's performance. Our results revealed the redundancy between different spectral parameters and also the complementarity between temporal, spectral and non-linear parameters. This result is understandable, since these represent the different phenomena involved in uterine contraction efficiency: intensity, excitability and non-linear dynamic character [6].

As for non-linear parameters, Fele-Žorž et al. showed that SampEn computed in FWH bandwidth can discriminate preterm and term records [14]. Lemancewicz et al. found that Lempel-Ziv and approximate entropy computed at 0.24–4 Hz was significantly higher for women with threatened preterm labor who delivered in less than seven days than in those who delivered in more than seven days [9]. On the other hand, Lempel-Ziv computed at FWH bandwidth in women with threatened preterm labor decreased as labor approached [18]. Mas-Cabo et al. compared different non-linear parameters computed from the whole bandwidth (0.1–4 Hz) and fast wave high bandwidth (0.34–4 Hz) and concluded that the signal bandwidth in which non-linear parameters are computed may be a key factor in obtaining a robust and physically interpretable indicator for characterizing EHG signals [18]. In this work, we found that the non-linear parameters computed in different bandwidths provided complementary information. In addition, SampEn, FuzEn and Lempel-Ziv which represent the signal predictability and complexity in time domain, contained redundant information between each other. SpEn, which measures the flatness of the spectrum, seemed to provide additional information on the signal. Of the obstetric data, only gestational age was found to be relevant for predicting preterm labor, possibly due to the values of EHG features being intrinsically modified as gestational age increases [14]. However, maternal age, parity and abortions were irrelevant for the algorithm, although these latter have been associated with preterm labor risk factors [40]. We believe that other obstetric data, such as fetal fibronectin and cervical length may provide complementary information to EHG and improve preterm labor prediction performance.

The results revealed the complementarity of EHG features extracted from different channels, highlighting the utility of multichannel recording for preterm labor prediction. Firstly, a similar number of features from the three channels were included in the optimized feature subsets (Best chromosomes). Using the same method (Genetic algorithm + LDA, LR or KNN classifier), we also attempted to develop a preterm labor prediction system using the information extracted from individual channels and computing a mean efficiency index, which has been shown to be a more robust indicator of uterine electrical activity efficiency from multichannel recordings [41]. We found that the model performance based on individual channels was much inferior (<70%) to that obtained for multichannel recording, suggesting that multichannel recording may provide a more reliable electrophysiological state of the whole uterus [41]. Unlike the prediction system based on a neural network developed in a previous work, we obtained a better performance using all the EHG features extracted from the three individual channels (S1+S2+S3) than the mean efficiency index [41]. We believe this discrepancy may be associated with the greater degree of freedom in combining information from a multichannel recording and the genetic algorithm's capacity to optimize the information in feature space. Our results for base classifiers also outperformed those obtained by Fergus, who used PCA to reduce the data dimension of EHG features extracted from channel 3 in the 0.34–1 Hz bandwidth. An AUC of 66%, 86% and 84% was obtained from this latter to validate the dataset using LDA, LR and KNN classifiers, respectively [42]. Our results are also comparable with those obtained using a neural network [7] and PCA for data dimension reduction (AUC = 88.2%). In other words, the information optimization in the feature space eliminated the need to use complex classification methods. In comparison with other studies in the literature [43, 44], our prediction model's performance may be slightly lower, which could be due to the different method used to validate it. In this respect, the use of cross-validation and hold-out validation without preserving a testing dataset previously unseen by the model may overestimate the model performance [12].

We also evaluated the performance improvement in an ensemble classifier over base classifiers in predicting preterm labor. Ensemble methods combining the output of individual weak classifiers have successfully produced accurate predictions for many complex classification tasks [12]. The success of these methods is attributed to their ability to consolidate accurate predictions and correct errors across many diverse base classifiers [45, 46]. Successful ensemble methods make a balance between the ensemble's diversity and accuracy [47]. In this work we showed that a simple ensemble classifier aggregation whose meta-level only consisted of a majority voting strategy can further improve classification performance, obtain higher average metrics and reduce the different metric variabilities between partitions. In this respect, the use of other meta-learning algorithms may further improve classification performance. Again, it is preferred to use simple and easy-to-interpret algorithms to develop the meta-classifier. Future work will compare different meta-learning algorithms in terms of improved classification performance and to further validate the utility of these methods for predicting imminent labor and/or preterm labor in women with threatened preterm labor undergoing tocolytic treatment. Regardless of the improvement in performance achieved by using ensemble classifiers, obstetricians may find it significantly more difficult to interpret these algorithms, and this should be taken into account when transferring the EHG technique to clinical practice.

In spite of its promising results, the present study is not exempt from limitations. Firstly, the size of the samples of women who took regular check-ups and delivered at term or prematurely was highly imbalanced. We used the commonly-used SMOTE oversampling technique to minimize this problem [24]. Specific imbalanced data learning algorithms, such as weighted classifiers or boosting ensemble learning, could help to achieve more reliable preterm labor prediction systems. Secondly, due to the limited sample size, a larger database is still needed to further validate the performance of preterm labor prediction systems before transferring them to clinical practice. Despite these limitations, we believe that this work constitutes a significant step towards putting the EHG technique into clinical practice.

3.5 Conclusions

By optimizing feature subspace with genetic algorithms, we showed the feasibility of developing a preterm labor prediction system with a high generalization capacity using simple and easy-to-interpret algorithms such as LDA, LR and KNN, obtaining an average F1-score of 89.34 ±3.5%, 86.87 ±4.53% and 84.63 ±2.76%, respectively, for the testing dataset. We found that temporal, spectral and non-linear EHG parameters computed in different bandwidths provide complementary information for predicting preterm labor. In addition, we further proved that the information extracted from multichannel recordings was also complementary among channels. A simple aggregation ensemble classifier can obtain more reliable preterm labor prediction systems than individual weak base classifiers, achieving higher average metrics and lower variability between partitions. The average F1-score of the ensemble classifier was about 92.04 ±2.97% for an incoming new dataset previously unseen by the model, which was significantly higher than that of commonly used obstetric techniques.

The optimized feature subset requires a few further features to be computed, which, together with use of easily interpreted classifier algorithms, would contribute to implementing preterm labor prediction systems in real-time and improve clinical staff's acceptance of the EHG technique, thus promoting its transferability to clinical practice.

3.6 References

[1] C. Leung, Born too soon, *Neuroendocrinology Letters*, vol. 25, no. SUPPL. 1, J. L. CP Howson, MV Kinney, Ed., pp. 133–136, 2004, ISSN: 0172780X. DOI: 10.2307/3965140.

[2] G. T. Mandy, Short-term complications of the preterm infant, *UpToDate*, vol. 46, no. 4965, pp. 1–17, 2019.

[3] R. E. Garfield and W. L. Maner, Physiology and electrical activity of uterine contractions, *Seminars in Cell and Developmental Biology*, vol. 18, no. 3, pp. 289–295, 2007, ISSN: 10849521. DOI: 10.1016/j.semcdb.2007.05.004.

[4] V. Berghella, E. Hayes, J. Visintine, and J. K. Baxter, Fetal fibronectin testing for reducing the risk of preterm birth, *Cochrane Database of Systematic Reviews*, no. 4, 2008, ISSN: 1469493X. DOI: 10.1002/14651858.CD006843.pub2.

[5] M. Pandey, M. Chauhan, and S. Awasthi, Interplay of cytokines in preterm birth, *Indian Journal of Medical Research*, vol. 146, no. September, pp. 316–327, 2017, ISSN: 09715916. DOI: 10.4103/ijmr.IJMR_1624_14.

[6] 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, ISSN: 00029378. DOI: 10.1016/0002-9378(93)90456-S.

[7] 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, 02TR01, 2018, ISSN: 13616579. DOI: 10.1088/1361-6579/aaad56.

[8] J. Mas-Cabo *et al.*, Electrohysterogram for ANN-Based Prediction of Imminent Labor in Women with Threatened Preterm Labor Undergoing Tocolytic Therapy, *Sensors*, vol. 20, no. 9, p. 2681, 2020, ISSN: 14248220. DOI: 10.3390/s20092681.

[9] A. Lemancewicz *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, ISSN: 02085216. DOI: 10.1016/j.bbe.2015.11.005.

[10] M. Brennan, M. Palaniswami, and P. Kamen, Do existing measures of Poincareé plot geometry reflect nonlinear features of heart rate variability? *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 11, pp. 1342–1347, 2001, ISSN: 00189294. DOI: 10.1109/10.959330.

[11] M. Son and E. S. Miller, Predicting preterm birth: Cervical length and fetal fibronectin, *Seminars in Perinatology*, vol. 41, no. 8, pp. 445–451, 2017, ISSN: 1558075X. DOI: 10.1053/j.semperi.2017.08.002.

[12] K. P. Murphy, *Machine Learning: A Probabilistic Perspective.* 2012, pp. 1–100, ISBN: 978026208029.

[13] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, *Electronics (Switzerland)*, vol. 8, no. 8, 2019, ISSN: 20799292. DOI: 10.3390/electronics8080832.

[14] 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 pre-term delivery groups, *Medical and Biological Engineering and Computing*, vol. 46, no. 9, pp. 911–922, 2008, ISSN: 01400118. DOI: 10.1007/s11517-008-0350-y.

[15] F. Jager, S. Libenšek, and K. Geršak, Characterization and automatic classification of preterm and term uterine records, *PLoS ONE*, vol. 13, no. 8, O. Uthman, Ed., e0202125, 2018, ISSN: 19326203. DOI: 10.1371/journal.pone.0202125.

[16] C. Marque, J. Gondry, J. Rossi, N. Baaklini, and J. Duchêne, Surveillance des grossesses à risque par électromyographie utérine, *RBM-Revue Europeenne de Technologie Biomedicale*, vol. 17, no. 1, pp. 25–31, 1995, ISSN: 02220776. DOI: 10.1016/S0222-0776(00)88906-3.

[17] J. Mas-Cabo *et al.*, Robust Characterization of the Uterine Myoelectrical Activity in Different Obstetric Scenarios, *Entropy*, vol. 22, no. 7, p. 743, 2020, ISSN: 10994300. DOI: 10.3390/e22070743.

[18] 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 and Biological Engineering and Computing*, vol. 57, no. 2, pp. 401–411, 2019, ISSN: 17410444. DOI: 10.1007/s11517-018-1888-y.

[19] Y. Ye-Lin, J. Alberola-Rubio, G. Prats-boluda, A. Perales, D. Desantes, and J. Garcia-Casado, Feasibility and Analysis of Bipolar Concentric Recording of Electrohysterogram with Flexible Active Electrode, *Annals of Biomedical Engineering*, vol. 43, no. 4, pp. 968–976, 2015, ISSN: 15739686. DOI: 10.1007/s10439-014-1130-5.

[20] A. Diaz-Martinez *et al.*, A Comparative Study of Vaginal Labor and Caesarean Section Postpartum Uterine Myoelectrical Activity, *Sensors*, vol. 20, no. 11, p. 3023, 2020, ISSN: 1424-8220. DOI: 10.3390/s20113023.

[21] W. L. Maner, L. B. MacKay, G. R. Saade, and R. E. Garfield, Characterization of abdominally acquired uterine electrical signals in humans, using a non-linear analytic method, *Medical and Biological Engineering and Computing*, vol. 44, no. 1-2, pp. 117–123, 2006, ISSN: 01400118. DOI: 10.1007/s11517-005-0011-3.

[22] M. J. Katz, Fractals and the analysis of waveforms, *Computers in Biology and Medicine*, vol. 18, no. 3, pp. 145–156, 1988, ISSN: 00104825. DOI: 10.1016/0010-4825(88)90041-8.

[23] R. Esteller, G. Vachtsevanos, J. Echauz, and B. Litt, A Comparison of waveform fractal dimension algorithms, *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 48, no. 2, pp. 177–183, 2001, ISSN: 10577122. DOI: 10.1109/81.904882.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, ISSN: 10769757. DOI: 10.1613/jair.953. arXiv: 1106.1813.

[25] D. Alamedine, M. Khalil, and C. Marque, Comparison of different EHG feature selection methods for the detection of preterm labor, *Computational and Mathematical Methods in Medicine*, vol. 2013, 2013, ISSN: 17486718. DOI: 10.1155/2013/485684.

[26] O. Babatunde, L. Armstrong, J. Leng, and D. Diepeveen, A Genetic Algorithm-Based Feature Selection, *International Journal of Electronics Communication and Computer Engineering*, vol. 5, no. 4, pp. 899–905, 2014, ISSN: 2278-4209.

[27] W. Bouaguel, A New Approach for Wrapper Feature Selection Using Genetic Algorithm for Big Data, in Intelligent and Evolutionary Systems, 2016, pp. 75–83. DOI: 10.1007/978-3-319-27000-5_6.

[28] O. Okun, *Feature selection and ensemble methods for bioinformatics: Algorithmic classification and implementations*. 2011, pp. 1–445, ISBN: 9781609605575. DOI: 10.4018/978-1-60960-557-5.

[29] G. Doquire and M. Verleysen, A comparison of multivariate mutual information estimators for feature selection, in ICPRAM 2012 - Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods, vol. 1, 2012, pp. 176–185, ISBN: 9789898425980. DOI: 10.5220/0003726101760185.

[30] F. Wilcoxon, Individual Comparisons by Ranking Methods, *Biometrics Bulletin*, vol. 1, no. 6, p. 80, 1945, ISSN: 00994987. DOI: 10.2307/3001968.

[31] D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. 1989, vol. 27, ISBN: 9780201157673.

[32] H. Li, D. Yuan, X. Ma, D. Cui, and L. Cao, Genetic algorithm for the optimization of features and neural networks in ECG signals classification, *Scientific Reports*, vol. 7, 2017, ISSN: 20452322. DOI: 10.1038/srep41011.

[33] P. Fischer, T. G. Amaral, O. P. Dias, A. R. Wołczowski, and M. Lichtenstein, Genetic algorithm based optimization for EMG pattern recognition system, *IFAC Proceedings Volumes (IFAC-PapersOnline)*, vol. 14, no. PART 1, pp. 53–58, 2009, ISSN: 14746670. DOI: 10.3182/20090819-3-pl-3002.00011.

[34] D. Beasley, D. R. Bull, and R. R. Martin, An overview of genetic algorithms : Part 1, fundamentals, *University Computing*, vol. 2, no. 15, pp. 1–16, 1993, ISSN: 0265-4385.

[35] M. Hollander, D. A. Wolfe, and E. Chicken, *Nonparametric Statistical Methods* (Wiley Series in Probability and Statistics). 2015, ISBN: 9780470387375. DOI: 10.1002/9781119196037.

[36] M. Pett, N. Lackey, and J. Sullivan, *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. 2003. DOI: 10.4135/9781412984898.

[37] C. Benalcazar-Parra *et al.*, Prediction of Labor Induction Success from the Uterine Electrohysterogram, *Journal of Sensors*, vol. 2019, pp. 1–12, 2019, ISSN: 1687-725X. DOI: 10.1155/2019/6916251.

[38] D. Alamedine, M. Khalil, and C. Marque, Comparison of Feature selection for Monopolar and Bipolar EHG signal, in Journees Recherche en Imagerie et Technologies pour la Santé (RITS 2015), 2015, pp. 100–101.

[39] S. P. Lim and H. Haron, Performance comparison of genetic algorithm, differential evolution and particle swarm optimization towards benchmark functions, *2013 IEEE Conference on Open Systems, ICOS 2013*, pp. 41–46, 2013. DOI: 10.1109/ICOS.2013.6735045.

[40] H. A. Frey and M. A. Klebanoff, The epidemiology, etiology, and costs of preterm birth, *Seminars in Fetal and Neonatal Medicine*, vol. 21, no. 2, pp. 68–73, 2016, ISSN: 18780946. DOI: 10.1016/j.siny.2015.12.011.

[41] J. Mas-Cabo, G. Prats-Boluda, J. Garcia-Casado, J. Alberola-Rubio, A. Perales, and Y. Ye-Lin, Design and Assessment of a Robust and Generalizable ANN-Based Classifier for the Prediction of Premature Birth by means of Multichannel Electrohysterographic Records, *Journal of Sensors*, vol. 2019, pp. 1–13, 2019, ISSN: 16877268. DOI: 10.1155/2019/5373810.

[42] 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, e77154, 2013, ISSN: 19326203. DOI: 10.1371/journal.pone.0077154.

[43] J. Alberola-Rubio *et al.*, Prediction of labor onset type: Spontaneous vs induced; role of electrohysterography? *Computer Methods and Programs in Biomedicine*, vol. 144, pp. 127–133, 2017, ISSN: 18727565. DOI: 10.1016/j.cmpb.2017.03.018.

[44] M. Liu, D. Zhang, and D. Shen, Ensemble sparse classification of Alzheimer's disease, *NeuroImage*, vol. 60, no. 2, pp. 1106–1116, 2012, ISSN: 10538119. DOI: 10.1016/j.neuroimage.2012.01.055.

[45] I. K. Ludmila and J. W. Christopher, Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy, *Machine Learning*, vol. 51, no. 2, pp. 181–207, 2003.

[46] Dietterich Thomas G., An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization, *Machine Learning*, vol. 40, no. 2, pp. 139–157, 2000.

[47] K. M. Ting and I. H. Witten, Issues in stacked generalization, *Journal of Artificial Intelligence Research*, vol. 10, pp. 271–289, 1999, ISSN: 10769757. DOI: 10.1613/jair.594.

Chapter 4

Assessment of Dispersion and Bubble Entropy Measures for Enhancing Preterm Birth Prediction Based on Electrohysterographic Signal

Félix Nieto-del-Amor[1], Raja Beskhani[1], Yiyao Ye-Lin[1,*], Javier Garcia-Casado[1], Alba Diaz-Martinez[1], Rogelio Monfort-Ortiz[2], Vicente Jose Diago-Almela[2], Dongmei Hao[3] and Gema Prats-Boluda[1]

[1] Centro de Investigación e Innovación en Bioingeniería, Universitat Politècnica de València, 46022 Valencia, Spain; feniede@ci2b.upv.es; rabes@ci2b.upv.es; jgarciac@ci2b.upv.es; adiaz@ci2b.upv.es; gprats@ci2b.upv.es

[2] Servicio de Obstetricia, H.U.P. La Fe, 46026 Valencia, Spain; monfort_isaort@gva.es; diago_vicalm@gva.es

[3] Faculty of Environment and Life, Beijing University of Technology, Beijing International Science and Technology Cooperation Base for Intelligent Physiological Measurement and Clinical Transformation, Beijing 100124, China; haodongmei@bjut.edu.cn

* Correspondence: yiye@ci2b.upv.es

Abstract

One of the remaining challenges for the scientific-technical community is predicting preterm births, for which electrohysterography (EHG) has emerged as a highly sensitive prediction technique. Sample and fuzzy entropy have been used to characterize EHG signals, although they require optimizing many internal parameters. Both bubble entropy, which only requires one internal parameter, and dispersion entropy, which can detect any changes in frequency and amplitude, have been proposed to characterize biomedical signals. In this work, we attempted to determine the clinical value of these entropy measures for predicting preterm birth by analyzing their discriminatory capacity as an individual feature and their complementarity to other EHG characteristics by developing six prediction models using obstetrical data, linear and non-linear EHG features, and linear discriminant analysis using a genetic algorithm to select the features. Both dispersion and bubble entropy better discriminated between the preterm and term groups than sample, spectral, and fuzzy entropy. Entropy metrics provided complementary information to linear features, and indeed, the improvement in model performance by including other non-linear features was negligible. The best model performance obtained an F1-score of 90.1 ±2% for testing the dataset. This model can easily be adapted to real-time applications, thereby contributing to the transferability of the EHG technique to clinical practice.

Keywords: Electrohysterography; Uterine electromyogram; Uterine electrical activity; Preterm birth prediction; Feature selection; genetic algorithm; Bubble entropy; Dispersion entropy; Sample entropy; Fuzzy entropy

4.1 Introduction

Preterm birth (deliveries before 37 weeks of gestation [1]) affect more than 15 million persons worldwide, involving 5 to 18% of pregnancies [2]. It is one of the leading causes of infant mortality, varying from 90% (<28 weeks) to 10% as gestation time advances [2]. Two-thirds of preterm births happen after the spontaneous onset of labor, while the remainder are medically indicated because of maternal or fetal complications [3]. Being born too soon increases the risk of neurodevelopmental impairments and respiratory and gastrointestinal complications [3]. In the survivors, this has been associated with 20% mental retardation, 50% cerebral palsy, and 33% eye injuries [4]. Preterm births also have a serious economic impact on public health systems; the average cost of a preterm birth is 5–10 times higher than a term birth [5]. As for an extremely preterm baby, born before 28 weeks of gestation in Canada, it costs an average of $67,467 for the first ten years of its life [6]. Saving a baby weighing less than 750 g costs more than $117,000, the highest costing procedure in Canada's public health system [7].

Several techniques have been proposed in the literature, such as monitoring uterine dynamics by tocodynamometry (TOCO), cervix length, Bishop score, and biochemical markers, to determine the risk of preterm birth [8]. Cervical length is seen as one of the best birth predicting methods [8], although several studies report that this criterion has proven to be insufficient or inaccurate [9, 10]. Due to its low positive predictive values and sensitivities, routine cervical length assessment is not recommended in women at low risk of preterm birth [9]. Monitoring uterine activity is routinely used by obstetricians during labor. The two most widespread current methods are: directly through an intrauterine pressure catheter (IUPC) and indirectly through external TOCO. However, both have serious disadvantages and limitations in their use. IUPC is an invasive method that can only be used during labor, which can increase the risk of infection and may even harm the fetus or the mother. Although TOCO is safe, it uses pressure transducers on the abdomen and has poor sensitivity and precision [11]. Biological fluids have been used as biochemical markers to predict preterm births, although systematic reviews have indicated that no single biomarker or combination of such could be identified to reliably predict the preterm birth risk or pregnancy outcome [12]. None of them has been proven to objectively and precisely estimate the time of delivery and whether or not it will be premature [8]. Electrohysterography (EHG) has emerged as a promising technique to identify the risk of preterm birth due to its high sensitivity [13]. EHG is the recording of changes in bioelectrical potential of the uterine myometrial cells and can be picked up on the human abdominal surface. During pregnancy, the uterine myometrial cells undergo a process of increased excitability and bioelectric propagability due to the larger number of gap-junctions, which end up leading to coordinated high-intensity contractions that give rise to labor. These electrophysiological changes have been shown to be reflected in an increased EHG signal amplitude associated with the number of uterine cells involved in the contractions [13]. On the other hand, the shift of the EHG signal spectral content to higher frequencies has been associated with increased cell excitability [13, 14]. The spectral content of the EHG signal has been widely studied and categorized by the frequency bandwidth: the whole bandwidth (WBW) ranges from 0.1–4 Hz, the slow wave, which is related to uterine contractions, and the fast wave which is usually subdivided into two components: fast wave low (0.13–0.26 Hz), which has been associated with signal propagation, and fast wave high (FWH) (0.34–0.88 Hz), which is related to cell excitability [13]. FWH is usually extended for study to 0.34–4 Hz [15].

Due to the non-linear nature of the biological system, other authors have proposed the use of different entropy measures to characterize EHG signals [16, 17]. The approximate entropy algorithm aims at obtaining a statistically valid measure of entropy for noisy biomedical time series and represents the probability that similar patterns (delay vectors) in a time series will remain similar once the pattern lengths are increased (extended delay vectors), thereby providing a natural measure of the time series regularity [18]. Lemancevicz et al. found that approximate entropy computed in the 0.24–4 Hz bandwidth for women with threatened preterm birth who delivered in less than 7 days was significantly higher than those who finally delivered in more than 7 days, suggesting that the EHG signal becomes more irregular as pregnancy progresses [19]. However, approximate entropy has been shown to be a biased estimator and highly sensitive to the number of data samples [20], while sample entropy is a modification of approximate entropy, in which self-matches are not included in calculating the probability [21]. A lower value of sample entropy also indicates more self-similarity in the time series. Sample entropy is more independent of data samples and behaves more consistently than approximate entropy [21]. This latter has been widely used to characterize EHG signals for discriminating preterm and term birth records [15] and also for distinguishing between women with threatened preterm birth undergoing tocolytic therapy who finally deliver in less or more than 7 days [17, 22].

In contrast to [19], the results of these works pointed to a reduction of entropy measurements (signal complexity) as labor approaches. Despite the promising results, sample entropy has been shown to have some drawbacks, e.g., it may be unstable and obtain unreliable results for short time series. It has also been shown to be sensitive to the configuration of its internal parameter values (embedding dimension m and scaling factor r) and can be too time-consuming for long data [23].

To deal with some of the deficiencies of sample entropy, fuzzy entropy is based on the concept of fuzzy sets. It presents a stronger relative consistency and shows less dependence on data length than sample entropy [23]. It has also been shown that the soft and continuous boundaries of fuzzy functions ensure continuity. Nevertheless, there are even more degrees of freedom for choosing internal parameters than for sample entropy, since both membership function and fuzzy power were introduced to define the boundary [24].

Since selecting the internal parameters is a critical issue in obtaining any entropy metrics, Manis et al. introduced a new definition of entropy known as bubble entropy, which is derived from permutation entropy, in which the vectors in the embedding space of the time series are ranked [25]. It quantifies the effort (number of swaps) required by the permutation process, which is carried out by the "bubble sort" algorithm, from which this measure obtains its name. By counting the number of swaps performed for each vector, a more coarse-grained distribution is obtained to compute the conditional Rényi entropy, which is the combination of the conditional permutation entropy [26] and Rényi permutation entropy [25, 27]. Bubble entropy has been shown to be almost free of internal parameters, since the scaling factor r is totally eliminated in its definition, and the importance of embedding dimension m is significantly reduced. It has also been proven to be remarkably stable and has a greater power to distinguish subjects with congestive heart failure and normal sinus rhythm than both sample entropy and permutation entropy [25]. Manis et al. proposed using bubble entropy for discriminating congestive heart failure from a healthy control group [25].

In 2016, Rostaghi and Azami proposed using dispersion entropy, another entropy measure, to quantify the complexity of a time series [28]. Dispersion entropy is based on the symbolic dynamics or pattern and Shannon entropy to quantify the randomness of the times series. The concept of symbolic dynamic arises from a coarse-graining of the signal measurements, i.e., the signal is transformed into a new one with only a small number of patterns. The transformation is achieved by using mapping functions such as a linear or normal cumulative distribution function, among others [28, 29]. In this way the study of the dynamic patterns of the time series is simplified to a distribution of symbol sequences. Dispersion entropy can detect a change of simultaneous frequency and amplitude and is relatively insensitive to noise since a small change in amplitude value will not vary the class label of the pattern [28]. Again, this entropy measure depends on the selection of the embedding dimension m and the number of classes c. For lower c values, there are few patterns to which to assign the time series, thus, underestimating the signal complexity, while if c is too high, small variations in the signal can cause a change of class, making it sensitive to noise [15]. It has been shown to perform better in detecting abrupt signal and noise robustness testing, has better stability for both simulated and real-word signals [30], and requires less computation time than sample entropy [29]. Dispersion entropy is also faster than other related entropy measures due to the fact that it does not need to either sort the amplitude values of each embedding vector or calculate every distance between any two composite delay vectors with embedding dimensions m and m + 1 [28]. Kafantaris et al. found that dispersion entropy obtained significantly higher values for both atrial premature beats and premature ventricular contraction electrocardiogram signals than healthy subjects [31]. Dispersion entropy of the discrete wavelet transformed EEG has also been used for differential diagnosis of health control, mild cognitive impairment, and Alzheimer's disease [32]. Tripathy et al. developed an automated sleep stage classification system, in which after some transformation of temporal EEG signals, they computed dispersion and bubble entropy. Their results could discriminate between different sleep stages with greater accuracy (>85%) [33].

To date, there has been no evidence of the utility of either bubble entropy or dispersion entropy for characterizing EHG signals; the aim of the present study was, therefore, to analyze the discriminatory capacity of bubble entropy and dispersion entropy EHG signals to differentiate between women who deliver at term and prematurely. We also attempted to determine whether these entropy measures can further improve the EHG feature space for predicting preterm birth by analyzing their complementary information to other EHG characteristics.

4.2 Materials and Methods

4.2.1 Database Description

A total of 326 EHG registers were analyzed from two public databases available in Physionet conducted on pregnant women between 22 and 37 weeks of gestation: "Term-Preterm EHG Database" (TPEHG DB) [15] and the "The Term-Preterm EHG Dataset with tocogram" (TPEHGT DS) [34] obtained by the Department of Obstetrics and Gynecology of the Ljubljana University Medical Center. Of the total, there were 275 term births and 51 preterm births (<37 weeks of gestations). The protocol used to obtain the EHG registers consisted of placing four electrodes (E1, E2, E3, and E4) on the woman's abdomen to obtain three bipolar channels (S1, S2, and S3), with a pairwise distance of 7 cm (see Figure 4.1) [15]. The sampling rate was set to 20 Hz. The signals were further band-pass filtered between 0.1 and 4 Hz using a fifth-order digital zero-phase Butterworth filter.

4.2.2 EHG Signal Analysis

We first excluded from the study the corrupt signal segments from the recordings (motion-artifacts and respiratory interference) by a double-blind process conducted

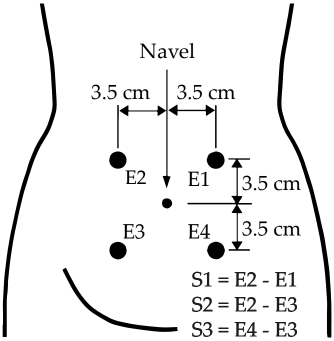


Figure 4.1: Recording protocol of EHG signals. Modified from [34].

by two experts. A whole-window analysis was then carried out to characterize the EHG signals, since this has been shown to provide relevant information on the uterine electrophysiological state without the need to identify the EHG-bursts associated with contractions embedded in the records, which is more suitable for future "real-time" applications [15, 16, 17]. The analysis used 120 s moving windows with a 50% overlap, a good trade-off between information-loss and computational cost [22].

Since we aimed to determine whether the different entropy metrics could further enhance preterm birth prediction by analyzing their complementary information to other EHG characteristics, we computed a set of 46 temporal, spectral, and non-linear features [8] to characterize EHG signals, the analysis windows, and EHG recordings. These were organized in different feature groups, as shown in Table 4.1. We then computed the median value of the total analyzed windows to obtain a unique representative value for each recording session and channel. We calculated EHG signal peak-to-peak amplitude (App) in both the whole EHG (WBW) bandwidth 0.1–4 Hz and Fast Wave High (FWH) bandwidth 0.34–4 Hz since this latter seems to be more sensitive to labor proximity [15]. As the EHG signal is mainly distributed between 0.2 and 1 Hz, we computed various spectral features [16, 17] to quantify the signal's energy distribution, including dominant frequency DF1 and DF2 computed in the 0.2–1 Hz and 0.34–1 Hz ranges, respectively; mean frequency (App) and power spectrum deciles (D1, . . . , D9) in the 0.2–1 Hz bandwidth; normalized subband energy (NormEn) (0.2–0.34 Hz, 0.34–0.6 Hz and 0.6–1 Hz); high (0.34–1 Hz)-to-low (0.2–0.34 Hz) frequency energy ratio (H/L Ratio); teager energy, and spectral moment ratio (SpMR). Since the analysis bandwidth is a key factor in estimating the non-linear features [16, 17], we calculated them in both whole EHG and FWH bandwidths. These latter included: Lempel-Ziv index (binary (LZBin) and multi-state n = 6 (LZMulti) Lempel-Ziv index, which evaluates time series complexity by measuring how "diverse" the patterns embedded in a time series are; time reversibility (TimeRev), which estimates the dynamic flows' similarity in forward (natural) time and reverse time and can be considered a measurement of the degree of signal nonlinearity [35]; Katz fractal dimension (KFD), as the uterine myoelectric activity has also been shown to possess fractal properties, which is another way of measuring self-similarity [36]; the Poincaré ellipse metrics were computed since the "present" EHG signal amplitude might significantly influence the "following" values. We, therefore, represented the Poincaré ellipse of consecutive EHG signal amplitudes (EHG[n] vs. EHG[n-1]) and extracted the main metrics (minor axis (SD1), major axis (SD2), square root of variance (SDRR, $\sqrt{(\mathrm{SD1}^2 + \mathrm{SD2}^2)/2}$, and SD1/SD2 ratio) [37]; spectral entropy (SpEn); sample entropy (SampEn); fuzzy entropy (FuzEn); dispersion entropy (DispEn), and bubble entropy (BubbEn). We performed an internal parameter sweep of the entropy measures to optimize their performance in discriminating preterm and term delivery by selecting the internal parameter combination associated with the lowest Wilcoxon Rank-Sum Test p-value when comparing preterm and term groups. For SampEn, the embedding dimension m was grid searched from 2 to 5, while the scaling factor r swept from 0.05 to 0.3 with a step of 0.05 times the standard deviation of the time series. Both the m and r ranges were considered to achieve reliable results for physiological data [38]. Due to the high number of degrees of freedom required for internal parameter selection to estimate fuzzy entropy (embedding dimension m, scaling factor r, fuzzy power n, and membership function), we evaluated several fuzzy membership functions as proposed in [24]: triangular, trapezoidal, Z-shaped, bell-shaped, gaussian, constant-gaussian, and exponential functions. For each membership function, we used the optimized parameter r and power n for discriminating biomedical signals [24], sweeping the embedding dimension from 2 to 5. We varied the dispersion entropy internal parameter m from 2 to 5 and the class number c between 3 and 9. The range for m and c was selected according to the literature, satisfying cm $<$ Z, Z being the length of the time series [29]. We assessed the performance of five mapping functions to compute DispEn: linear, normal cumulative distribution function, tangent sigmoid, logarithm sigmoid, and the sorting method. Finally, we varied the embedding dimension of bubble entropy m from 2 to 40, since stability has been shown to improve as m increases [25]. For SampEn, FuzEn, SpEn, and DisEn, the time delay parameter was fixed to 1 to avoid loss of information of high frequency components without excessively increasing the computational cost [24, 28]. Table 4.2 shows the optimized internal parameters of the entropy measures best able to discriminate preterm and term records (lowest p-value of Wilcoxon Rank-Sum Test). Each optimized entropy was included for further analysis to predict preterm birth.

To analyze the clinical value of different entropy measures for predicting preterm birth, we designed prediction models 1 and 2 using a subset of entropy measures previously used in EHG analysis: $\mathrm{En}_{\mathrm{SFS}}$ set (sample entropy, fuzzy, and spectral entropy) and all entropy measures $\mathrm{En}_{\mathrm{ALL}}$ (see Table 4.3). We also defined models 3–6 to determine whether entropy measures provided complementary information to other EHG characteristics and obstetric data. Table 4.3 shows the input features of the six preterm birth prediction models developed in this work.

Table 4.1: EHG features for predicting preterm birth including the composition of each feature group. SampEn, FuzEn, and SpEn were included in $\mathrm{En}_{\mathrm{SFS}}$ and $\mathrm{En}_{\mathrm{ALL}}$ subsets to analyze the additional value of bubble entropy and dispersion entropy for predicting preterm birth in relation to other entropy measures.

	Linear Features (L)	**Non-Linear Features (NL)**	**EnSFS**	**EnALL**	**Obstetric Data**
Number of features	20/channel	16/channel	6/channel	10/channel	5
Included features	App MeanF. DF1, DF2 NormEn H/L Ratio [D1–D9] Teager Energy SpecMR	LZBin LZMulti (n = 6) TimeRev KFD SD1 SD2 SDRR SD1/SD2	SampEn FuzEn SpEn	SampEn FuzEn SpEn DispEn BubbEn	Maternal age Parity Abortions Weight Week of gestation (WOG)

Table 4.2: Optimum internal parameters of each entropy measure in 0.1–4 Hz and 0.34–4 Hz bandwidths for channels S1, S2, and S3.

	Channel S1	**Channel S2**	**Channel S3**
$\mathrm{SampEn}_{\mathrm{WBW}}$	m = 3, r = 0.15	m = 3, r = 0.1	m = 2, r = 0.1
$\mathrm{SampEn}_{\mathrm{FWH}}$	m = 2, r = 0.3	m = 3, r = 0.1	m = 2, r = 0.3
$\mathrm{FuzEn}_{\mathrm{WBW}}$	m = 5, r = 0.0077, n = 3, exponential function	m = 5, r = 0.0077, n = 3, exponential function	m = 2, r = 0.0077, n = 3, exponential function
$\mathrm{FuzEn}_{\mathrm{FWH}}$	m = 5, r = 0.0077, n = 3, exponential function	m = 5, r = 0.0077, n = 3, exponential function	m = 2, r = 0.0077, n = 3, exponential function
$\mathrm{DispEn}_{\mathrm{WBW}}$	m = 2, c = 3, linear	m = 2, c = 3, linear	m = 2, c = 3, linear
$\mathrm{DispEn}_{\mathrm{FWH}}$	m = 2, c = 3, linear	m = 3, c = 4, linear	m = 2, c = 7, logsig
$\mathrm{BubbEn}_{\mathrm{WBW}}$	m = 23	m = 23	m = 26

4.2.3 Classifier Design and Evaluation

Due to the fact that only about 12% of the women undergoing regular check-ups deliver prematurely, there is a high imbalance rate between the two target classes in the original database, term and preterm. So as to mitigate the bias of conventional classification algorithms towards the majority class, obtaining low sensitivity for true preterm birth, we used the synthetic minority oversampling technique (SMOTE, k = 5) to obtain balanced preterm and term birth data [39]. This latter consisted of generating synthetic samples for the minority class, taking into account the original feature space. The conventional holdout method (30 partitions) was used to design and val-

Table 4.3: Composition of feature model depending on feature group considered.

Model	Acronym	Input EHG Features	Obstetrical Data	Initial Features
1	$\mathrm{En}_{\mathrm{SFS}}$	$\mathrm{En}_{\mathrm{SFS}}$	No	18
2	$\mathrm{En}_{\mathrm{ALL}}$	$\mathrm{En}_{\mathrm{ALL}}$	No	30
3	Linear	Linear	Yes	65
4	LNL	Linear, NL	Yes	113
5	$\mathrm{LEn}_{\mathrm{ALL}}$	Linear, $\mathrm{En}_{\mathrm{ALL}}$	Yes	95
6	$\mathrm{LNLEn}_{\mathrm{ALL}}$	Linear, NL, $\mathrm{En}_{\mathrm{ALL}}$	Yes	143

idate the classifiers. For each partition, the whole balanced database was randomly split into training (1/3), validation (1/3), and testing (1/3) with the same proportion between classes for designing, validating, and testing the classifier. We used the same partitions for designing and testing the classifier to compare the performance of the different models (see Table 4.3) for predicting preterm birth.

Since we attempted to evaluate the complementary information between entropy metrics and linear and other non-linear EHG features, we preferred to use a feature selection technique rather than dimensionality reduction. Feature selection is the process of obtaining a subset of relevant features to construct a machine learning model, it removes "irrelevant" features that do not contribute much to the classification problem and keeps the most relevant and complementary information to discriminate the target classes. The computational cost is also reduced by removing some of the features [40]. To optimize feature subset selection, we used the genetic algorithm, which is a random search strategy that provides a trade-off between classification performance and search complexity for a moderate and/or large number of features [41]. Both population size and genome length were fixed to the number of the model's input features (N) [42]. The tournament function was established with a size of 2 and an elite count of 2 to create the next generation population [42]. The crossover probability of combining the genetic information of parents to generate new offspring was typically assumed between 0.6 and 1, increasing the randomness of the children generation for a lower value [43]. The convergence to a lower minimum is better with low values (<0.1) of mutation probability, which is used to maintain the genetic diversity between generations [43, 44]. Arithmetic crossover was used with a probability of 0.8 and uniform mutation with a probability of 0.01. Finally, the genetic algorithm's termination condition was achieved if the fitness function did not improve noticeably for 150 consecutive generations (differential tolerance: 10^{-6}). To analyze the complementary information between input features we preferred to use linear classification methods without any complex data transformation by the linear discriminant classifier (LDA) to design the preterm birth prediction model.

Figure 4.2 shows a scheme of the optimization of features. Initially, we started balancing the original imbalanced dataset for each initial feature of the model (see Table 4.3). In the first step of the genetic algorithm, a set of randomly generated

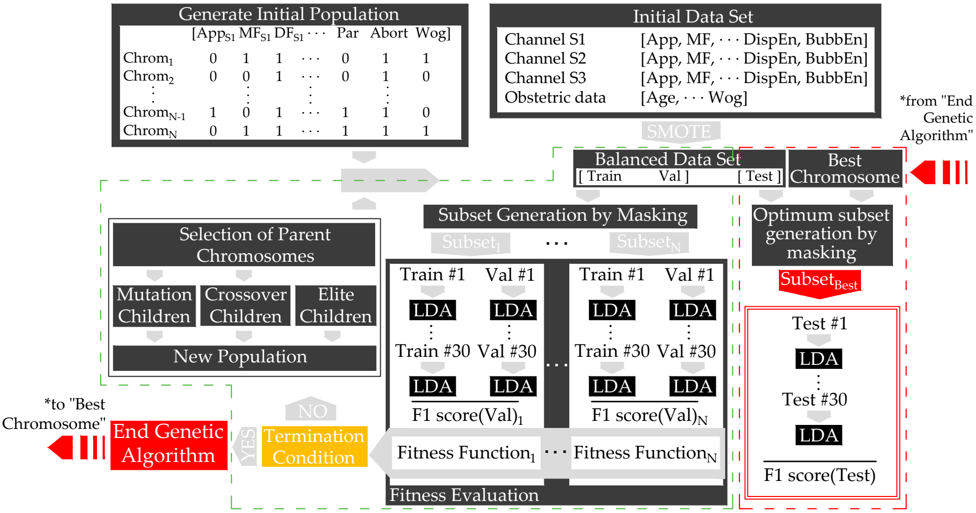

Generate Initial Population
[App$_{S1}$ MF$_{S1}$ DF$_{S1}$ ··· Par Abort Wog]
Chrom$_1$ 0 1 1 ··· 0 1 1
Chrom$_2$ 0 0 1 ··· 0 1 0
Chrom$_{N-1}$ 1 0 1 ··· 1 1 0
Chrom$_N$ 0 1 1 ··· 1 1 1
Initial Data Set
Channel S1 [App, MF, ··· DispEn, BubbEn]
Channel S2 [App, MF, ··· DispEn, BubbEn]
Channel S3 [App, MF, ··· DispEn, BubbEn]
Obstetric data [Age, ··· Wog]
*from "End Genetic Algorithm"
SMOTE
Balanced Data Set
[Train Val] [Test]
Best Chromosome
Subset Generation by Masking
Subset$_1$ ··· Subset$_N$
Optimum subset generation by masking
Subset$_{Best}$
Selection of Parent Chromosomes
Mutation Children
Crossover Children
Elite Children
New Population
Train #1 Val #1
LDA LDA
Train #30 Val #30
LDA LDA
F1 score(Val)$_1$
F1 score(Val)$_N$
Test #1
LDA
Test #30
LDA
F1 score(Test)
*to "Best Chromosome"
End Genetic Algorithm
YES
NO
Termination Condition
Fitness Function$_1$ ··· Fitness Function$_N$
Fitness Evaluation

Figure 4.2: Diagram of the genetic algorithm for selecting the optimized feature subset to predict preterm birth based on EHG (green dashed line). The red dashed line represents the calculation of the performance of the test group masked by the best chromosome obtained from the optimization of the genetic algorithm, considering: training dataset (Train), validation dataset (Val), testing dataset (Test), chromosome (Chrom), population size (N).

chromosomes masked the balanced data set, creating a feature subset. The mask (selected features), which corresponded to an i-chromosome, was set to the balanced data set obtaining the i-subset, with 1 ≤ i ≤ N, N being the model input features. We then used the LDA classifier to design the prediction model using each feature i-subset for the training dataset. The model was then scored by a fitness function in the validation dataset, defined as the mean F1-score of the 30 validation datasets weighted by the number of features being used in each iteration.

$$\text{Fitness function} = \max\{\overline{\text{F1-score}} \cdot (\text{NFeat} - \text{NCFeat})\} \tag{4.1}$$

where:

- NFeat is the number of features of the initial set.
- NCFeat is the number of features of the current subset.

The two best scored chromosomes (elite children) and new ones derived from mutation and crossover processes created a new population. This process was repeated until the termination condition was achieved, giving rise to the optimum feature subset (best chromosome).

For the different performance comparisons of the prediction models, we also computed the following metrics for the dataset testing: F1-score, accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under curve (AUC). Likewise, we carried out the Friedman nonparametric test to analyze statistical differences in different metrics for the different models. The Wilcoxon Rank-Sum test was then used for pair-wise evaluation of the classifiers, checking the similarity of their performance.

4.3 Results

Figure 4.3 shows box and whisker plots of different entropy measures using the optimal configuration of their internal parameters for both the whole and FWH bandwidths for preterm and term birth records. In general, the different entropy measures from the preterm group showed lower values than those from the term group, suggesting increased signal predictability as labor approaches, although some controversial results were obtained with contradictory tendencies for sample, fuzzy, and spectral entropy in channel S1. Regardless of the recording channel, the entropy measures computed from the FWH bandwidth offered better separability between preterm and term groups and obtained lower p-values. The spectral entropy of the preterm group in the FWH bandwidth was significantly lower than that of the term group for channel S3. In sample and fuzzy entropy, statistically significant differences were found between the preterm and term groups for both channel S2 and S3. Fuzzy entropy seemed to offer slightly better separability for channel S3 when compared with sample entropy. Dispersion entropy obtained significantly different values in distinguishing preterm and term records for both channel S1 and S3, the latter obtaining the greatest separability. Finally, bubble entropy obtained significantly lower values for the preterm group than the term group for all the recording channels and the two bandwidths studied. Again, channel S3 obtained the best differentiation outcome between the preterm and term groups.

Table 4.4 shows the optimized feature subsets obtained for each of the feature models. SpEn was selected to consider only classical entropy metrics (En_{SFS}). When extending the entropy characteristics with DispEn and BubbEn (En_{ALL}), only the latter was selected. In the rest of the models some features were shared between the optimum feature subsets. A large number of input features were computed for each model in different channels and EHG bandwidths. Due to the complexity of the multidimensional feature space and redundancy or complementarity between features, the genetic algorithm may reach the best chromosome with different combination of features for each model [44]. In spite of these issues, many features are part of the best chromosome of the models developed. For linear metrics, DF1, NormEn in 0.1–0.34 Hz, D6, D8, and SpMR were selected for every optimum subset. KFD in the whole bandwidth was selected in the two models that included it as input features, i.e., LNL and $\text{LNLEn}_{\text{ALL}}$. Regarding entropy metrics, only bubble entropy appeared in every optimum subset. Only week of gestation was chosen in all cases for obstetric features. Features such as Lempel-Ziv multistate, time reversibility, or Poincaré ellipse metrics

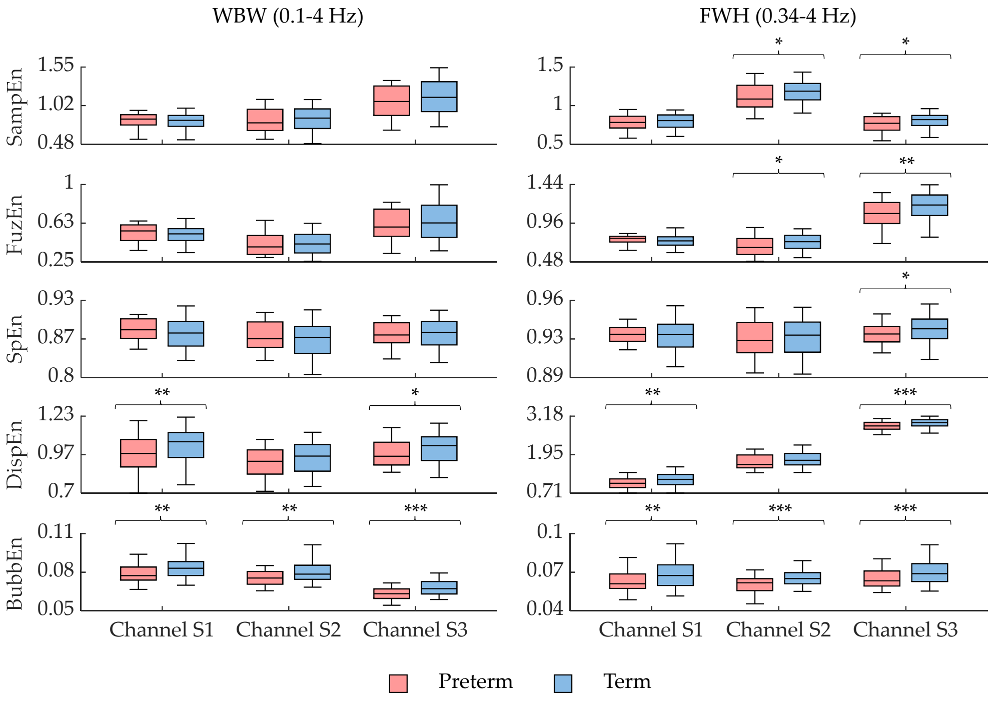


Figure 4.3: Box and whisker plot distributions of SampEn, FuzEn, SpEn, DispEn, and BubbEn using the optimal configuration of internal parameters indicated in Table 4.2 computed from EHG signals in different bandwidths and channels. *, ** and *** mean significant statistical difference (p-value ≤ 0.05, ≤ 0.01, and ≤ 0.001, respectively) between preterm and term records.

were never selected from any channel or bandwidth, which can be attributed to the fact that they contain redundant information with linear and entropy metrics.

As expected, in the preterm birth prediction models, the training dataset always obtained higher performance than the validation and testing datasets, the performance of these two latter being similar. Since the testing dataset performance denotes the model generalization capacity for the "never seen" incoming data, we only show in Table 4.5 the average performance of the testing dataset for the different preterm birth prediction models. Figure 4.4 shows the outcome of pairwise comparisons of the performance metrics of these prediction models using the Wilcoxon Rank-Sum test. Moderate average performance with relatively high variability was achieved when using only entropy measures for predicting preterm birth, the average F1-score being 63.7 ±5.1% and 76.8 ±3.2% for $\mathrm{En}_{\mathrm{SFS}}$ and $\mathrm{En}_{\mathrm{ALL}}$, respectively. The inclusion of dispersion and bubble entropy significantly enhanced the prediction model performance, and only $\mathrm{BubbEn}_{\mathrm{WBW}}$, S2, $\mathrm{BubbEn}_{\mathrm{WBW}}$, S3 are part of the

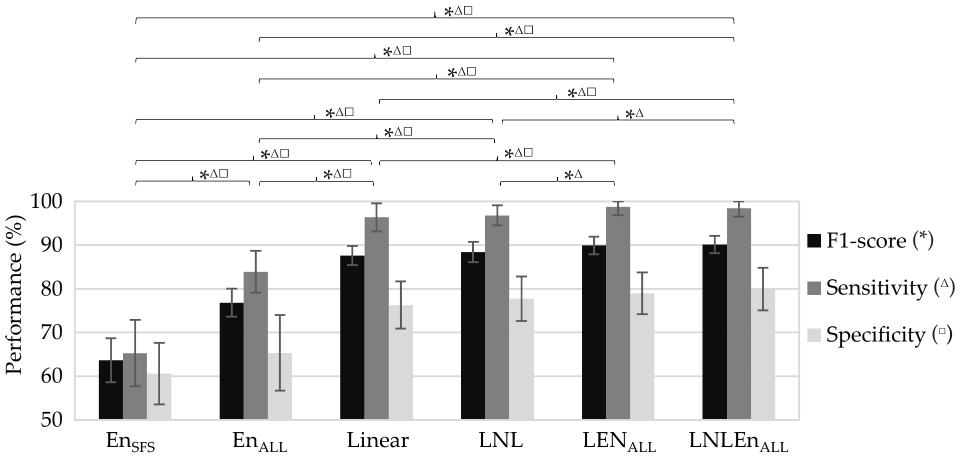


Figure 4.4: Comparison of the metrics' performance in the testing dataset for different preterm birth prediction models $*$, $\triangle$, and $\square$ mean a significant statistical difference (p-value ≤ 0.05) between classifiers' performance in F1-score, sensitivity, and specificity, respectively, by the Wilcoxon Rank-Sum test.

En_{ALL} best chromosome. The Linear model, which used linear EHG features and obstetric data, provided a significantly higher F1-score (87.6 ±2.2%) than En_{SFS} and En_{ALL}. The inclusion of other non-linear features (LNL model with binary and multistate Lempel-Ziv, Time reversibility, Katz fractal dimension, and Poincaré ellipse metrics) only provided a slightly better prediction outcome than Linear, without any significant difference (Linear 87.6 ±2.2% vs. LNL 88.4 ±2.3%), while entropy measures seemed to complement linear EHG features, obtaining a higher performance (Linear 87.6 ±2.2% vs. LEn_{ALL} 89.9 ±2%) with a statistically significant improvement in sensitivity, specificity, and F1-score. The best performance was achieved by the $\text{LNLEn}_{\text{ALL}}$, which used both linear and all non-linear EHG features, including all the entropy measures, with an average F1-score of 90.1 ±2%. The $\text{LNLEn}_{\text{ALL}}$ metrics were significantly higher than those of the other models (see Figure 4.4), except for LEn_{ALL}, for which no significant difference was found in any metric. The variability of $\text{LNLEn}_{\text{ALL}}$ metrics between partitions, especially for sensitivity, was lower than other models. Figure 4.5) shows the $\text{LNLEn}_{\text{ALL}}$ curve of $\text{LNLEn}_{\text{ALL}}$ for training, validation, and testing a dataset. The $\text{LNLEn}_{\text{ALL}}$ curve shows how the training partition presents the best performance with a larger area under the curve. The validation and testing partitions obtained a lower area under the curve than the training partition and an almost overlapping curve, which suggests a high power of generalization and, therefore, minimizes bias and variance in the classification.

Table 4.4: Optimum feature subset reached for the different initial feature models, defined in Table 3. The number of features obtained in each model feature repeated in the best chromosome of all the classifiers that use them as inputs are marked in bold.

Input Features Acronym	Selected Feature Subset	No. of Features
En$_{\text{SFS}}$	SpEn$_{\text{WBW, S2}}$, SpEn$_{\text{WBW, S3}}$	2
En$_{\text{ALL}}$	BubbEn$_{\text{WBW, S2}}$, BubbEn$_{\text{WBW, S3}}$	2
Linear	App$_{\text{WBW, S2}}$, DF1$_{\text{S2}}$, **NormEn$_{\text{0.2–0.34Hz, S2}}$**, NormEn$_{\text{0.2–0.34Hz, S3}}$, H/Lratio$_{\text{S1}}$, D3$_{\text{S1}}$, **D6$_{\text{S2}}$**, **D8$_{\text{S3}}$**, SpMR$_{\text{S3}}$, **WoG**	10
LNL	**DF1$_{\text{S2}}$**, DF1$_{\text{S3}}$, **NormEn$_{\text{0.2–0.34Hz, S2}}$**, NormEn$_{\text{0.2–0.34Hz, S3}}$, H/Lratio$_{\text{S1}}$, D3$_{\text{S1}}$, **D6$_{\text{S2}}$**, **D8$_{\text{S2}}$**, D8$_{\text{S3}}$, D9$_{\text{S2}}$, **SpMR$_{\text{S3}}$**, LZBin$_{\text{WBW, S3}}$, **KFD$_{\text{WBW, S1}}$**, **WoG**	14
LEn$_{\text{ALL}}$	App$_{\text{FWH}}$, S2, **DF1$_{\text{S3}}$**, DF2$_{\text{S1}}$, **NormEn$_{\text{0.2–0.34Hz, S2}}$**, NormEn$_{\text{0.2–0.34Hz, S3}}$, **D6$_{\text{S2}}$**, **D8$_{\text{S3}}$**, **SpMR$_{\text{S3}}$**, **BubbEn$_{\text{FWH, S3}}$**, Abortions, **WoG**	11
LNLEn$_{\text{ALL}}$	**DF1$_{\text{S2}}$**, DF2$_{\text{S1}}$, **NormEn$_{\text{0.2–0.34Hz, S2}}$**, D3$_{\text{S1}}$, **D6$_{\text{S2}}$**, **D8$_{\text{S2}}$**, D9$_{\text{S2}}$, **SpMR$_{\text{S3}}$**, **KFD$_{\text{WBW, S1}}$**, FuzEn$_{\text{FWH, S1}}$, **BubbEn$_{\text{WBW, S2}}$**, **BubbEn$_{\text{FWH, S3}}$**, **WoG**	12

Table 4.5: Average performance on testing dataset for the different feature subset models.

Input Features Acronym	F1-Score (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	AUC (%)
En$_{\text{SFS}}$	63.7 ±5.1	63 ±4.6	65.3 ±7.6	60.6 ±7	62.5 ±4.6	63.8 ±5.4	66.3 ±4.71
En$_{\text{ALL}}$	76.8 ±3.2	74.6 ±4.2	83.9 ±4.8	65.4 ±8.7	71.1 ±5	80.4 ±4.5	80.8 ±4.76
Linear	87.6 ±2.2	86.3 ±2.6	96.4 ±3.2	76.3 ±5.4	80.4 ±3.5	95.6 ±3.6	90 ±2.5
LNL	88.4 ±2.3	87.3 ±2.7	96.8 ±2.3	77.7 ±5.1	81.4 ±3.5	96.1 ±2.6	91.7 ±2.6
LEn$_{\text{ALL}}$	89.9 ±2	88.9 ±2.4	98.7 ±1.9	79 ±4.8	82.6 ±3.3	98.5 ±2.3	91.6 ±2.8
LNLEn$_{\text{ALL}}$	90.1 ±2	89.2 ±2.4	98.4 ±1.9	79.9 ±4.9	83.2 ±3.3	98.2 ±2.2	93.6 ±2.3

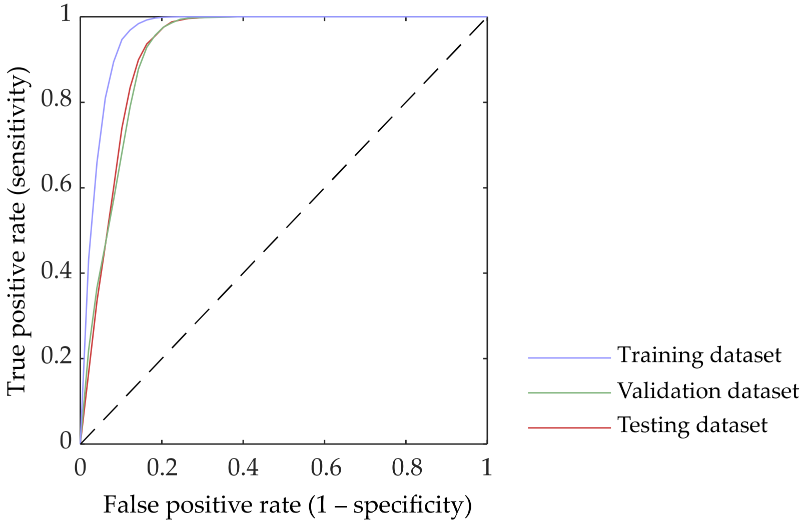


Figure 4.5: ROC curves of the LNLEn$_{\mathrm{ALL}}$ model that used linear, all non-linear EHG features and obstetric data including optimized feature subset for the training, validation, and testing dataset.

4.4 Discussion

In this work, we compared five entropy measures from three EHG channels computed in both FWH and WBW bandwidths for distinguishing between preterm and term delivery records. Our results showed that the EHG metrics from channel S3 generally obtained lower p-values between preterm and term delivery metrics, suggesting greater class separability. This finding is consistent with those found by other authors, who attempted to predict preterm birth using information extracted from the S3 channel due to its higher signal-to-noise ratio [45, 46, 47, 48]. We also confirmed that the different entropy measures computed in the FWH bandwidth provided higher separability between preterm and term delivery records than the WBW bandwidth, as we found in a previous work [16]. As for entropy measures, sample entropy was widely used for characterizing EHG signals acquired in women who had had regular check-ups, women with threatened preterm birth, and those who underwent labor induction [8, 22, 49]. Both fuzzy entropy and spectral entropy were previously proposed to distinguish preterm and term records [17]. As far as we know, this is the first time EHG has been characterized using dispersion entropy and bubble entropy, which have also been used for quantifying the regularity of other biomedical signals [25, 31, 32, 33].

With regard to the application of dispersion and bubble entropy to EHG characterization, despite eliminating the scale factor r in bubble entropy that makes it easier to search for optimization parameters, it still presents a certain dependency on embedding dimension m. In our application, preterm records obtained lower bubble entropy values than term records for high embedding dimension m, suggesting increasing signal predictability as labor approaches [8, 13]. In contrast, a low embedding dimension may lead to physiological misinterpretation throughout pregnancy. This finding was consistent with the observation made by Manis et al., with respect to the increase in the stability of the entropy measure as the embedding dimension increases [25]. We also found bubble entropy to be less sensitive to the signal bandwidth considered in the computation. In contrast, dispersion entropy was more sensitive to the signal bandwidth in which we computed this measure (see Figure 4.3, WBW vs. FWH). This may be due to the fact that dispersion entropy not only detects the signal complexity, but also instantaneous amplitude and frequency fluctuations [29].

We found that bubble entropy outperformed dispersion entropy and fuzzy entropy in discriminating preterm and term delivery patients and that these latter outperformed sample and spectral entropy. Our results agree with those of other authors who stated that dispersion entropy was found to be more consistent than sample entropy in characterizing the effect of age on the intrinsic stride-to-stride dynamics for gait maturation evaluation and in discriminating the non-invasive blood pressure signals of Dahl salt-sensitive hypertensive rats and rats protected from high-salt-induced hypertension [29]. Azami et al. also showed that dispersion entropy outperformed both sample entropy and fuzzy entropy for characterizing resting-state magnetoencephalogram regularity in Alzheimer's disease [50]. Fuzzy entropy is more effective than sample entropy and approximate entropy for distinguishing Alzheimer patients from normal subjects [51].

We also attempted to determine the redundancy and complementary information between input features, since redundant and/or irrelevant features may lead to high computational complexity and overfitting problems, thereby increasing the variance of the prediction model without reducing its bias [52]. Information redundancy can be detected by analyzing mutual information in a multidimensional feature space to obtain a high correlation between the chosen features subsets and the target class [53]. Nevertheless, estimating the mutual information (especially through estimating probability density functions) between high-dimensional variables is a hard task in practice due to the limited number of available data points for real-world problems [52]. In this work, we used a wrapper method based on a genetic algorithm for selecting complementary features to enhance the prediction model outcome while keeping redundancy features out. This latter has also been proven to outperform the filter method for predicting pregnancy and labor contractions [54].

While previous studies reported that non-linear and entropy features have been shown to better characterize the EHG signal than linear metrics [55, 56], our results showed that linear, non-linear, and entropy metrics complement each other for differentiating between preterm and term deliveries. This agrees with those of other authors who proposed using sample entropy together with linear features (root mean square, peak frequency, and median frequency) and obstetrical data for distinguishing term and preterm delivery records and achieved an AUC of 95% using a cross validation technique [45]. In a later work, feature ranking was proposed to determine the optimized feature subset, achieving a similar AUC of 94% using sample entropy, log detector, and other linear metrics as input features [47]. We found a great deal of redundant information between non-linear features, and only 4 of 78 non-linear and entropy features were included in the optimized feature subset for $\mathrm{LNLEn}_{\mathrm{ALL}}$. This could be associated with the fact that these latter attempted to quantify the same phenomena: signal regularity and complexity. Only entropy measures complement linear features, obtaining a significantly higher prediction performance (see Table 4.5 Linear vs. $\mathrm{LEn}_{\mathrm{ALL}}$); the improvement of prediction performance when including other non-linear features being negligible (Linear vs. LNL). For entropy measures, fuzzy and bubble entropy offered complementary information for predicting preterm and term delivery records, which were included in the optimized feature subset (see Table 4.4). Spectral, sample, and dispersion entropy were more likely to be redundant than fuzzy and bubble entropy. In our previous work [57], a similar input feature to the one included in this work (adding, in this case, dispersion and bubble entropy) was optimized using a genetic algorithm for feature selection and LDA for classifying. The results revealed that selected linear features maintain a correlation with those in this work. Dominant frequency in 0.1–1 Hz and 0.34–1 Hz, normalized energy in 0.1–0.34 Hz and spectral moment ratio were chosen in both studies. Decile 5 seemed to be a good discriminative feature, but in this case, it seems to be replaced by deciles 6, 8, and 9, indicating their complementarity and redundancy with decile 5. In contrast, peak-to-peak amplitude was selected by the genetic algorithm in other computed bandwidths not added to the input feature of the present work. In the non-linear parameters, only the Katz fractal dimension in the whole bandwidth appeared in the optimum feature subset in LNL and $\mathrm{LNLEn}_{\mathrm{ALL}}$. For entropy measures, the double selection of bubble entropy and the lack of other non-linear and entropy metrics in the final subset suggest that bubble entropy has a high discriminating power between term and preterm cases. As bubble entropy keeps enough redundant and complementary information with non-linear and other entropy metrics, they were not used in the optimum feature subset. Our results agree with those of Cuesta-Frau, who suggested that bubble entropy may offer complementary information to other entropy measures, such as permutation entropy, for predicting the risk of developing diabetes [58]. In this regard, fuzzy entropy was also found to complement entropy measures as the distribution entropy for differentiating both ictal and interictal EEG from normal EEG and for discriminating ictal from interictal EEG [59].

Our results are hardly comparable to many previous works in preterm birth prediction systems that used cross validation methods to design and validate the classifiers, without determining the real generalization capacity for incoming “never seen” data by the classifiers [45, 46, 47, 60]. The results of the prediction model obtained in this work even outperformed the results of our previous work, in which principal component analysis was used for dimension reduction of input features and multilayer perceptron artificial neural network to implement the classifier [61]. We believe that this prediction performance improvement was mainly due to the information optimization in feature space using the genetic algorithm, which can eliminate any redundant information and irrelevant features while keeping in the complementary information [41]. In addition, by optimizing information in feature space, we showed the feasibility of designing a preterm birth prediction system using simple linear classifiers, which are easily interpretable by clinicians. The complex classification algorithms, which can only be interpreted by experts, such as artificial neural network and/or support vector machines, can be dispensed with, which will considerably improve the transferability of the technique to clinical practice [62]. Instead of the mean efficiency index, which has been proposed as a robust indicator of uterine electrical activity efficiency from multichannel recordings [61], in this work, we used the EHG features extracted from the three individual channels since there is a greater degree of freedom in combining the information extracted from them.

Ahmed el al. used multivariate multiscale sample and fuzzy entropy in addition to univariate metrics to capture cross-channel dynamics of multichannel EHG recording and to characterize the interaction between the variates of complex systems to successfully discriminate between women who finally delivered at term and those who did so prematurely [63]. We found that the multivariate sample, fuzzy, and dispersion entropy measures obtained a relatively low model performance (∼60%) for the test dataset (result not shown here for the sake of brevity). The addition of these multivariate entropy measures to the univariate measures did not significantly improve the model performance, which means that no additional relevant information can be obtained from them.

In spite of its promising results, the present study has certain limitations that should be pointed out. There are various factors that make the transfer of the EHG technique to clinical practice difficult; the databases are small and highly imbalanced for the preterm birth class; for instance, in the public databases used in this work, the term/preterm ratio is around 7 to 1. A larger database will be needed to assess the robustness of these preterm birth prediction systems if they are to be used in clinical practice. Second, the lack of a standard protocol for the electrode position in EHG recordings is another factor that makes a shared database difficult. On the other hand, most of the prediction systems are based on neural networks or support vector machine, multilayer perceptron, or similar algorithms, which involve non-linear transformations of the input EHG features into high dimension space, in which data from the target classes offer better linear separability [40]. This could give rise to good prediction performance even when the input features apparently do not contain individually information to differentiate the target classes. Obstetricians often consider this type of classification algorithm as a “black box” or a “mathematician’s gadget” due to its being difficult to interpret [62] and so find it difficult to trust the predictions of these complex classifiers. In a previous study, we, therefore, attempted to develop a preterm birth prediction model using simple classifiers to avoid complex artificial intelligence algorithms, whose success depends mainly on the information embedded in the features [57]. Other clinically relevant measures, such as cervical length, fetal fibronectine, and/or interleukin 6, which has been proven to be one of the more effective techniques to predict preterm birth [8], were missing in these databases [15, 34]. This could partly be due to the fact that EHG signals were recorded from regular check-ups in women that did not show symptoms of preterm labor risk, and these measurements are not usually performed in this scenario. The inclusion of these additional clinical data to the predictor model could, therefore, further improve preterm birth prediction performance [17]. The commonly used SMOTE oversampling technique was employed to mitigate the imbalanced class problem [39]. Future studies should focus on the design and validation of preterm birth prediction systems using specifically imbalanced data learning algorithms.

4.5 Conclusions

Both dispersion and bubble entropy can be used to characterize EHG signals, providing a higher between-class distance for distinguishing between preterm and term delivery records than sample, fuzzy, or spectral entropy. A feature-selection method based on a genetic algorithm was used to determine redundant and complementary information between linear and non-linear EHG features. We found that non-linear features contained a great deal of redundant information, as did the different entropy measures. Nevertheless, the entropy measures offered complementary information to linear features and could achieve a significantly higher performance for predicting preterm birth. Bubble entropy was declared to be a high-performance term-preterm discriminator, even improving on dispersion entropy in individual and multidimensional approaches. By optimizing the information in the feature space using the genetic algorithm, we were able to design a preterm birth prediction system using a simple linear classifier that yielded an average F1-score of 90.1 ±2% for the test dataset. These results suggest that the proposed system has a high generalization capability for “never seen” incoming data and has great potential to bring the EHG technique closer to clinical practice.

4.6 References

[1] C. Leung, Born too soon, *Neuroendocrinology Letters*, vol. 25, no. SUPPL. 1, J. L. CP Howson, MV Kinney, Ed., pp. 133–136, 2004, ISSN: 0172780X. DOI: 10.2307/3965140.

[2] S. Chawanpaiboon *et al.*, Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis, *The Lancet Global Health*, vol. 7, no. 1, e37–e46, 2019, ISSN: 2214109X. DOI: 10.1016/S2214-109X(18)30451-0.

[3] 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, ISSN: 01406736. DOI: 10.1016/S0140-6736(08)60074-4.

[4] G. T. Mandy, Short-term complications of the preterm infant, *UpToDate*, vol. 46, no. 4965, pp. 1–17, 2019.

[5] S. Petrou, H. H. Yiu, and J. Kwon, Economic consequences of preterm birth: A systematic review of the recent literature (2009-2017), *Archives of Disease in Childhood*, vol. 104, no. 5, pp. 456–465, 2019, ISSN: 14682044. DOI: 10.1136/archdischild-2018-315778.

[6] K. M. Johnston *et al.*, The economic burden of prematurity in Canada, *BMC Pediatrics*, vol. 14, no. 1, pp. 1–10, 2014, ISSN: 14712431. DOI: 10.1186/1471-2431-14-93.

[7] N. X. Thanh, J. Toye, A. Savu, M. Kumar, and P. Kaul, Health Service Use and Costs Associated with Low Birth Weight - A Population Level Analysis, *Journal of Pediatrics*, vol. 167, no. 3, pp. 551–556, 2015, ISSN: 10976833. DOI: 10.1016/j.jpeds.2015.06.007.

[8] 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, 02TR01, 2018, ISSN: 13616579. DOI: 10.1088/1361-6579/aaad56.

[9] K. Lim *et al.*, Ultrasonographic Cervical Length Assessment in Predicting Preterm Birth in Singleton Pregnancies, *Journal of Obstetrics and Gynaecology Canada*, vol. 33, no. 5, pp. 486–499, 2011, ISSN: 17012163. DOI: 10.1016/S1701-2163(16)34884-8.

[10] T. Y. Euliano *et al.*, Monitoring uterine activity during labor: a comparison of three methods, *Am J Obstet Gynecol*, vol. 208, no. 1, pp. 66–67, 2013. DOI: 10.1016/j.ajog.2012.10.873.Monitoring.

[11] T. Y. Euliano, M. T. Nguyen, S. Darmanjian, J. D. Busowski, N. Euliano, and A. R. Gregg, Monitoring Uterine Activity during Labor: Clinician Interpretation of Electrohysterography versus Intrauterine Pressure Catheter and Tocodynamometry, *American Journal of Perinatology*, vol. 33, no. 9, pp. 831–838, 2016, ISSN: 10988785. DOI: 10.1055/s-0036-1572425.

[12] V. Berghella, E. Hayes, J. Visintine, and J. K. Baxter, Fetal fibronectin testing for reducing the risk of preterm birth, *Cochrane Database of Systematic Reviews*, no. 4, 2008, ISSN: 1469493X. DOI: 10.1002/14651858.CD006843.pub2.

[13] 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, ISSN: 00029378. DOI: 10.1016/0002-9378(93)90456-S.

[14] D. Schlembach, W. L. Maner, R. E. Garfield, and H. Maul, Monitoring the progress of pregnancy and labor using electromyography, *European Journal of Obstetrics and Gynecology and Reproductive Biology*, vol. 144, no. SUPPL 1, pp. 2–8, 2009, ISSN: 18727654. DOI: 10.1016/j.ejogrb.2009.02.016.

[15] 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 pre-term delivery groups, *Medical and Biological Engineering and Computing*, vol. 46, no. 9, pp. 911–922, 2008, ISSN: 01400118. DOI: 10.1007/s11517-008-0350-y.

[16] J. Mas-Cabo *et al.*, Robust Characterization of the Uterine Myoelectrical Activity in Different Obstetric Scenarios, *Entropy*, vol. 22, no. 7, p. 743, 2020, ISSN: 10994300. DOI: 10.3390/e22070743.

[17] J. Mas-Cabo *et al.*, Electrohysterogram for ANN-Based Prediction of Imminent Labor in Women with Threatened Preterm Labor Undergoing Tocolytic Therapy, *Sensors*, vol. 20, no. 9, p. 2681, 2020, ISSN: 14248220. DOI: 10.3390/s20092681.

[18] S. M. Pincus, Approximate entropy as a measure of system complexity, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 88, no. 6, pp. 2297–2301, 1991, ISSN: 00278424. DOI: 10.1073/pnas.88.6.2297.

[19] A. Lemancewicz *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, ISSN: 02085216. DOI: 10.1016/j.bbe.2015.11.005.

[20] M. Ferrario, M. G. Signorini, G. Magenes, and S. Cerutti, Comparison of entropy-based regularity estimators: Application to the fetal heart rate signal for the identification of fetal distress, *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 1, pp. 119–125, 2006, ISSN: 00189294. DOI: 10.1109/TBME.2005.859809.

[21] J. S. Richman and J. R. Moorman, Physiological time-series analysis using approximate entropy and sample entropy, *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, H2039–H2049, 2000, ISSN: 0363-6135. DOI: 10.1152/ajpheart.2000.278.6.H2039.

[22] 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 and Biological Engineering and Computing*, vol. 57, no. 2, pp. 401–411, 2019, ISSN: 17410444. DOI: 10.1007/s11517-018-1888-y.

[23] A. Humeau-Heurtier, Evaluation of Systems' Irregularity and Complexity: Sample Entropy, Its Derivatives, and Their Applications across Scales and Disciplines, *Entropy*, vol. 20, no. 10, p. 794, 2018, ISSN: 1099-4300. DOI: 10.3390/e20100794.

[24] H. Azami, P. Li, S. E. Arnold, J. Escudero, and A. Humeau-Heurtier, Fuzzy Entropy Metrics for the Analysis of Biomedical Signals: Assessment and Comparison, *IEEE Access*, vol. 7, pp. 104 833–104 847, 2019, ISSN: 2169-3536. DOI: 10.1109/access.2019.2930625.

[25] G. Manis, M. Aktaruzzaman, and R. Sassi, Bubble entropy: An entropy almost free of parameters, *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 11, pp. 2711–2718, 2017, ISSN: 15582531. DOI: 10.1109/TBME.2017.2664105.

[26] A. M. Unakafov and K. Keller, Conditional entropy of ordinal patterns, *Physica D: Nonlinear Phenomena*, vol. 269, pp. 94–102, 2014, ISSN: 01672789. DOI: 10.1016/j.physd.2013.11.015. arXiv: 1407.5390.

[27] Z. Liang *et al.*, EEG entropy measures in anesthesia, *Frontiers in Computational Neuroscience*, vol. 9, no. JAN, pp. 1–17, 2015, ISSN: 16625188. DOI: 10.3389/fncom.2015.00016.

[28] M. Rostaghi and H. Azami, Dispersion Entropy: A Measure for Time-Series Analysis, *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 610–614, 2016, ISSN: 10709908. DOI: 10.1109/LSP.2016.2542881.

[29] H. Azami and J. Escudero, Amplitude- and fluctuation-based dispersion entropy, *Entropy*, vol. 20, no. 3, p. 210, 2018, ISSN: 10994300. DOI: 10.3390/e20030210.

[30] Y. Li, X. Gao, and L. Wang, Reverse Dispersion Entropy: A New Complexity Measure for Sensor Signal, *Sensors*, vol. 19, no. 23, p. 5203, 2019, ISSN: 14248220. DOI: 10.3390/s19235203.

[31] E. Kafantaris, I. Piper, T. Y. M. Lo, and J. Escudero, Application of Dispersion Entropy to Healthy and Pathological Heartbeat ECG Segments, in In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019, pp. 2269–2272, ISBN: 9781538613115. DOI: 10.1109/EMBC.2019.8856554.

[32] J. P. Amezquita-Sanchez, N. Mammone, F. C. Morabito, and H. Adeli, A New dispersion entropy and fuzzy logic system methodology for automated classification of dementia stages using electroencephalograms, *Clinical Neurology and Neurosurgery*, vol. 201, 2021, ISSN: 18726968. DOI: 10.1016/j.clineuro.2020.106446.

[33] R. K. Tripathy, S. K. Ghosh, P. Gajbhiye, and U. R. Acharya, Development of automated sleep stage classification system using multivariate projection-based fixed boundary empirical wavelet transform and entropy features extracted from multichannel eeg signals, *Entropy*, vol. 22, no. 10, pp. 1–23, 2020, ISSN: 10994300. DOI: 10.3390/e22101141.

[34] F. Jager, S. Libenšek, and K. Geršak, Characterization and automatic classification of preterm and term uterine records, *PLoS ONE*, vol. 13, no. 8, O. Uthman, Ed., e0202125, 2018, ISSN: 19326203. DOI: 10.1371/journal.pone.0202125.

[35] Y. Ye-Lin, J. Garcia-Casado, G. Prats-Boluda, J. Alberola-Rubio, and A. Perales, Automatic identification of motion artifacts in EHG recording for robust analysis of uterine contractions, *Computational and Mathematical Methods in Medicine*, vol. 2014, 2014, ISSN: 1748670X. DOI: 10.1155/2014/470786.

[36] M. J. Katz, Fractals and the analysis of waveforms, *Computers in Biology and Medicine*, vol. 18, no. 3, pp. 145–156, 1988, ISSN: 00104825. DOI: 10.1016/0010-4825(88)90041-8.

[37] M. Brennan, M. Palaniswami, and P. Kamen, Do existing measures of Poincareé plot geometry reflect nonlinear features of heart rate variability? *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 11, pp. 1342–1347, 2001, ISSN: 00189294. DOI: 10.1109/10.959330.

[38] J. Xiong, X. Liang, T. Zhu, L. Zhao, J. Li, and C. Liu, A new physically meaningful threshold of sample entropy for detecting cardiovascular diseases, *Entropy*, vol. 21, no. 9, p. 830, 2019, ISSN: 10994300. DOI: 10.3390/e21090830.

[39] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, ISSN: 10769757. DOI: 10.1613/jair.953. arXiv: 1106.1813.

[40] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. 2012, pp. 1–100, ISBN: 978026208029.

[41] W. Bouaguel, A New Approach for Wrapper Feature Selection Using Genetic Algorithm for Big Data, in Intelligent and Evolutionary Systems, 2016, pp. 75–83. DOI: 10.1007/978-3-319-27000-5_6.

[42] O. Babatunde, L. Armstrong, J. Leng, and D. Diepeveen, A Genetic Algorithm-Based Feature Selection, *International Journal of Electronics Communication and Computer Engineering*, vol. 5, no. 4, pp. 899–905, 2014, ISSN: 2278-4209.

[43] D. Beasley, D. R. Bull, and R. R. Martin, An overview of genetic algorithms : Part 1, fundamentals, *University Computing*, vol. 2, no. 15, pp. 1–16, 1993, ISSN: 0265-4385.

[44] D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. 1989, vol. 27, ISBN: 9780201157673.

[45] 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, e77154, 2013, ISSN: 19326203. DOI: 10.1371/journal.pone.0077154.

[46] 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, ISSN: 13616579. DOI: 10.1088/0967-3334/36/2/341.

[47] 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, ISSN: 18728286. DOI: 10.1016/j.neucom.2015.01.107.

[48] 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, e0132116, 2015, ISSN: 19326203. DOI: 10.1371/journal.pone.0132116.

[49] J. Mas-Cabo, G. Prats-Boluda, Y. Ye-Lin, J. Alberola-Rubio, A. Perales, and J. Garcia-Casado, Characterization of the effects of Atosiban on uterine electromyograms recorded in women with threatened preterm labor, *Biomedical Signal Processing and Control*, vol. 52, pp. 198–205, 2019, ISSN: 17468108. DOI: 10.1016/j.bspc.2019.04.001.

[50] H. Azami, M. Rostaghi, A. Fernandez, and J. Escudero, Dispersion entropy for the analysis of resting-state MEG regularity in Alzheimer's disease, in 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2016, pp. 6417–6420, ISBN: 9781457702204. DOI: 10.1109/EMBC.2016.7592197.

[51] S. Simons, P. Espino, and D. Abásolo, Fuzzy Entropy analysis of the electroencephalogram in patients with Alzheimer's disease: Is the method superior to Sample Entropy? *Entropy*, vol. 20, no. 1, 2018, ISSN: 10994300. DOI: 10.3390/e20010021.

[52] G. Doquire and M. Verleysen, A comparison of multivariate mutual information estimators for feature selection, in ICPRAM 2012 - Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods, vol. 1, 2012, pp. 176–185, ISBN: 9789898425980. DOI: 10.5220/0003726101760185.

[53] L. Fang *et al.*, Feature selection method based on mutual information and class separability for dimension reduction in multidimensional time series for clinical data, *Biomedical Signal Processing and Control*, vol. 21, pp. 82–89, 2015, ISSN: 17468108. DOI: 10.1016/j.bspc.2015.05.011.

[54] D. Alamedine, M. Khalil, and C. Marque, Comparison of different EHG feature selection methods for the detection of preterm labor, *Computational and Mathematical Methods in Medicine*, vol. 2013, 2013, ISSN: 17486718. DOI: 10.1155/2013/485684.

[55] S. M. Naeem, A. F. Seddik, and M. A. Eldosoky, New technique based on uterine electromyography nonlinearity for preterm delivery detection, *Journal of Engineering and Technology Research*, vol. 6, no. 7, 107–114, November 2014, 2014, ISSN: 2006-9790. DOI: 10.5897/JETR2013.0332.

[56] M. Hassan, J. Terrien, C. Marque, and B. Karlsson, Comparison between approximate entropy, correntropy and time reversibility: Application to uterine electromyogram signals, *Medical Engineering and Physics*, vol. 33, no. 8, pp. 980–986, 2011, ISSN: 13504533. DOI: 10.1016/j.medengphy.2011.03.010.

[57] F. Nieto-del-Amor *et al.*, Optimized Feature Subset Selection Using Genetic Algorithm for Preterm Labor Prediction Based on Electrohysterography, *Sensors*, vol. 21, no. 10, p. 3350, 2021. DOI: 10.3390/s21103350.

[58] D. Cuesta-Frau and B. Vargas, Permutation entropy and bubble entropy: Possible interactions and synergies between order and sorting relations, *Mathematical Biosciences and Engineering*, vol. 17, no. 2, pp. 1637–1658, 2020, ISSN: 15510018. DOI: 10.3934/mbe.2020086.

[59] P. Li, C. Karmakar, J. Yearwood, S. Venkatesh, M. Palaniswami, and C. Liu, Detection of epileptic seizure based on entropy analysis of short-term EEG, *PLoS ONE*, vol. 13, no. 3, e0193691, 2018, ISSN: 19326203. DOI: 10.1371/journal.pone.0193691.

[60] U. R. Acharya *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, ISSN: 18790534. DOI: 10.1016/j.compbiomed.2017.04.013.

[61] J. Mas-Cabo, G. Prats-Boluda, J. Garcia-Casado, J. Alberola-Rubio, A. Perales, and Y. Ye-Lin, Design and Assessment of a Robust and Generalizable ANN-Based Classifier for the Prediction of Premature Birth by means of Multichannel Electrohysterographic Records, *Journal of Sensors*, vol. 2019, pp. 1–13, 2019, ISSN: 16877268. DOI: 10.1155/2019/5373810.

[62] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, *Electronics (Switzerland)*, vol. 8, no. 8, 2019, ISSN: 20799292. DOI: 10.3390/electronics8080832.

[63] 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, ISSN: 10994300. DOI: 10.3390/e19010002.

Chapter 5

Combination of Feature Selection and Resampling Methods to Predict Preterm Birth Based on Electrohysterographic Signals from Imbalance Data

Félix Nieto-del-Amor[1], Gema Prats-Boluda[1,*], Javier Garcia-Casado[1], Alba Diaz-Martinez[1], Vicente Jose Diago-Almela[2], Rogelio Monfort-Ortiz[2], Dongmei Hao[3] and Yiyao Ye-Lin[1]

[1] Centro de Investigación e Innovación en Bioingeniería, Universitat Politècnica de València, 46022 Valencia, Spain; feniede@ci2b.upv.es; jgarciac@ci2b.upv.es; adiaz@ci2b.upv.es; yiye@ci2b.upv.es

[2] Servicio de Obstetricia, H.U.P. La Fe, 46026 Valencia, Spain; diago_vicalm@gva.es; monfort_isaort@gva.es

[3] Faculty of Environment and Life, Beijing University of Technology, Beijing International Science and Technology Cooperation Base for Intelligent Physiological Measurement and Clinical Transformation, Beijing 100124, China; haodongmei@bjut.edu.cn

* Correspondence: gprats@ci2b.upv.es

Abstract

Due to its high sensitivity, electrohysterography (EHG) has emerged as an alternative technique for predicting preterm labor. The main obstacle in designing preterm labor prediction models is the inherent preterm/term imbalance ratio, which can give rise to relatively low performance. Numerous studies obtained promising preterm labor prediction results using the synthetic minority oversampling technique. However, these studies generally overestimate mathematical models' real generalization capacity by generating synthetic data before splitting the dataset, leaking information between the training and testing partitions and thus reducing the complexity of the classification task. In this work, we analyzed the effect of combining feature selection and resampling methods to overcome the class imbalance problem for predicting preterm labor by EHG. We assessed undersampling, oversampling, and hybrid methods applied to the training and validation dataset during feature selection by genetic algorithm, and analyzed the resampling effect on training data after obtaining the optimized feature subset. The best strategy consisted of undersampling the majority class of the validation dataset to 1:1 during feature selection, without subsequent resampling of the training data, achieving an AUC of 94.5 ±4.6%, average precision of 84.5 ±11.7%, maximum F1-score of 79.6 ±13.8%, and recall of 89.8 ±12.1%. Our results outperformed the techniques currently used in clinical practice, suggesting the EHG could be used to predict preterm labor in clinics.

Keywords: Genetic algorithm; Imbalance data learning; Electrohysterography; Preterm labor prediction; Resampling methods; Uterine electromyography; Machine learning

5.1 Introduction

5.1.1 Preterm Labor

The World Health Organization defines preterm labor (prevalent in more than 11% of total births) as labor before 37 completed weeks of gestation [1]. It is the leading cause of death in children, accounting for approximately 35% of newborn deaths and 16% of children under five years of age [2]. In the case of survivors, shorter term consequences involve respiratory difficulties, sepsis, neurological conditions, feeding difficulties, as well as visual and hearing problems [3]. Long-term complications include poorer neurodevelopmental outcomes, higher rates of hospital admissions, as well as behavioral, social–emotional, and learning difficulties in childhood [2]. As the average cost of preterm birth is 5–10 times higher than a term birth, preterm birth also has a significant economic impact on public health systems, the average cost of preterm birth is 5–10 times higher than a term birth [4], with an average cost of 64,815 USD per premature baby [5]. For an extremely preterm baby born before 28 weeks of gestation the average cost per baby amounts to 74,009 USD for the first year of life in Germany [6].

Various methods are currently used to predict preterm labor in clinical practice, including: uterine dynamics monitoring by tocodynamometry, cervix length, Bishop score, and bio-chemical markers [7] such as fetal fibronectin and interleukin 6 [8]. None of these techniques can precisely predict true preterm labor, and their clinical values mainly lies in their negative predictive value thanks to their ability to identify patients who are not at risk of preterm labor [7]. Due to its high sensitivity, electrohysterography (EHG) is emerging as a promising technique to identify the risk of preterm birth [9]. This non-invasive technique records the electrical activity of billions of uterine myometrial cells on the maternal abdominal wall.

5.1.2 Electrohysterography for Preterm Labor Prediction

Previous studies showed that the EHG signal distributes its energy within 0.1–4 Hz and is made up of two components: fast wave low (0.2–0.34 Hz), which has been associated with signal propagation, and fast wave high (0.34–4 Hz), which is related to cell excitability [9, 10]. Since EHG mainly distributes its energy below 1 Hz, many authors preferred to analyze the signal within the 0.34–1 Hz range to minimize respiratory and cardiac interference [11]. Uterine myometrial cell excitability and bioelectric propagability rise due to progressive formation of gap-junctions, which end up leading to coordinated high-intensity contractions that give rise to labor.

A set of temporal, spectral, and non-linear parameters have been proposed in the literature to characterize these electrophysiological changes. As pregnancy progresses, EHG amplitude increases and is associated with a larger number of uterine cells involved in the contractions [9]. The EHG signal spectral content also shifts towards higher frequencies as delivery approaches, suggesting increased cell excitability [9, 12]. Previous studies found increased signal regularity, thus reduced complexity, by analyzing Lempel-Ziv and different entropy measures [10, 13, 14, 15, 16, 17], although controversial results were obtained due to the limited database with different compositions depending on the inclusion criteria and the analysis bandwidth, among others. Time reversibility and Poincaré plot-derived parameters were also used for characterizing the EHG signal [13, 14, 18], with an increased signal non-linearity degree and less randomness as pregnancy progresses.

The latest research studies focused on the development of preterm birth prediction systems and have obtained promising results, with an accuracy of more than 90% [11, 15, 19, 20, 21]. However, they have not had a significant impact on clinical praxis. Firstly, most preterm labor prediction systems used complex classifiers that involve the non-linear transformation of input features into higher dimension space to better separate the target classes [22]. Obstetricians find the prediction results difficult to interpret and hard to trust, since these algorithms achieve good performance even when the input features are highly overlapped between the target classes [19, 23]. In this regard, we have shown the feasibility of predicting preterm labor with the synthetic minority oversampling technique (SMOTE) on a balanced dataset using simple classification algorithms such as the K-nearest-neighbor, logistic regression, and linear discriminant analysis by feature subspace optimization using a genetic algorithm [15, 19]. Secondly, due to the highly imbalanced data between the two target classes (11% preterm labor vs. 89% term labor), conventional classification algorithms are often biased towards the majority class and fail to correctly identify the minority class, obtaining a higher misclassification rate of true preterm labor in predicting premature deliveries [21, 24, 25, 26]. This phenomenon is due to the fact that conventional machine learning algorithms are designed to optimize the overall performance (accuracy) instead of considering the predictive capability of each class [27]. The majority class data are relatively excessively distributed than the minority class data, thus invading the minority class area and hindering the correct setting of the decision boundary [25].

5.1.3 Resampling Methods for Imbalance Data Learning

Rebalancing to equal the distribution of data classes is a commonly used strategy to mitigate the above imbalanced learning problems. Most previous studies used SMOTE, which consisted of synthesizing new samples by interpolating the original minority class observations [28], achieving promising results [11, 15, 19, 26, 29, 30, 31, 32]. Nevertheless, according to a recent study [26], these works may overestimate preterm labor prediction performance due to their methodological bias. Application of the SMOTE technique prior to data partition would give rise to the data structure correlation between training and test dataset, and tends to overestimate the real generalization capacity of the model [26]. In fact, Vandewiele et al. attempted to reproduce the preterm labor prediction system method of 11 published studies and analyzed the model's performance difference between applying SMOTE before and after data partition [26]. When balancing data before partition, they obtained an AUC ranging from 85% to 99% which was very close to the reported evaluation metrics. In contrast, when applying SMOTE to training data after partitioning, prediction performance decreased drastically, with an AUC below 65% using the same input features and classification algorithms [26]. Due to the underlying assumption of the homogeneity of the clusters of minority observations, SMOTE can inappropriately alter the class distribution when factors such as disjoint data distributions, noise, and outliers are present [33]. In addition to the SMOTE technique, other resampling methods have also been proposed to mitigate the imbalanced data problem, including undersampling and oversampling/undersampling hybrid methods [21]. Undersampling is a non-heuristic method that consists of removing instances from the majority class to alleviate the skewed class distribution problem. This latter is limited to a moderate or low imbalanced dataset and is not recommended for highly imbalanced datasets because of its high potential of underfitting due to information loss [34]. If the size of the minority class sample is small the classifier performance may be greatly impaired [34]. However, other authors have proposed hybrid oversampling/undersampling methods to reduce the class overlap problem, which usually consists of cleaning the majority class observations in proximity to the minority instances by the undersampling method before or after SMOTE [35, 36, 37, 38].

Studies in different application areas have attempted to determine the optimal resampling method from a database set with variable numbers and/or type characteristics [34, 39, 40]. Napierala & Stefanowski studied types of minority class distribution in real imbalanced datasets and their influence on learning classifiers [39]. Zhou analyzed the effect of sampling methods on the performance of quantitative bankruptcy prediction models on real highly imbalanced dataset and confirmed that the proper sampling method in developing prediction models mainly depended on the size of the training sample [40]. With hundreds of minority observations in the dataset, the undersampling was superior to the oversampling method in terms of computation time, although SMOTE was found to be a better choice with only a few dozen minority instances. A combination of SMOTE and undersampling could be a good alternative for a large training sample [40]. Loyola-González et al. analyzed the impact of resampling methods for contrast pattern-based classifiers on imbalanced databases and provided a guide for the selection of the resampling method regarding the class imbalance ratio [34]. Despite these previous studies, no resampling method always outperforms the others [41]. It is difficult to determine a specific optimal rate of undersampling or oversampling which always leads to better results for a specific application [41].

Other authors have proposed combining feature selection, resampling, and ensemble learning to deal with multiclass imbalanced data learning, and obtained results that outperformed or were comparable to several state-of-the art algorithms [42]. In the classification task, high-dimensional features may lead to overfitting, which can limit the model's generalization capability [43]. Removing irrelevant features may reduce the noise information in the training space and also model complexity and training time. In imbalanced scenarios, high-dimensionality could have a greater impact; as minority class samples can easily be discarded as noise [42], eliminating irrelevant features may also reduce the risk of treating the minority class as noise. High-dimensionality can even lead to class overlapping, which makes the design of discriminative rules extremely difficult in imbalanced data scenarios [44]. Ramos-Pérez et al. analyzed the combination effects of resampling and feature selection techniques on high-dimensional and low instance imbalanced data, also determining whether resample data should be before or after feature selection [45]. The contribution of feature selection to specific preterm labor prediction from imbalanced data remains unclear.

The aim of this work was to determine the effect of combining feature selection and resampling methods on preterm labor prediction from imbalanced data. We first confirmed that the application of resampling methods before data partition considerably reduced the complexity of the classification task. We showed the feasibility of combining both the feature selection using genetic algorithm and resample methods in the same iterative process to deal with imbalanced data, in contrast to resampling before or after feature selection. Our results suggested that undersampling the validation set turned out to be the best strategy for preterm labor prediction in an imbalanced scenario, achieving a recall ranging from 79.6% to 89.8%, which is considerably higher than the techniques commonly used in clinical practice and also than the unbiased preterm labor prediction performance reported by Vandewiele et al.

5.2 Materials and Methods

5.2.1 Database Description

300 EHG records from "Term-Preterm EHG Database" (TPEHG DB) [10] and 26 EHG records from "The Term-Preterm EHG Dataset with tocogram" (TPEHGT DS)

[46] obtained between 22 and 37 weeks of gestation were analyzed in the study. This ensemble database was highly imbalanced in terms of preterm labor: 275 term labor (84%) vs. 51 preterm (16%). Both datasets used the same recording protocol, which consisted of placing four electrodes (E1, E2, E3 and E4) on the abdomen to obtain three bipolar channels (S1, S2 and S3), with a pairwise distance of 7 cm. All the signals were sampled at 20 Hz and then pre-processed by band-pass filtering between 0.1 and 4 Hz using a fifth-order digital zero-phase Butterworth filter (see Figure 5.1). We also used obstetric data available from both databases, such as maternal age, parity, number of previous abortions, maternal weight and weeks of gestation on recording.

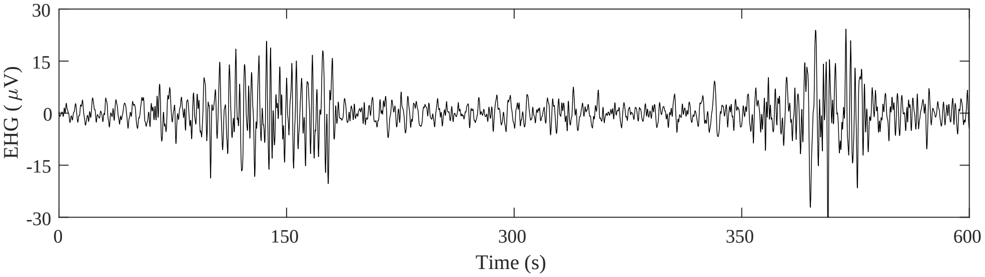


Figure 5.1: Example of preprocessed EHG signal recorded from women with 30 weeks of gestation who finally delivered at preterm. Two EHG-bursts associated with uterine contraction can be clearly seen (around 150 s and 400 s) with increased amplitude and frequency contents with respect to basal activity when the uterus is at rest.

5.2.2 EHG Signal Analysis

As EHG signal recordings may not only contain uterine myoelectrical activity, but also corrupt segments such as motion-artifacts and respiratory interference, EHG records were reviewed by two experts in a double-blind process to remove all the corrupted signal segments. A whole windows analysis with sliding windows of 120 s length and 50% overlap was then performed to characterize the EHG recordings [10, 13, 14], and proved to be a good trade-off between computational cost and information loss [47]. This type of analysis was able to identify relevant information in the EHG signal without identifying EHG-bursts associated with uterine contractions [47], which could be very challenging in EHG records taken far from delivery. After obtaining all the features of the analysis windows of a whole recording, we computed the median value as the representative data of this process.

We used a widely used set of temporal, spectral, and non-linear parameters for EHG signal characterization. First, we calculated EHG signal peak-to-peak amplitude (App) in the following four bandwidths: 0.1–4 Hz, 0.2–0.34 Hz, 0.34–4 Hz and 0.34–1 Hz. Since EHG spectral content mainly distributed its energy in the 0.2–1 Hz bandwidth, we estimated dominant frequency DF1 in the range 0.2–1 Hz, DF2 in 0.34–1 Hz, normalized sub-band energy (NormEn) (0.2–0.34 Hz, 0.34–0.6 Hz and 0.6–1 Hz) and high (0.34–1 Hz)-to low (0.2–0.34 Hz) frequency energy ratio (H/L Ratio). We also calculated mean frequency (App), power spectrum deciles (D1, . . . , D9), Teager energy and spectral moment ratio (SpMR) in 0.2–1 Hz. Likewise, we computed the following parameters to quantify the non-linear degree, signal complexity and regularity: binary and multistate Lempel-Ziv index (LZBin and LZMulti n = 6), time reversibility (TimeRev), Katz fractal dimension (KFD), Poincaré ellipse metrics (minor axis (SD1), major axis (SD2), square root of variance (SDRR, (SD12+SD22)/2) and SD1/SD2 ratio), sample entropy (SampEn), fuzzy entropy (FuzEn), spectral entropy (SpEn), dispersion entropy (DispEn), and bubble entropy (BubbEn) [15]. Since non-linear parameters estimated from different bandwidths may contain complementary information for predicting preterm labor [14], we computed the non-linear parameters in the same four bandwidths as the signal amplitude. In total, each record was characterized by a set of 222 EHG features ((4 temporal, 18 spectral, and 52 non-linear parameters per channel) × 3 channels = 222) and the 5 obstetric patient data. Table 5.1 summarizes all the parameters described in this section, which constituted the input features of the preterm labor prediction system.

5.2.3 Classifier Design and Evaluation

Our specific application was first characterized by a total of 227 high-dimensional input features, with few and imbalanced sample data between the target classes (326 EHG records, with an imbalanced ratio of 51/275 preterm/term cases). We used the conventional holdout method (200 partitions) to design and validate the classifier. For each partition, the whole imbalanced database was randomly split into training (80%) and testing (20%), preserving the skewness between the preterm and term classes (preterm/term samples = 51/275). The training partition was then further split into training (64%) and validation datasets (16%). As mentioned above, we attempted to evaluate the effect of combining feature selection and resampling methods for predicting preterm labor in an imbalanced scenario. As there is still no general agreement in the literature as to which strategy with imbalanced data obtains the best performance, we compared the different strategies by balancing training or validation data using the following resampling methods: oversampling (SMOTE, k = 5), undersampling, and over/undersampling hybrid, the preterm/term instance ratio after data balancing being 1:1. We used the neighborhood cleaning rule (NCL) for the undersampling method; this uses Wilson's edited nearest neighbor rule to remove noise instances, as it identifies the boundary samples to the decision boundary to avoid overfitting [48].

Step 1: effect of resampling strategy for feature selection. We used the genetic algorithm to optimize feature subspace, which has been proven to successfully preserve complementary information for predicting preterm labor in the SMOTE balanced database, while discarding redundant, irrelevant and noise information [15, 19]. This

Table 5.1: EHG features and obstetrical data included as input data to the classifier to discriminate preterm from term deliveries. The number of features per channel depends on the frequency band- widths and were computed: 0.1–4 Hz, 0.2–0.34 Hz, 0.34–4 Hz and 0.34–1 Hz for temporal and non-linear features, and 0.2–1 Hz for spectral features. Considering: peak to peak amplitude (APP), dominant frequency (DF1) in the range 0.2–1 Hz, (DF2) in 0.34–1 Hz, normalized sub-band energy (NormEn) (0.2–0.34 Hz, 0.34–0.6 Hz and 0.6–1 Hz) and high (0.34–1 Hz)-to low (0.2–0.34 Hz) frequency energy ratio (H/L ratio), power spectrum deciles (D1, . . . , D9), spectral moment ratio (SpMR), binary and multistate Lempel-Ziv index (LZBin and LZMulti), time reversibility (TimeRev), Katz fractal dimension (KFD), Poincaré ellipse metrics (minor axis (SD1), major axis (SD2), square root of variance (SDRR, $\sqrt{(\mathrm{SD1}^2 + \mathrm{SD2}^2)/2}$ and SD1/SD2 ratio), sample entropy (SampEn), fuzzy entropy (FuzEn), spectral entropy (SpEn), dispersion entropy (DispEn), and bubble entropy (BubbEn).

EHG Temporal Features	**EHG Spectral Features**	**EHG Non-Linear Features**	**EHG Obstetrical Data**
4 per channel	18 per channel	52 per channel	5
App	MeanF DF1, DF2 NormEn H/L Ratio [D1–D9] Teager Energy SpecMR	LZBin LZMulti (n = 6) TimeRev KFD SD1 SD2 SDRR SD1/SD2 SampEn FuzEn SpEn DispEn BubbEn	Maternal age Parity Abortions Weight Week of gestation (WOG)

Table 5.2: Configuration parameters used in genetic algorithm.

Parameter	**Value**	**Parameter**	**Value**
Population size	N=222	Mutation	Uniform
Genome length	N=222	Mutation Probability	0.01
Number of generations	500	Selection scheme	Tournament of size 2
Crossover	Arithmetic	Elite count	2
Crossover Probability	0.8	Termination condition	No fitness function improvement for 150 consecutive iterations (differential tolerance: 10^{-6})

algorithm (GA) is an optimization technique, a population-based heuristic search method that simulates the natural evolutionary process. It is an iterative procedure that manipulates a population of chromosomes (solution candidates) to produce a new population through genetic functions such as crossing over and mutation. These algorithms have been shown to be able to escape from local minima to reach global minima in complex functions [49]. We used the same GA configuration parameters as in our previous studies (see Table 5.2) [15, 19].

As for the classification method, in this work we used the simple easily interpreted linear discrimination analysis (LDA) to discriminate the target classes, which has obtained good results for predicting preterm birth in previous works [15, 19]. The mathematical formulation of LDA classification methods can be found in previous works [22].

All the chromosomes in the total population were evaluated to determine model goodness by the fitness function, which we defined as the mean F1-score of the 200 validation datasets weighted by the number of features used in each iteration [15, 19, 49]. This was used in preference to accuracy, since the F1-score is the geometric mean of precision and recall and obtains the correct classification of the preterm observation, without ignoring term observations.

$$\text{Fitness function} = \text{mean}\left\{\text{F1-score} \times (\text{NFeat} - \text{NCFeat})\right\} \tag{5.1}$$

where NFeat and NCFeat are the number of features in the initial set and the current subset, respectively. The six best chromosomes which optimized feature subsets were thus obtained by considering the following assumptions: resampling the training partition by oversampling (FS_{TO}), undersampling (FS_{TU}), or under/oversampling hybrid (FS_{TH}) method; resampling the validation partition by oversampling (FS_{VO}), undersampling (FS_{VU}), or under/oversampling (FS_{VH}) hybrid method. Figure 5.2 shows the flowchart that assesses the effect of combining feature selection by the genetic algorithm and the different resampling methods for imbalanced data learning.

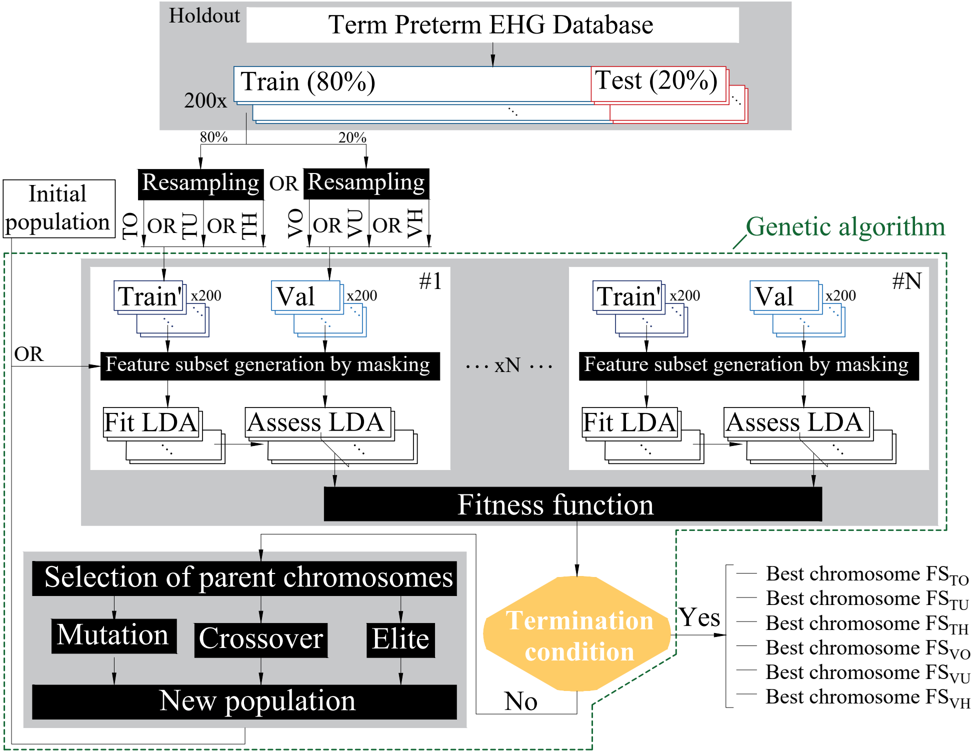

Holdout
Term Preterm EHG Database
Train (80%)
Test (20%)
200x
80%
20%
Initial population
Resampling
OR
Resampling
TO
OR
TU
OR
TH
VO
OR
VU
OR
VH
Genetic algorithm
#1
Train'
x200
Val
x200
OR
Feature subset generation by masking
Fit LDA
Assess LDA
··· xN ···
#N
Train'
x200
Val
x200
Feature subset generation by masking
Fit LDA
Assess LDA
Fitness function
Selection of parent chromosomes
Mutation
Crossover
Elite
New population
Termination condition
No
Yes
Best chromosome FS$_{TO}$
Best chromosome FS$_{TU}$
Best chromosome FS$_{TH}$
Best chromosome FS$_{VO}$
Best chromosome FS$_{VU}$
Best chromosome FS$_{VH}$

Figure 5.2: Flowchart to assess the effect of combining feature selection by the genetic algorithm and resampling methods to deal with the imbalanced data problem. The training or validation partitions are resampled by oversampling (TO, VO), undersampling (TU, VU), or applying hybrid methods (TH, VH). The initial population of N chromosomes masks the training and validation partitions. For each chromosome, LDA classifiers are trained and evaluated with the respective validation partitions by its fitness function. A new population of chromosomes is generated from the processes of mutation, crossing over, and selection of the elite chromosomes from the previous iteration until the termination condition was satisfied, obtaining its corresponding best chromosome.

Table 5.3: Resampling method for predicting preterm labor.

Approach	Resampling Technique	Model
No resampling	Not applicable	RN
Oversampling	SMOTE	RO
Undersampling	NeighborhoodCleaningRule	RU
Hybrid	SMOTE + NeighborhoodCleaningRule	RH

Step 2: effect of resampling strategy for training the prediction model. For each optimized feature subset, we further assessed the influence of the different resampling methods (RN, RO, RU and RH, see Table 5.3) applied to the total of 80% of training dataset (see Figure 5.3). Each training and test partition was masked by the optimized feature subset $\mathrm{FS}_{\mathrm{TO}}$, $\mathrm{FS}_{\mathrm{TU}}$, $\mathrm{FS}_{\mathrm{TH}}$, $\mathrm{FS}_{\mathrm{VO}}$, $\mathrm{FS}_{\mathrm{VU}}$ or $\mathrm{FS}_{\mathrm{VH}}$ or not (all features, AF). We then trained the LDA classifier using the resampled training partition and evaluated its average performance for the testing dataset, which represents the new incoming data never seen by the model and could be used to determine the real model generalization capability, using two threshold independent metrics to evaluate the model performance: the area under the $\mathrm{LNLEn}_{\mathrm{ALL}}$ Curve (AUC) and average precision (AP). This was because the threshold-dependent metrics have been shown to be biased towards the majority class in an imbalanced scenario, whereas AUC and AP avoid this bias [50]. AUC and AP are mathematically formulated in Equations (5.2) and (5.3).

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR}) \times \mathrm{dFPR} \tag{5.2}$$

$$\mathrm{AP} = \sum_n (\mathrm{R}_n - \mathrm{R}_{n-1}) \times \mathrm{P}_n \tag{5.3}$$

where TPR and FPR are true positive rate and false positive rate, and P_n and R_n are the precision and recall at the nth threshold.

We then analyzed the statistically significant difference between the different model performances to determine the best strategy to achieve the highest average AUC and AP scores ((AUC + AP)/2) for the testing dataset. We first confirmed the normal data distribution (D'Agostino's k-squared test [51]) for both AUC and AP scores of the 200 partitions for each combination of feature subset and resampling method. Then we assessed the statistically significant difference of the (AUC + AP)/2 between different resampling methods for each feature subset by one-way analysis of variance with repeated measures (RANOVA, $\alpha = 0.05$) followed by Tukey's multiple comparison test and evaluated the statistically significant difference of the (AUC + AP)/2 between the different feature subsets for all the resampling methods (RN + RO + RU + RH) by the same statistical method ($\alpha = 0.05$).

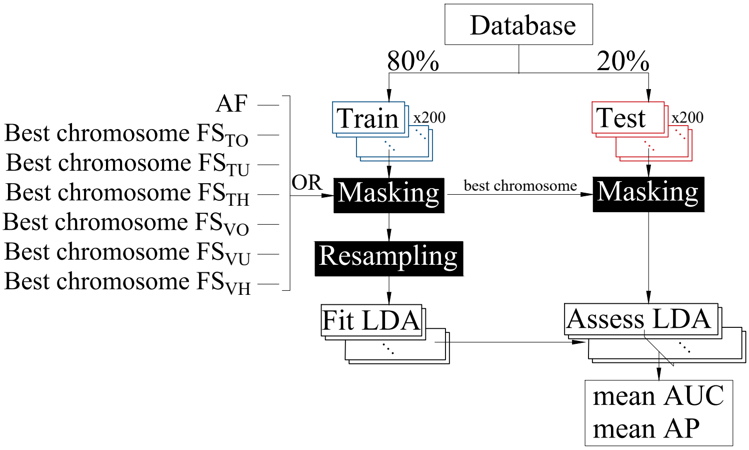

Database
80%
20%
AF
Best chromosome FS$_{TO}$
Best chromosome FS$_{TU}$
Best chromosome FS$_{TH}$
Best chromosome FS$_{VO}$
Best chromosome FS$_{VU}$
Best chromosome FS$_{VH}$
OR
Train
x200
Test
x200
Masking
best chromosome
Masking
Resampling
Fit LDA
Assess LDA
mean AUC
mean AP

Figure 5.3: Flow diagram of the training process and evaluation of the prediction models.

Step 3. Effect of imbalance ratio for feature extraction. We also assessed the influence of the post-resampling preterm/term instance ratio (imbalance ratio) for the best strategy of steps 1 and 2 (resampling methods for feature subset and training of prediction model). The process shown in Figure 5.2 was again used to obtain nine best chromosomes with an imbalance ratio of from 20 to 100% with a 10% step. We determined the statistically significant differences of the model performances between the different imbalanced ratios using the same statistical method ($\alpha = 0.05$).

Finally, for the best strategy, i.e., the best (AUC + AP)/2, we determined the threshold-dependent scores of the test partitions for the operative point that maximizes the F1-score and G-mean: F1-score, G-mean, precision, recall, and specificity. Recall metric denotes the true preterm birth predicted by the algorithm with respect to the total of preterm labor women in the testing partition. Precision represents the true preterm birth with respect to the total preterm birth predicted by the algorithm. Specificity refers to the true negative rates over the total negative cases predicted by the algorithm. F1-score is the harmonic average of recall and precision, which is a trade-off between false positives and false negatives. G-mean was defined as the geometric average of recall and specificity [52]. All these metrics were mathematically formulated in the Equations (5.4)–(5.8) [53].

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.4}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{5.5}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5.6}$$

$$\text{F1-score} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \tag{5.7}$$

$$\text{G-mean} = \sqrt{\text{recall} \times \text{specificity}} \tag{5.8}$$

where TP is the true positive, TN is the true negative, FP is the false positive, and FN is the false negative.

5.3 Results

Table 5.4 shows the average AUC and AP scores for the testing dataset to predict preterm labor in an imbalanced scenario using each combination of the resampling method for the feature subset and training of the prediction model. Figure 5.4 shows the violin plot of the score (AUC + AP)/2 for the four resampling methods of each set of input features. The average values of AUC, AP and (AUC + AP)/2 are also shown in this figure. When using all features (AF) for designing the model, SMOTE (RO) did not enhance the prediction capacity of the base classifier with AUC~52% and AP~21%. Both the undersampling (RU) and hybrid (RH) methods performed significantly better, achieving an AUC of ~65% and AP of ~12%. When using FS_{TO}, FS_{TU} and FS_{TH} as input features, different resampling methods yield similar performance with no significant difference. For FS_{VO}, FS_{VU} and FS_{VH}, the no resampling (RN) and oversampling (RO) versions performed significantly better than the undersampling and hybrid versions. When using the optimized feature subset achieved by the genetic algorithm (FS_{TO}, FS_{TU}, FS_{TH}, FS_{VO}, FS_{VU} and FS_{VH}), none of the resampling methods proposed for the training dataset of the models significantly improved the model performance without additional resampling (RO, RU, RH vs. RN).

The different optimized feature subsets obtained by the genetic algorithm significantly improved the mean score of AUC and AP over AF. Undersampling or hybrid methods during feature selection achieved significantly higher mean AUC and AP scores than those obtained by the oversampling method when used in the training or validation subsets ($\text{FS}_{\text{TU}} \approx \text{FS}_{\text{TH}} > \text{FS}_{\text{TO}}$, and $\text{FS}_{\text{VU}} \approx \text{FS}_{\text{VH}} > \text{FS}_{\text{VO}}$). Regarding whether to balance training or validation datasets during feature selection, better performance metrics were obtained for the latter in all cases (except for $\text{AP}(\text{FS}_{\text{TO}})$ vs. $\text{AP}(\text{FS}_{\text{VO}})$). Undersampling the validation dataset significantly outperformed the rest ($\text{FS}_{\text{VU}} > \text{FS}_{\text{VH}} > \text{FS}_{\text{VO}}$). Our results showed that the best preterm labor prediction strategy in an imbalanced scenario was undersampling the validation dataset for feature selection, with no further resampling method (base-classifier RN of FS_{VU}).

We also evaluated the effect on the model performance of post-resampling the imbalance ratio of the validation dataset. Table 5.5 shows AUC and AP values for the optimized features subset achieved using different validation dataset ratios. Besides, Figure 5.5 shows violin plots of the score (AUC + AP)/2. Ratios of from 20% to 40%

Table 5.4: AUC and AP scores for the testing datasets for each feature subset and resampling method. Homogenous groups of resampling method with similar performance with no statistically significant differences are shown in different shades of grey. Bottom row shows the average value of the 4 resampling methods for each feature subset.

		AF	**FS$_{\mathrm{TO}}$**	**FS$_{\mathrm{TU}}$**	**FS$_{\mathrm{TH}}$**	**FS$_{\mathrm{VO}}$**	**FS$_{\mathrm{VU}}$**	**FS$_{\mathrm{VH}}$**
AUC (%)	**RN**	52.1 ±12.3	86.7 ±8.2	90.3 ±6.6	89.9 ±6.6	88.8 ±5.5	94.5 ±4.6	93.5 ±4.4
	RO	52.8 ±12.2	86.5 ±8.0	90.6 ±6.0	90.2 ±6.4	89 ±5.8	93.7 ±4.8	92.5 ±5.2
	RU	65.6 ±11.7	86.7 ±8.3	90.9 ±6.3	89.2 ±7.2	85.8 ±6.9	92.9 ±5.3	91.2 ±5.5
	RH	65.1 ±12.2	85.9 ±7.9	91.5 ±5.3	89.9 ±6.93	87.4 ±6.2	92.3 ±5.7	89.9 ±6.3
	mean	*59.2 ±13.9*	*86.5 ±8.1*	*90.8 ±6.1*	*89.9 ±6.8*	*88.2 ±5.9*	*93.4 ±5.2*	*91.8 ±5.5*
		AF	**FS$_{\mathrm{TO}}$**	**FS$_{\mathrm{TU}}$**	**FS$_{\mathrm{TH}}$**	**FS$_{\mathrm{VO}}$**	**FS$_{\mathrm{VU}}$**	**FS$_{\mathrm{VH}}$**
AP (%)	**RN**	22.9 ±8.9	66.5 ±16.1	70.3 ±15.4	70.6 ±14.5	63.6 ±14.5	84.8 ±11.7	77.8 ±14.4
	RO	21.6 ±7.2	65.3 ±15.8	67.1 ±15.6	70.2 ±15	65.4 ±14.7	82.9 ±11.9	75.6 ±14.3
	RU	36.7 ±15.5	66.7 ±15.7	69.1 ±15.5	69.6 ±15.9	57.4 ±14.9	78.4 ±14.1	71.5 ±15.6
	RH	36.7 ±14.7	64.9 ±15	67.3 ±15.0	70.7 ±14.9	60.6 ±15.1	77.7 ±14.4	69.1 ±15.9
	mean	*29.9 ±14.5*	*66 ±15.8*	*68.5 ±15.3*	*71.0 ±15.0*	*62.7 ±14.9*	*81.0 ±13.4*	*73.5 ±15.4*

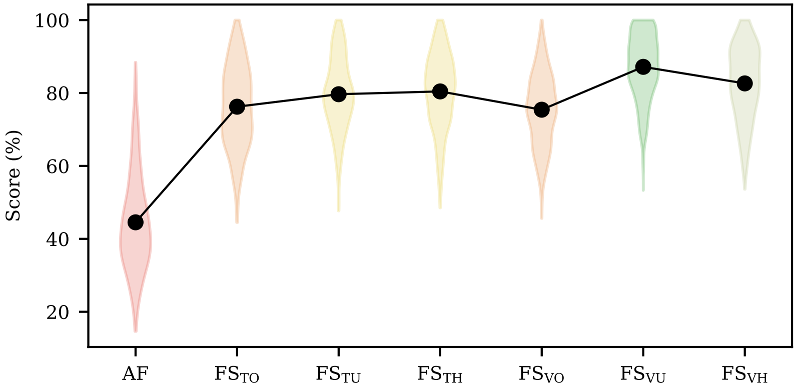


Figure 5.4: Violin plots represent the distribution of (AUC + AP)/2 for the four resampling methods for each set of input features and average value of (AUC + AP)/2 in black line. Violin color represents homogenous group with similar performance without significant difference (p-value > 0.05).

performed significantly worse than the other imbalance ratios. The model performance increased from an imbalance ratio of 50%, with the best result achieved when the validation partition was totally balanced. The statistical analysis showed that imbalance ratios of 90% and 100% significantly outperformed those of 50–80%, with no statistically significant differences between them. The number of features included in each best chromosome was 30, 44, 35, 34, 46, 57, 55, 59, and 58 for imbalance ratios of from 20% to 100%, respectively.

Figure 5.6 shows the average LNLEn$_{\mathrm{ALL}}$ and precision-recall curve for the best strategy to deal with the imbalanced data problem (FS$_{\mathrm{VU}}$, imbalance ratio 100% and no resampling method) for the testing dataset. Table 5.6 shows the threshold-dependent scores for the test partitions for the operative point that maximizes the F1-score (threshold = 0.85) and G-mean (threshold = 0.01) shown in Figure 5.5. The maximum F1-score for preterm labor prediction in an imbalanced scenario was 79.6 ±13.8%, with a recall of 79.6 ±17.4% and precision of 81.9 ±14.9% for testing dataset. By maximizing the G-mean we can further improve the recall score to 89.8 ±12.1% with a specificity of 94 ±5.4%.

Table 5.5: AUC and AP scores for the testing datasets for optimum feature subset obtained from undersampling validation partition with different imbalance ratios.

Imbalance ratio (%)	**20**	**30**	**40**	**50**	**60**
AUC (%)	86.9 ±8	88.9 ±7.4	86.5 ±8.5	89.2 ±6.9	90.4 ±7.2
AP (%)	70.7 ±15.1	71.3 ±15.1	72.1 ±14.6	72.7 ±14.9	81.6 ±13
Imbalance ratio (%)	**70**	**80**	**90**	**100**	**-**
AUC (%)	88.5 ±8.6	91.4 ±5.7	92.5 ±5.7	94.5 ±4.6	-
AP (%)	76 ±14.2	75.5 ±14.5	81.2 ±12.7	84.8 ±11.7	-

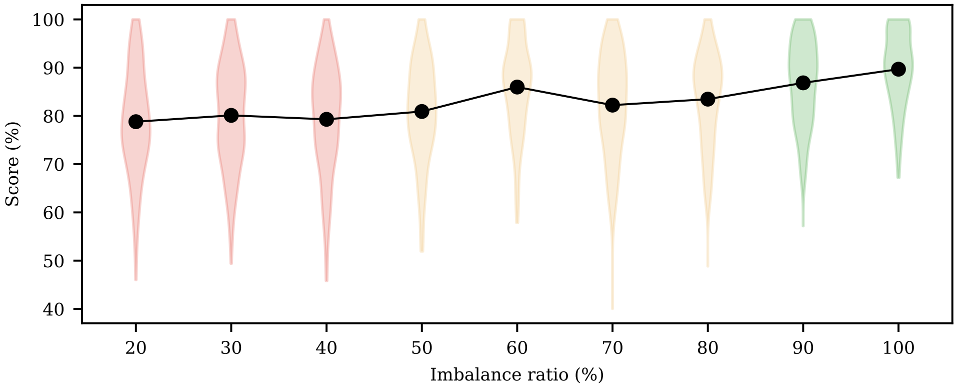


Figure 5.5: (AUC + AP)/2 distribution of testing partition for optimum feature subset obtained from undersampling validation partition with different imbalance ratios and the average value of (AUC + AP)/2 (black line). Violin colors represent homogenous groups with similar performance and no significant differences (p-value > 0.05).

Table 5.6: Threshold-dependent metrics for the best model (FS$_{\text{VU}}$, imbalance ratio 100% and no resampling method).

Maximizing criteria	**F1-score (%)**	**G-mean (%)**	**Precision (%)**	**Recall (%)**	**Specificity (%)**
F1-score	79.6 ±13.8	87.7 ±10.1	81.9 ±14.9	79.6 ±17.4	97.9 ±2.7
G-mean	71.5 ±17.8	91.6 ±6.7	61.4 ±21.5	89.8 ±12.1	94.0 ±5.4

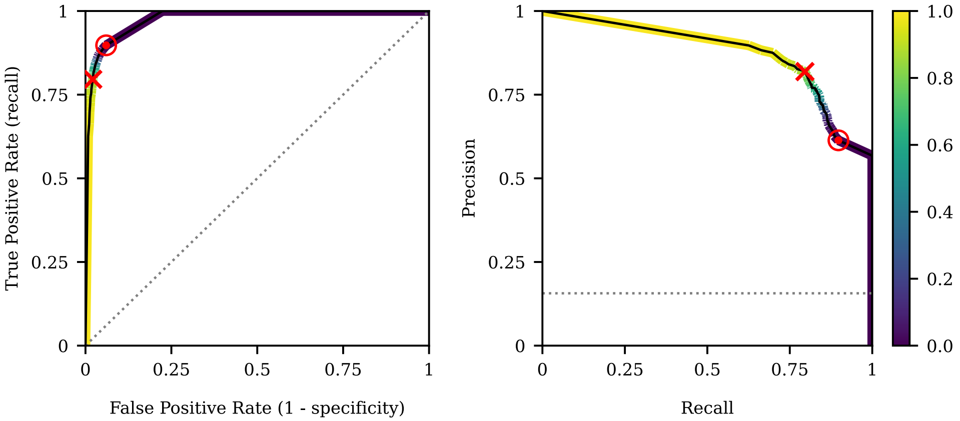


Figure 5.6: Average ROC curve (left) and precision-recall curve (right) of the testing dataset for the best combination of feature subset and resampling method (FSVU, imbalance ratio 100% and no resampling). The red "×" and "⊙" markers show the operative points that maximize the F1-score and G-mean, respectively. The threshold level was shown for each point of the color curves (more blue means closer to 0 and more yellow closer to 1). The dotted lines in the graphs represent the ROC baseline and precision-recall curves (random classifier).

5.4 Discussion

5.4.1 Imbalanced Data Learning

This paper describes different resampling methods for dealing with the imbalanced class problem to predict preterm labor from EHG records and obstetrical data and identified their realistic generalization capability for new incoming data. To avoid data structure correlation by oversampling the whole database before data partition, this was carried out before resampling. Regardless of the resampling method, we found that poor results were obtained when using all input features due to high dimensionality, achieving an AUC of less than 65% and AP below 40%. This result may suggest the existence of noise information that could give rise to high data overlapping between the target classes. These results were comparable with those obtained by Vandewiele et al., who obtained an AUC<65% applying SMOTE after data partition without optimizing the feature subspace [26]. Other authors found that oversampling before data partition significantly reduced the classification task complexity [54], i.e., training and testing data have a similar and correlated data structure, overestimating the model's generalization capability [26, 54]. Indeed, we found the classification task complexity considerably increased with respect to oversampling before data partition. In fact, in the present work the optimum feature set (FS_{VU} and no resampling method) for preterm labor prediction was compounded by 58 features, which were much more than the 12-feature subset using the SMOTE balanced dataset before partition [15].

Regardless of the resampling method applied to the training or validation data, the feature optimization of the subspace by the genetic algorithm may reduce the overlapping data between the target classes and classification task complexity [55, 56], thus significantly increasing both AUC and AP. Our results revealed the importance of feature quality in correctly discriminating target classes in an imbalanced data scenario. The optimized feature subset achieved by balancing data using the oversampling method performed worse than the undersampling method. This may be due to the ability of the latter method to remove noisy observations close to the decision boundary, thus increasing the visibility of the minority class and reducing classification task complexity [25, 57]. The reduced data overlap enhanced sensitivity, highly desirable in the medical context, while offering good trade-offs between the majority and minority class accuracy rates [57]. By contrast, SMOTE may alter class distribution in the presence of noise and/outlier instances [33], unavoidable in medical data, giving rise to blurring of the decision boundary between the target classes [58]. We also found that the undersampling validation dataset performed significantly better than the balancing training data (FS_{TU} vs. FS_{VU} columns, Table 5.4), which by the undersampling method eliminate a great deal of information of the majority class for the training model. The total sample size used to design the model was thus too small to statistically represent their population, worsening the quality of the feature subspace and impairing the classifier performance [34]. However, the hybrid resampling method performed significantly better than the oversampling method, being slightly, but significantly, worse than undersampling, suggesting that the latter is the main cause of the relative improvement of the model performance in hybrid implementations.

After obtaining the optimized feature subset after balancing the validation data by the undersampling method, it was no longer necessary to apply the resampling method to the training data. In fact, similar results were obtained for the original data without the resampling and oversampling method. Again, as applying the undersampling method to the training data could even worsen the model performance due to information loss [21, 24] (see AP: RN vs. RU, Table 5.4), there was an insufficient sample size to design a robust preterm labor prediction system. Our results suggest that both the feature selection and resampling methods are effective to solve the classification task in imbalanced scenarios. These results agree with other authors who studied the combined feature selection and resampling method for imbalance data learning and found that in 79% of the study cases, balancing before feature selection improves the results [59]. We also showed the feasibility of combining both the feature selection and resampling methods in the same iterative process to deal with imbalanced data, in contrast to resampling before or after feature selection [59, 60]. Balancing validation data to deal with the imbalance data problem was similar to the strategy proposed by Jain et al., who used a weighted sum of recall and specificity as the fitness function [61]. By adding more weight to the recall metric, minority samples became more representative in the fitness function, thus to some extent overcoming the bias of the classifier towards the majority class [61].

Conventional accuracy is known to be unsuitable for evaluating classifier performance in an imbalanced scenario, although in the literature both the F1-score and G-mean have been widely used for this purpose [24, 52, 62]. Many studies highlight the weakness of the threshold-dependent metric in comparison to threshold-independent metrics such as AUC and AP in imbalanced scenarios [63, 64]. Jeni et al. compared a broad range of metrics that included both threshold-dependent metrics (accuracy, F1-score, Cohen's kappa, and Krippendorf's alpha) and threshold-independent metrics such as AUC of the LNLEn$_{\mathrm{ALL}}$ curve and precision-recall curve [62]. They found that all other metrics except threshold-independent metrics were attenuated by skewed distributions. Although the area under the LNLEn$_{\mathrm{ALL}}$ is a popular and strong measure to assess the performance of binary classifiers, it has been found that the LNLEn$_{\mathrm{ALL}}$ curve may provide an overly optimistic view when dealing with imbalanced data [27, 65]. By contrast, precision-recall curves can be more informative than LNLEn$_{\mathrm{ALL}}$ and have become the basis for assessing performance imbalanced data learning [27, 65]. In fact, a very different precision-recall would be obtained for the same LNLEn$_{\mathrm{ALL}}$ in these scenarios (see Table 5.5). In the present work we used threshold-independent metrics, as suggested by other authors [27, 65, 66], to avoid a data-skewed bias. Threshold-independent metrics avoid the optimization of the threshold for class assignment and ease the preliminary comparison of different classifier performances. After obtaining the best strategy to achieve the highest AUC and AP mean score, we further determined the threshold-dependent metrics by maximizing both the F1-score and G-mean. By maximizing the latter mean, we considerably increased the recall score by reducing the false negative cases that consisted of true preterm labor patients misclassified as term cases, despite the fact that this necessarily involved less precision [62]. The false negative cases in our application are especially relevant in obstetrics, due to the serious consequences of preterm birth on the newborn's health.

5.4.2 Preterm Labor Prediction System

Using the optimized feature subset obtained by undersampling the validation dataset, our best results achieved an AUC∼94% and AP∼84%. Although this result may perform worse than most studies in the literature that attempted to predict preterm birth by balancing the data by SMOTE before data partition [11, 15, 19, 29, 32, 67], there is no comparison from the methodological point of view. We believe that the generalization capability of the preterm term prediction model in these studies is overestimated, due to the leaked information between the training and testing partitions [26]. Our model outperformed that obtained by Vandewiele et al. who, as in the present work, conducted data partition before the resampling method [26].

The fact that we did not obtain even better results was due to diverse main factors. In addition to a small database with an imbalance problem, the features from the preterm and term classes were highly overlapping, since the EHG data was recorded a considerable time before delivery. Our results agree with other authors who found that the impact of class imbalance on sensitivity greatly depends on the degree of class overlap [25, 68], i.e., class imbalance had a greater impact when class overlap was high and seemed insignificant when low. For the case under study, a total of 326 registers were considered when the imbalance ratio was (preterm cases/term cases = 51/275). There is some evidence that overlapping between classes is the main cause of misclassification for this amount of records and imbalance ratio [25]. Other difficult factors, such as small sample size, the presence of disjoint data distribution, outlier and noise observations, and high dimensionality features could be amplified by the data imbalance, making the classification task more challenging [25]. The influence of the data imbalance problem decreases for larger datasets; when the train data is large enough imbalanced distributions do not prevent correct classification, even when the imbalance level is very high [69, 70]. There are only a few dozen minority instances in our application that can cause a possible distribution discrepancy between the training, validation, and testing data. Considering that 10% of all births will deliver preterm, an effect size of 0.2, error margin of 5%, and confidence level of 95%, at least 27 preterm women were necessary in the training, validation, and testing data to statistically represent the overall population [71]. This means approximately 810 patients were required to design a robust and generalizable preterm labor prediction system for clinical use. There is currently an urgent requirement for a large database of EHG records to determine its clinical value for predicting preterm labor. In this regard, although there are other publicly available EHG databases, we were unable to join databases from different sources due to the lack of a standardized protocol for data acquisition [21]. In addition, these databases were obtained from women in regular check-ups, which means that some important preterm birth prediction measures, such as cervical length, fetal fibronectin and/or interleukin 6 [11] are missing from their obstetric data [10, 46]. Including these additional clinical data in the classifier could therefore further enhance preterm labor prediction performance [13].

5.4.3 Limitations, Future Works and Practical Implications

Our results suggest that the best strategy to mitigate imbalanced data learning in highly overlapping classification tasks with small samples, which is very frequent in the medical data context, is to undersample the validation dataset to 1:1 during feature selection. Despite the promising results, the present work is not exempt from limitations: in addition to the limited sample size, we only tested our method by LDA classification methods in a specific application. This general recommendation should be further corroborated by future studies that seek to deal with imbalanced data learning using other classification methods and/or to be used for other classification tasks.

Future work may be directed toward the use of other strategies to mitigate the imbalanced data problem, such as cost-sensitive or ensemble learning [21], which to date has only been used to predict preterm births from EHG records using balanced data by oversampling before data partition. In spite of the limitations of our study, we believe that the results faithfully represent a realistic generalization capacity for new incoming data, with a recall ranging from 79.6% to 89.8%, which is considerably higher than the techniques commonly used in clinical practice [7, 72, 73, 74]. Our results contribute to more accurate prediction and prevention of preterm labor, which is highly relevant in clinical practice. Accurate prediction of preterm labor would allow screening out almost 75% of false threatened preterm labor cases, with an estimated cost of 20,372 USD/patient [75], which would give rise to substantial savings for public health systems. It would also allow clinicians to provide better and more personalized care to real preterm labor cases, potentially contribute to increasing the survival rate in cases of extreme prematurity by prolonging pregnancy, and reduce long-term morbidity and lifelong disabilities in survivors.

5.5 Conclusions

In the present work we have shown the feasibility of combining different resampling methods in feature selection and training the prediction model during the same iterative process to deal with the imbalanced data problem. We found that overlapping data between the target classes was the main problem in predicting preterm labor and was amplified by the data imbalance scenario. Feature selection by the genetic algorithm and intrinsically balancing the validation partition could significantly reduce data overlap between target classes and improve the model performance. This result highlights the importance of the feature quality for preterm labor prediction. Using the best chromosome by the genetic algorithm, subsequent resampling of the training dataset did not improve decision making, suggesting that the same feature subset was already optimally arranged to avoid information loss and noise between observations.

We also determined that the undersampling method during feature selection outperformed the oversampling method, thanks to its ability to enhance the visibility of the minority class by eliminating noisy observations close to the decision boundary, while undersampling seemed to be the main contribution of the model performance improvement in hybrid implementations. The best strategy to mitigate imbalanced data consisted of undersampling the validation dataset to 1:1 during feature selection, achieving an AUC~94% and AP~84%. The maximum F1-score was around 80%, with a recall of ~80%. By maximizing the G-mean, the best model achieved a recall of ~90%, with an F1-score around 72%. Our results represent a realistic estimation of the EHG technique's generalization capability for predicting preterm labor and outperform the current techniques used in clinical practice to detect true preterm labor cases, thus constituting a useful tool for clinical use for preterm labor prevention.

5.6 References

[1] WHO, Recommended definitions, terminology and format for statistical tables related to the perinatal period and use of a new certificate for cause of perinatal deaths. Modifications recommended by FIGO as amended October 14, 1976. *Acta Obstetricia et Gynecologica Scandinavica*, vol. 56, no. 3, pp. 247–253, 1977, ISSN: 16000412. DOI: 10.3109/00016347709162009.

[2] J. P. Vogel, S. Chawanpaiboon, A.-B. Moller, K. Watananirun, M. Bonet, and P. Lumbiganon, The global epidemiology of preterm birth, *Best Practice & Research Clinical Obstetrics & Gynaecology*, vol. 52, pp. 3–12, 2018. DOI: 10.1016/j.bpobgyn.2018.04.003.

[3] G. T. Mandy, Short-term complications of the preterm infant, *UpToDate*, vol. 46, no. 4965, pp. 1–17, 2019.

[4] S. Petrou, H. H. Yiu, and J. Kwon, Economic consequences of preterm birth: A systematic review of the recent literature (2009-2017), *Archives of Disease in Childhood*, vol. 104, no. 5, pp. 456–465, 2019, ISSN: 14682044. DOI: 10.1136/archdischild-2018-315778.

[5] N. J. Waitzman, A. Jalali, and S. D. Grosse, Preterm birth lifetime costs in the United States in 2016: An update, *Seminars in Perinatology*, vol. 45, no. 3, p. 151 390, 2021, ISSN: 1558075X. DOI: 10.1016/j.semperi.2021.151390.

[6] J. Jacob, M. Lehne, A. Mischker, N. Klinger, C. Zickermann, and J. Walker, Cost effects of preterm birth: a comparison of health care costs associated with early preterm, late preterm, and full-term birth in the first 3 years after birth, *The European Journal of Health Economics*, vol. 18, no. 8, pp. 1041–1046, 2017, ISSN: 1618-7601.

[7] R. E. Garfield and W. L. Maner, Physiology and electrical activity of uterine contractions, *Seminars in Cell and Developmental Biology*, vol. 18, no. 3, pp. 289–295, 2007, ISSN: 10849521. DOI: 10.1016/j.semcdb.2007.05.004.

[8] A. Leaños-Miranda, A. G. Nolasco-Leaños, R. I. Carrillo-Juárez, C. J. Molina-Pérez, I. Isordia-Salas, and K. L. Ramírez-Valenzuela, Interleukin-6 in amniotic fluid: A reliable marker for adverse outcomes in women in preterm labor and intact membranes, *Fetal Diagnosis and Therapy*, vol. 48, no. 4, pp. 313–320, 2021, ISSN: 14219964. DOI: 10.1159/000514898.

[9] 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, ISSN: 00029378. DOI: 10.1016/0002-9378(93)90456-S.

[10] 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 pre-term delivery groups, *Medical and Biological Engineering and Computing*, vol. 46, no. 9, pp. 911–922, 2008, ISSN: 01400118. DOI: 10.1007/s11517-008-0350-y.

[11] 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, 02TR01, 2018, ISSN: 13616579. DOI: 10.1088/1361-6579/aaad56.

[12] D. Schlembach, W. L. Maner, R. E. Garfield, and H. Maul, Monitoring the progress of pregnancy and labor using electromyography, *European Journal of Obstetrics and Gynecology and Reproductive Biology*, vol. 144, no. SUPPL 1, pp. 2–8, 2009, ISSN: 18727654. DOI: 10.1016/j.ejogrb.2009.02.016.

[13] J. Mas-Cabo *et al.*, Electrohysterogram for ANN-Based Prediction of Imminent Labor in Women with Threatened Preterm Labor Undergoing Tocolytic Therapy, *Sensors*, vol. 20, no. 9, p. 2681, 2020, ISSN: 14248220. DOI: 10.3390/s20092681.

[14] J. Mas-Cabo *et al.*, Robust Characterization of the Uterine Myoelectrical Activity in Different Obstetric Scenarios, *Entropy*, vol. 22, no. 7, p. 743, 2020, ISSN: 10994300. DOI: 10.3390/e22070743.

[15] F. Nieto-del-amor, R. Beskhani, Y. Ye-lin, J. Garcia-casado, and A. Diaz-martinez, Assessment of Dispersion and Bubble Entropy Measures for Enhancing Preterm Birth Prediction Based on Electrohysterographic Signals, *Sensors*, vol. 21, no. 18, 2021. DOI: 10.3390/s21186071.

[16] A. Lemancewicz *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, ISSN: 02085216. DOI: 10.1016/j.bbe.2015.11.005.

[17] J. Vrhovec and A. Macek, An Uterine Electromyographic Activity as a Measure of Labor Progression, in Applications of EMG in Clinical and Sports Medicine, Jan. 2012. DOI: 10.5772/25526.

[18] M. Hassan, J. Terrien, C. Marque, and B. Karlsson, Comparison between approximate entropy, correntropy and time reversibility: Application to uterine electromyogram signals, *Medical Engineering and Physics*, vol. 33, no. 8, pp. 980–986, 2011, ISSN: 13504533. DOI: 10.1016/j.medengphy.2011.03.010.

[19] F. Nieto-del-Amor *et al.*, Optimized Feature Subset Selection Using Genetic Algorithm for Preterm Labor Prediction Based on Electrohysterography, *Sensors*, vol. 21, no. 10, p. 3350, 2021. DOI: 10.3390/s21103350.

[20] J. Mas-Cabo, G. Prats-Boluda, J. Garcia-Casado, J. Alberola-Rubio, A. Perales, and Y. Ye-Lin, Design and Assessment of a Robust and Generalizable ANN-Based Classifier for the Prediction of Premature Birth by means of Multichannel Electrohysterographic Records, *Journal of Sensors*, vol. 2019, pp. 1–13, 2019, ISSN: 16877268. DOI: 10.1155/2019/5373810.

[21] T. Włodarczyk *et al.*, Machine learning methods for preterm birth prediction: A review, *Electronics (Switzerland)*, vol. 10, no. 5, pp. 1–24, 2021, ISSN: 20799292. DOI: 10.3390/electronics10050586.

[22] K. P. Murphy, *Machine Learning: A Probabilistic Perspective.* 2012, pp. 1–100, ISBN: 978026208029.

[23] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, *Electronics (Switzerland)*, vol. 8, no. 8, 2019, ISSN: 20799292. DOI: 10.3390/electronics8080832.

[24] Y. Sun, A. K. Wong, and M. S. Kamel, Classification of imbalanced data: A review, *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 4, pp. 687–719, 2009, ISSN: 02180014. DOI: 10.1142/S0218001409007326.

[25] M. Denil and T. Trappenberg, Overlap versus imbalance, in Canadian conference on artificial intelligence, A. Farzindar and V. Kešelj, Eds., vol. 6085 LNAI, 2010, pp. 220–231, ISBN: 3642130585. DOI: 10.1007/978-3-642-13059-5_22.

[26] G. Vandewiele *et al.*, Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling, *Artificial Intelligence in Medicine*, vol. 111, p. 101 987, 2021, ISSN: 18732860. DOI: 10.1016/j.artmed.2020.101987. arXiv: 2001.06296.

[27] S. Vluymans, Learning from imbalanced data, *Studies in Computational Intelligence*, vol. 807, no. 9, pp. 81–110, 2019, ISSN: 1860949X. DOI: 10.1007/978-3-030-04663-7_4.

[28] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, ISSN: 10769757. DOI: 10.1613/jair.953. arXiv: 1106.1813.

[29] 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, ISSN: 18728286. DOI: 10.1016/j.neucom.2015.01.107.

[30] 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, ISSN: 13616579. DOI: 10.1088/0967-3334/36/2/341.

[31] 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, e77154, 2013, ISSN: 19326203. DOI: 10.1371/journal.pone.0077154.

[32] 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, e0132116, 2015, ISSN: 19326203. DOI: 10.1371/journal.pone.0132116.

[33] M. Koziarski, M. Woźniak, and B. Krawczyk, Combined Cleaning and Resampling algorithm for multi-class imbalanced data with label noise, *Knowledge-Based Systems*, vol. 204, 2020, ISSN: 09507051. DOI: 10.1016/j.knosys.2020.106223. arXiv: 2004.03406.

[34] O. Loyola-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and M. García-Borroto, Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases, *Neurocomputing*, vol. 175, pp. 935–947, 2016, ISSN: 18728286. DOI: 10.1016/j.neucom.2015.04.120.

[35] Y. Liu, A. An, and X. Huang, Boosting prediction accuracy on imbalanced datasets with SVM ensembles, in Pacific-Asia conference on knowledge discovery and data mining, vol. 3918 LNAI, 2006, pp. 107–118, ISBN: 3540332065. DOI: 10.1007/11731139_15.

[36] N. Junsomboon and T. Phienthrakul, Combining over-sampling and under-sampling techniques for imbalance dataset, in 9th International Conference on Machine lLarning and Computing, vol. Part F1283, 2017, pp. 243–247, ISBN: 9781450348171. DOI: 10.1145/3055635.3056643.

[37] S. Park and H. Park, Combined oversampling and undersampling method based on slow-start algorithm for imbalanced network traffic, *Computing*, vol. 103, no. 3, pp. 401–424, 2021, ISSN: 14365057. DOI: 10.1007/s00607-020-00854-1.

[38] K. Fujiwara *et al.*, Over- and Under-sampling Approach for Extremely Imbalanced and Small Minority Data Problem in Health Record Analysis, *Frontiers in Public Health*, vol. 8, no. May, pp. 1–15, 2020, ISSN: 22962565. DOI: 10.3389/fpubh.2020.00178.

[39] K. Napierala and J. Stefanowski, Types of minority class examples and their influence on learning classifiers from imbalanced data, *Journal of Intelligent Information Systems*, vol. 46, no. 3, pp. 563–597, 2016, ISSN: 15737675. DOI: 10.1007/s10844-015-0368-1.

[40] L. Zhou, Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods, *Knowledge-Based Systems*, vol. 41, pp. 16–25, 2013, ISSN: 09507051. DOI: 10.1016/j.knosys.2012.12.007.

[41] M. Bekkar and T. A. Alitouche, Imbalanced Data Learning Approaches Review, *International Journal of Data Mining & Knowledge Management Process*, vol. 3, no. 4, pp. 15–33, 2013, ISSN: 2231007X. DOI: 10.5121/ijdkp.2013.3402.

[42] L. Yijing, G. Haixiang, L. Xiao, L. Yanan, and L. Jinling, Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data, *Knowledge-Based Systems*, vol. 94, pp. 88–104, 2016, ISSN: 09507051. DOI: 10.1016/j.knosys.2015.11.013.

[43] J. P. Cunningham and Z. Ghahramani, Linear dimensionality reduction: Survey, insights, and generalizations, *Journal of Machine Learning Research*, vol. 16, pp. 2859–2900, 2015, ISSN: 15337928. arXiv: 1406.0873.

[44] G.-h. Fu, F. Xu, B.-y. Zhang, and L.-z. Yi, Chemometrics and Intelligent Laboratory Systems Stable variable selection of class-imbalanced data with precision-recall criterion, *Chemometrics and Intelligent Laboratory Systems*, vol. 171, no. September, pp. 241–250, 2017, ISSN: 0169-7439.

[45] I. Ramos-Pérez, Á. Arnaiz-González, J. J. Rodríguez, and C. García-Osorio, When is resampling beneficial for feature selection with imbalanced wide data? *Expert Systems with Applications*, vol. 188, 2022, ISSN: 09574174. DOI: 10.1016/j.eswa.2021.116015.

[46] F. Jager, S. Libenšek, and K. Geršak, Characterization and automatic classification of preterm and term uterine records, *PLoS ONE*, vol. 13, no. 8, O. Uthman, Ed., e0202125, 2018, ISSN: 19326203. DOI: 10.1371/journal.pone.0202125.

[47] 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 and Biological Engineering and Computing*, vol. 57, no. 2, pp. 401–411, 2019, ISSN: 17410444. DOI: 10.1007/s11517-018-1888-y.

[48] J. Laurikkala, Improving Identification of Difficult Small Classes by Balancing Class Distribution, in Conference on artificial intelligence in medicine in Europe, June 2001, vol. 2101, 2001, pp. 63–66, ISBN: 3540422943. DOI: 10.1007/3-540-48229-6_9.

[49] O. Babatunde, L. Armstrong, J. Leng, and D. Diepeveen, A Genetic Algorithm-Based Feature Selection, *International Journal of Electronics Communication and Computer Engineering*, vol. 5, no. 4, pp. 899–905, 2014, ISSN: 2278-4209.

[50] M. H. Nguyen, Impacts of unbalanced test data on the evaluation of classification methods, *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 3, pp. 497–502, 2019, ISSN: 21565570. DOI: 10.14569/IJACSA.2019.0100364.

[51] R. B. D'Agostino, An omnibus test of normality for moderate and large size samples, *Biometrika*, vol. 58, no. 2, pp. 341–348, 1971, ISSN: 00063444. DOI: 10.1093/biomet/58.2.341.

[52] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, Review of classification methods on unbalanced data sets, *IEEE Access*, vol. 9, pp. 64 606–64 628, 2021. DOI: 10.1109/ACCESS.2021.3074243.

[53] M. B. Bin Heyat *et al.*, Wearable Flexible Electronics Based Cardiac Electrode for Researcher Mental Stress Detection System Using Machine Learning Models on Single Lead Electrocardiogram Signal, *Biosensors*, vol. 12, no. 6, p. 427, 2022, ISSN: 2079-6374. DOI: 10.3390/bios12060427.

[54] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, and J. Santos, Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches, *IEEE Computational Intelligence Magazine*, vol. 13, no. 4, pp. 59–76, 2018, ISSN: 15566048. DOI: 10.1109/MCI.2018.2866730.

[55] S. Maldonado, R. Weber, and F. Famili, Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines, *Information Sciences*, vol. 286, pp. 228–246, 2014, ISSN: 00200255. DOI: 10.1016/j.ins.2014.07.015.

[56] W. J. Lin and J. J. Chen, Class-imbalanced classifiers for high-dimensional data, *Briefings in Bioinformatics*, vol. 14, no. 1, pp. 13–26, 2013, ISSN: 14675463. DOI: 10.1093/bib/bbs006.

[57] P. Vuttipittayamongkol and E. Elyan, Overlap-Based Undersampling Method for Classification of Imbalanced Medical Datasets, in IFIP International Conference on Artificial Intelligence Applications and Innovations, I. Maglogiannis, L. Iliadis, and E. Pimenidis, Eds., 2020, pp. 358–369, ISBN: 9783030491864. DOI: 10.1007/978-3-030-49186-4_30.

[58] R. Alizadehsani *et al.*, Handling of uncertainty in medical data using machine learning and probability theory techniques: a review of 30 years (1991–2020), *Annals of Operations Research*, no. 0123456789, pp. 1–42, 2021, ISSN: 15729338. DOI: 10.1007/s10479-021-04006-2.

[59] R. Martín-Félez and R. A. Mollineda, On the suitability of combining feature selection and resampling to manage data complexity, in Conference of the Spanish Association for Artificial Intelligence, vol. 5988 LNAI, 2009, pp. 141–150, ISBN: 364214263X. DOI: 10.1007/978-3-642-14264-2_15.

[60] M. W. Huang, C. H. Chiu, C. F. Tsai, and W. C. Lin, On combining feature selection and over-sampling techniques for breast cancer prediction, *Applied Sciences (Switzerland)*, vol. 11, no. 14, 2021, ISSN: 20763417. DOI: 10.3390/app11146574.

[61] A. Jain, S. Ratnoo, and D. Kumar, Addressing class imbalance problem in medical diagnosis: A genetic algorithm approach, in 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC), 2018, pp. 1–8, ISBN: 9781509063130. DOI: 10.1109/ICOMICON.2017.8279150.

[62] L. A. Jeni, J. F. Cohn, and F. De La Torre, Facing imbalanced data - Recommendations for the use of performance metrics, in 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, 2013, pp. 245–251, ISBN: 9780769550480. DOI: 10.1109/ACII.2013.47.

[63] N. Japkowicz, Assessment metrics for imbalanced learning, *Imbalanced Learning: Foundations, Algorithms, and Applications*, pp. 187–206, 2013. DOI: 10.1002/9781118646106.ch8.

[64] H. R. Sofaer, J. A. Hoeting, and C. S. Jarnevich, The area under the precision-recall curve as a performance metric for rare binary events, *Methods in Ecology and Evolution*, vol. 10, no. 4, pp. 565–577, 2019, ISSN: 2041210X. DOI: 10.1111/2041-210X.13140.

[65] T. Saito and M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, *PLoS ONE*, vol. 10, no. 3, pp. 1–21, 2015, ISSN: 19326203. DOI: 10.1371/journal.pone.0118432.

[66] Y. Yuan, W. Su, and M. Zhu, Threshold-Free Measures for Assessing the Performance of Medical Screening Tests, *Frontiers in Public Health*, vol. 3, no. April, 2015, ISSN: 2296-2565. DOI: 10.3389/fpubh.2015.00057.

[67] U. R. Acharya *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, ISSN: 18790534. DOI: 10.1016/j.compbiomed.2017.04.013.

[68] P. Vuttipittayamongkol, E. Elyan, and A. Petrovski, On the class overlap problem in imbalanced data classification, *Knowledge-Based Systems*, vol. 212, p. 106 631, 2021, ISSN: 09507051. DOI: 10.1016/j.knosys.2020.106631.

[69] N. Japkowicz, Class imbalances: are we focusing on the right issue, in Proc. the ICML 2003 Workshop on Learning from Imbalanced Data Sets (II), 2003.

[70] T. Jo and N. Japkowicz, Class imbalances versus small disjuncts, *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 40–49, 2004. DOI: https://doi.org/10.1145/1007730.1007737.

[71] C. C. Serdar, M. Cihan, D. Yücel, and M. A. Serdar, Sample size, power and effect size revisited: Simplified and practical approachin pre-clinical, clinical and laboratory studies, *Biochemia Medica*, vol. 31, no. 1, pp. 1–27, 2021, ISSN: 13300962. DOI: 10.11613/BM.2021.010502.

[72] V. Berghella, E. Hayes, J. Visintine, and J. K. Baxter, Fetal fibronectin testing for reducing the risk of preterm birth, *Cochrane Database of Systematic Reviews*, no. 4, 2008, ISSN: 1469493X. DOI: 10.1002/14651858.CD006843.pub2.

[73] M. Pandey, M. Chauhan, and S. Awasthi, Interplay of cytokines in preterm birth, *Indian Journal of Medical Research*, vol. 146, no. September, pp. 316–327, 2017, ISSN: 09715916. DOI: 10.4103/ijmr.IJMR_1624_14.

[74] M. Sean Esplin *et al.*, Predictive accuracy of serial transvaginal cervical lengths and quantitative vaginal fetal fibronectin levels for spontaneous preterm birth among nulliparous women, *Obstetrical and Gynecological Survey*, vol. 72, no. 7, pp. 397–399, 2017, ISSN: 15339866. DOI: 10.1097/OGX.0000000000000455.

[75] M. Lucovnik, L. R. Chambliss, and R. E. Garfield, Costs of unnecessary admissions and treatments for ”threatened preterm labor”, *American Journal of Obstetrics and Gynecology*, vol. 209, no. 3, 217.e1–217.e3, 2013, ISSN: 10976868. DOI: 10.1016/j.ajog.2013.06.046.

Chapter 6

General Discussion

6.1 EHG characterization

In this study, our aim was to develop an EHG preterm labor forecasting system to promote its transfer to clinical settings. The classical approach consists of extracting relevant information from the EHG used as the input features of classification algorithms. Previous studies have proposed a wide range of features to characterize EHG changes throughout pregnancy, divided into four main groups: temporal, spectral, nonlinear, and synchronization features [1, 2]. These four feature categories encompass the different phenomena involved in the efficiency of uterine myoelectrical activity (intensity, excitability and synchronization) and their non-linear nature, which allows an objective global assessment of the uterine electrophysiological state for better preventing preterm labor [1, 2].

6.1.1 Analysis of temporal and spectral EHG features

Amplitude-related features such as peak-to-peak amplitude or RMS are commonly used temporal parameters to characterize EHG signals. These latter reflect the number of uterine cells involved in the contraction and therefore are indirect measures of uterine contraction intensity. While it is widely accepted that the amplitude of the EHG signal tends to increase as labor approaches [3, 4, 5], the utility of amplitude-based features for preterm labor prediction has provided controversial results. Amplitude-related metrics are effective in discriminating the imminence of delivery in EHG recordings from women with threaten preterm labor [6, 7]. In contrast, other authors found that women who delivered prematurely exhibited a similar amplitude to term delivery cases with no significant difference [8, 9], which agrees with our findings [10]. This is mainly due to the fact that surface-recorded EHG depends not only on the internal source signal amplitude but is also affected by other factors, such as body mass index and skin preparation, among others [4], so that surface-recorded amplitude is not a reliable measure of uterine contraction intensity.

Spectral parameters were found to be more reliable than temporal parameters in characterizing EHG changes throughout pregnancy [4, 8]. Physiologically, uterine contractility depends on the uterine myocyte excitability and the propagation of its electrical activity to the whole uterus, which plays a crucial role in shaping the spectral content of the EHG signal [3, 8]. Previous studies showed that spectral features, such as dominant frequency, median frequency, decile, high-to-low frequency energy ratio and spectral moment ratio, effectively increase as labor approaches [4, 5, 8, 11, 12]. This means that uterine cells become more excitable, giving rise to the shift of spectral content towards higher frequencies [4, 8]. Preterm delivery was also found to have a significantly higher dominant frequency than term delivery in both the FWH and WBW bandwidth in women who have regular check-ups [5, 8, 11, 12, 13]. We confirmed that dominant frequency is a relevant discriminator between preterm and term deliveries [10]. Also, the median frequency is considered an effective feature for distinguishing between term and preterm deliveries within the FWH, but not in the WBW [8, 11, 12]. In the present thesis our findings indicate that the median frequency in the range of 0.2–1Hz is a reliable discriminator between preterm and term labor [10]. We preferred to restrict the upper frequency limit to 1 Hz to reduce cardiac interference, while the lower frequency limit was set to minimize the influence of baseline fluctuation. Our study also revealed that the spectral moment ratio of preterm deliveries was significantly lower than term deliveries [10], which agrees with the smaller value for women during the active phase of labor as compared to antepartum patients [5]. Similarly, we found that the high-to-low frequency ratio and deciles were also effective in differentiating between preterm and term labor cases [10], while other spectral parameters were found to be less sensitive in detecting preterm delivery in women with regular check-ups, with no significant differences with the term-labor group, such as normalized energy or Teager energy [10].

6.1.2 Analysis of non-linear EHG features

The physiological mechanisms of biological systems are widely recognized to be non-linear processes that change over time. These systems can be modeled as non-linear dynamic systems due to the coupling between billions of interconnected cells and the complex feedback networks inherent to them. This nonlinearity is a fundamental characteristic of biological systems [4]. During the transition from pregnancy to labor, the properties of EHG signals such as complexity and randomness, experience a significant decline [5, 14], while regularity and predictability tend to increase [5, 14]. These physiological changes are reflected in different measures such as Lempel-Ziv, SD1/SD2 (Poincaré ellipse minor axis (SD1) and major axis (SD2)), sample, fuzzy and spectral entropy [5, 8, 14, 15, 16, 17]. Preterm EHG records are expected to show patterns that closely resemble those observed during labor rather than in term deliveries [4, 5].

Sample entropy has been widely used to detect preterm labor in women undergoing regular check-ups and identify imminent labor in cases of threatened preterm birth. Previous studies have indicated better discriminatory capability in the FWH than WBW bandwidth [4, 6, 8, 18], which aligns with the findings of the present thesis [10, 19]. Fuzzy entropy and spectral entropy obtained a similar performance in differentiating between term and preterm deliveries [16, 17]. We obtained a similar result [10, 19] to previous studies for Lempel-Ziv and Poincaré derived measures [5, 8].

Our results showed the Katz Fractal Dimension as an efficient discriminator between preterm and term labor, regardless of whether it was computed in the FWH or WBW bandwidths [10]. Our observations revealed that time reversibility could not distinguish between preterm and term labor [10], in agreement with other studies, which suggests that time reversibility may be insensitive to subtle electrophysiological changes occurring well before delivery, while it may be more suitable for identifying imminent labor in women with threatened preterm labor [5, 20, 21].

We propose two new biomarkers in this thesis (dispersion entropy and bubble entropy) to enhance the separability of the preterm and term groups [19]. Our findings suggest that bubble entropy better discriminates women who deliver at term from those who deliver prematurely than dispersion entropy, while both outperform sample, fuzzy, and spectral entropy [19]. These results agree with those of Azami et al., who showed that dispersion entropy outperformed classical entropy measures (sample, fuzzy and permutation entropy) in detecting Alzheimer's disease, with the additional advantage of less computation time [22]. Our findings also align with those of previous studies that found that dispersion entropy outperforms sample entropy in characterizing the intrinsic dynamics of gait maturation and discriminating noninvasive blood pressure signals in hypertensive rats [23]. However, they disagree with Romero-Morales et al., who found that dispersion entropy outperforms bubble entropy in distinguishing between preterm and term deliveries when computed in the FWH bandwidth [24]. This discrepancy may be due to various factors: firstly, Romero-Morales et al. only considered women in the 27$^{\text{th}}$ to 33$^{\text{rd}}$ week of gestation and subsampled term delivery records to match the sample size with a minority class (17 preterm vs. 17 term). Neither did they discard the corrupted EHG segments, which could have significantly altered the data distribution in both groups [24]. Based on these results, we speculate that dispersion entropy is in fact a robust measure against motion artifacts and would be suitable for real-time application. On the other hand, bubble entropy seems to be more vulnerable to motion and respiratory artifacts than dispersion entropy as it better discriminates artifact-free signals, although its performance can be compromised by noisy signals with both motion and breathing artifacts [19, 24]. Further research is required to validate this hypothesis and compare the effectiveness of these measures in EHG recordings with artifact removal.

6.1.3 Analysis of synchronization EHG features

During pregnancy, uterine myoelectrical activity is reduced and uncoordinated to facilitate fetal growth and development. As labor approaches, there is a down-regulation of progesterone, responsible for maintaining uterine quiescence, while the expression of hormone receptors such as prostaglandin and oxytocin, which promote uterine contractility, is up-regulated [25]. During the transition from pregnancy to labor, there is also an association with the formation of gap junctions comprising connexin proteins which facilitate a low-resistance pathway for the propagation of inter-cell action potentials. This enhanced electrical coupling propagates action potentials throughout the uterus, giving rise to high-intensity and better synchronized uterine electrical activity [3, 26]. The synchronization of the different uterine regions thus plays a crucial role in determining labor proximity.

Bivariate analysis of multichannel EHG recorded from electrode arrays has shown this increased coupling. Measures such as the non-linear correlation coefficient, the imaginary part of coherency, and normalized permutation and cross mutual information have shown this association [27, 28], and, as expected, a significant difference was also found in uterine electrical coupling during labor and for 1 week postpartum [29].

It is recommended to strategically position the electrodes across a wide area of the uterus, rather than focusing on a localized region, to analyze synchronization, also to maintain an interelectrode-distance greater than 5-6 cm to minimize the blurring effect of volume conduction, regardless of the synchronization measure [30]. However, multichannel EHG recordings increase the recording system's complexity, which hinders the system's transfer to clinical practice. In this thesis we therefore focused on temporal, spectral and non-lineal features to characterize EHG signals and develop a clinical preterm labor prediction system [10, 19, 31].

6.2 Overcoming the curse of dimensionality

6.2.1 Effective feature selection strategies for preterm labor prediction

As mentioned above, a large number of features have been used to characterize different phenomena involved in uterine contraction efficiency, including intensity, excitability, and non-linear dynamics [3]. They are often used as input features in preterm labor prediction systems, giving rise to the phenomenon known as the "curse of dimensionality" [32]. High-dimensional input feature data not only pose a significant computational burden, but also increase the risk of overfitting, compromising the performance of the prediction model. Since input data could contain complementary, redundant and noisy features, their identification and processing are necessary to overcome this issue. The principal component analysis (PCA) has been widely used for reducing the dimensionality of input features to predict preterm labor [4]. However, it is not very useful in determining complementarity and redundancy in the features or discarding noisy information, since PCA essentially results in linear combinations of the original data [33]. However, feature selection techniques have been shown to be more effective than PCA in dealing with redundancy and noise [34]. In this regard, filter methods have been used to deal with the features that cannot distinguish between term and preterm labor [4, 35]. Other studies have suggested that filter methods performed worse than wrapper methods, which find the most informative feature subset by minimizing misclassifying classes by the machine learning algorithm [36, 37, 38].

Sequential forward selection is a widely used wrapped method for feature selection in preterm labor prediction from EHG signals [24, 36, 39]. This latter starts with an empty feature set and iteratively adds one feature at a time, based on the performance of a machine learning model [40]. Alternatively, Particle Swarm Optimization (PSO) and genetic algorithm are other wrapper methods, while evolutionary algorithms, which evaluate a fitness function for multiple random feature subsets to combine the feature subsets with the best performance to iteratively converge in the optimum feature subset [40]. PSO and the genetic algorithm mainly differ in the way they create the new population of feature subsets [41]. The genetic algorithm is more suitable for discrete optimization, as in feature selection, while PSO is better for continuous optimization [41]. According to Alamedine et al., SFS performs worse than (PSO) in predicting labor and pregnancy contractions by LDA, QDA, and KNN as classification methods [36]. Benalcazar et al. also used PSO and neural networks to predict labor induction success [42]. Similarly, the genetic algorithm has been shown to perform better than filter methods for predicting pregnancy and labor contractions by KNN [43] and outperformed both forward and backward selection when predicting central nervous system embryonal tumor outcomes based on gene expression [37]. In the present thesis, we show the genetic algorithm's potential to optimize the feature subspace with the maximum complementary information between them, eliminating redundant and noisy features [10, 19]. However, we also consider that PSO could yield similar results to the genetic algorithm in this context, as has been found in similar applications [40, 41].

6.2.2 Genetic algorithm and imbalanced datasets: tailoring feature selection for preterm birth prediction

Starting from an initial set with more than 140 features, we obtained a different optimized feature subspace using KNN, LDA and logistic regression [10]. This agrees with other authors who found that the wrapper method drastically reduced feature subspace dimensionality [39]. In fact, there may not be an absolute optimal feature subset with the best performance, as research has shown that the presence of complementary and redundant information can lead to multiple optimized feature subsets (multiple local solutions in an optimization problem) of similar performance [36, 44, 45, 46]. Moreover, it is worth noting that in our specific application, there could be a large number of features with redundant information. In this case, the absence of one feature could easily be replaced by another with similar classification performance. So, we found that few features appeared repeatedly in the four different optimum feature subsets (Dec5, spectral entropy, Lempel-Ziv, SDRR, and week of gestation) [10, 19] and thus seem to provide significant information to predict preterm labor. It is worth noting that Dec5 has been shown to be a reliable predictor of preterm labor in previous studies [4, 5, 8] and has also been included in the optimal feature subset of a similar research project [36], while bubble entropy was the most frequently selected non-linear feature in the initial feature set [19]. Romero-Morales et al. reported that dispersion and bubble entropy were also selected repeatedly in their optimal feature subset for predicting preterm labor [24]. However, in terms of obstetric data, only gestational age was found to be relevant in predicting preterm labor [10, 19], probably because the values of EHG features change intrinsically as gestational age increases [8, 47, 48], while maternal age, parity and abortions were determined to be irrelevant for the algorithm [10, 19], despite being associated in the literature with risk factors for preterm delivery [49]. Future work with a larger database should be carried out to confirm this.

Our findings also indicate that the strategy used to deal with imbalanced data has a strong impact on feature subspace representation. In other words, the optimum feature set obtained by the genetic algorithm consisted of 12 and 58 features when using resampling-partitioning [19] and partitioning-resampling [31] pipeline, respectively. These results align with a previous study on imbalanced real-life heart failure prediction, which showed that the resampling-partitioning pipeline outperformed the partitioning-resampling approach in terms of achieving better performance with a lower-dimensional feature subset [50]. This suggests that data oversampling before splitting could lead to an information leakage across partitions, thereby reducing the complexity of the classification task [51] and giving rise to a lower-dimensional optimized feature subspace. However, when using the partitioning-resampling pipeline in highly imbalanced datasets, the classification task would be more challenging, which is reflected in a high-dimensional optimized feature subset (58 vs. 12 features) [19, 31]. As the greater the imbalance ratio and overlap between classes, the more complex the classification task [52], so that a larger number of features will be necessary to increase the sensitivity of the minority class and improve the separability of the minority and majority class [53].

In addition, we examined the impact of resampling (undersampling, oversampling and hybrid methods) of the training and validation partition on the performance of the genetic algorithm. Our results indicated a significant improvement in the performance of the genetic algorithm when using a balanced-by-undersampling validation set as compared to the imbalanced set [31]. This suggests a better approximation to the objective function of the genetic algorithm, i.e. allowing it to focus on preterm births and increase the visibility of preterm samples with respect to term samples. This approach of balancing the target data to find the best feature subset has been proposed previously: balancing before feature selection yielded better results than balancing after feature selection in 79% of the study cases [54]. Other works have tailored the genetic algorithm by modifying the fitness function to enhance the "visibility" of the minority class [53, 55, 56]. Our opinion is that similar results can be obtained by modifying the fitness function as when balancing the validation partitions (our approach), provided that the fitness function is modified to allow the genetic algorithm to focus on the minority class, i.e. premature cases. For instance, Jain et al. used a weighted sum of recall and specificity as the fitness function, with more weight assigned to recall [55]. By doing so, the minority samples became more representative in the fitness function, enhancing their classification [55]. Our conclusion is therefore that using the genetic algorithm for feature selection can be an effective approach in addressing imbalanced datasets, a method that allows for the selection of meaningful features while reducing classification complexity and improving the visibility of the minority class [31].

6.3 Imbalanced data learning for a preterm labor prediction system: challenges and perspectives

6.3.1 Resampling before partitioning in preterm labor prediction from EHG signals

Various preterm labor prediction systems with promising results have been proposed [4, 57], although many of them overestimate the system performance from EHG registers due to the method used. Their main flaw is oversampling the preterm class by generating synthetic samples to match the number of term delivery samples before splitting the dataset into training and test partitions. This practice can lead to information leakage from the training partition to the testing partition, compromising the integrity of the evaluation process [51]. Most preterm delivery prediction systems published to date have been implemented have used this method. Despite their limitation, these studies still offer significant insights since they show the EHG potential for predicting preterm labor and encourage the scientific community to further study in this field. Regardless of the methodology bias, these studies serve as a fundamental basis for preterm labor prediction systems, providing valuable information for benchmarking new machine learning algorithms. In the present thesis, we therefore first developed a preterm delivery prediction system using this flawed approach to compare our specific pipeline with others in the literature.

Using a data resampling-partitioning strategy and the genetic algorithm to optimize the feature subspace, we demonstrated the feasibility of developing a preterm labor prediction system by assessing their generalization capacity with simple and easy to interpret algorithms such as logistic regression, LDA and KNN. These individual base classifiers achieved an F1-score of over 80% for the testing partition [10, 19]. Our results are thus hardly comparable to numerous prior works on preterm birth prediction systems that used cross-validation techniques to design and validate the classifiers without determining the true generalization capacity of the classifiers for incoming data that had never been encountered before [4, 11, 58, 59, 60]. Moreover, these works typically achieve accuracy scores and an AUC over 90% but have had little impact on clinical practice. This is mainly because of using complex algorithms like neural networks or support vector machines that are difficult to interpret, making it hard for obstetricians to trust their predictions [61]. However, thanks to the simplicity and easy-to-interpret algorithms, the classification algorithms in the present thesis could potentially boost obstetricians' confidence in the preterm labor prediction model's outcome and bring the EHG technique closer to clinical practice.

We also assessed the superior performance of an ensemble classifier to that of base classifiers for preterm labor prediction [10]. Ensemble methods, which combine the output of individual weak classifiers, have been effective in producing accurate predictions for various classification tasks [62]. The success of these methods is attributed to their ability to improve accuracy and rectify misclassifications across a multitude of simple and different base classifiers [63, 64]. Successful ensemble methods seek to achieve a balance between the diversity and accuracy of the base classifiers [65]. Ren et al. compared four simple classifiers and two ensemble classifiers to create a preterm birth classifier based on EHG features. The two ensemble classifiers, AdaBoost and Random Forest, obtained an AUC of 98.6% and 95.7% and outperformed the best simple classifier, a Bayesian Network, with an AUC of 91.2% [66]. Idowu et al. also achieved better results in preterm labor prediction by applying the Random Forest ensemble classifier, with an AUC of 94.2 ±1.4%, compared with the Penalized Logistic Regression classifier with an AUC of 91.9 ±3%. Moreover, the variability performance (σ/μ) of the simple classifier ($\sim$3.2) is twice that of the ensemble classifier ($\sim$1.5%) [67]. Our research supports the fact that adding a basic ensemble classifier, which relies on a majority voting strategy at the meta-level and uses outputs from easy-to-interpret base classifiers, can enhance individual classification performance of these latter, e.g. the best base classifier achieved an AUC of 94.7 ±2.5%, whereas the ensemble classifier achieved a notably higher value of 98.1 ±1.3% [10]. This approach yielded increased average metrics and minimized variations across partitions [67, 68, 69], which affects the reduction of the performance variability of ensemble classifiers with respect to base classifiers [10, 67].

6.3.2 Partitioning before resampling in preterm labor prediction from EHG signals

Applying resampling methods before data partitioning, the prediction system performance was drastically reduced to an AUC of <65% and average precision of <40% [31]. This is in line with Vandewiele et al. [51], who were the first to report that balancing the database by generating synthetic samples of preterm delivery records before splitting the dataset into training and test partitions could provide unrealistically promising results. The information from the training partition thus ends up by leaking into the testing partition, i.e. oversampling before data partition reduces the complexity of the classification task but overestimates generalization capability [70]. It is crucial for testing samples to be completely "new" and "unseen" throughout the entire machine learning pipeline, as the final evaluation is representative of the forecasting algorithm's performance in real-world situations. If the testing partitions contain leakage information from the training data, the performance results become unrealistic and is a hindrance to use in clinical practice [51].

Different strategies have been proposed to improve labor prediction results better than earlier studies, using the data partitioning-resampling pipeline [4, 51, 57]. While these studies were able to increase the AUC to between 65% and 80%, the metrics evaluating the precision of the algorithm to detect true preterm labor were still lower than 40% [71, 72]. Lou et al. reported the best benchmark scores with the partitioning-oversampling pipeline, achieving a sensitivity of 84 ±10%, specificity of 66 ±6%, and an AUC of 84 ±7% [73]. However, they did not report on precision, F1-score, or average precision, which are essential in assessing the classifier's ability to identify true preterm labor. Instead, the use of threshold-independent metrics like AUC in conjunction with AP is stronger [74, 75, 76, 77]. Conventional evaluation metrics such as accuracy, recall, precision or specificity are not suitable for evaluating classifier performance in imbalanced scenarios, presenting an overly optimistic view with imbalanced data [75, 76]. In a study comparing various classifier metrics, skewed distributions attenuated all of these except for threshold-independent metrics [74]. Precision-recall curves are now commonly used for evaluating performance in imbalanced data learning because they provide more informative results [78, 79]. It is important to note that in these scenarios, the same AUC curve can yield a different precision-recall curve [78, 79]. We therefore consider that our preterm labor prediction system, based on EHG features, outperforms all previous studies with a realistic and unseen partition for classifier assessment, achieving an AUC of 94.5 ±4.6% and AP of 84.8 ±11.7% [31]. The prediction model was developed by subsampling the validation partition during feature selection by a genetic algorithm. After obtaining the optimal feature subset, there is no need to resample the validation set [31]. Oversampling does not provide additional information and undersampling can worsen model performance due to information loss [31, 57, 80]. The proposed EHG-based preterm labor prediction method is expected to have high transferability to clinical use due to its simplicity: the final method relies only on a set of features and a straightforward LDA classifier, making it easy for clinical staff to understand and use it with confidence.

6.4 Limitations and future work

Despite promising results, our studies are not exempt from limitations. Firstly, the databases were not only highly imbalanced, but were also small [8, 81], which are common problems in the published databases of term and preterm EHG records. This problem mainly arises from the absence of a standard protocol for electrode placement in EHG recordings, which makes it difficult to compare databases [82]. It is thus crucial to establish a standard for EHG signal acquisition to generate bigdata, which can provide the groundwork for the development of reliable and transferable preterm birth prediction systems for clinical application. In this regard, IoT systems could help to create large databases by conducting multicenter studies [83].

Due to the nature of small databases, it is especially challenging to detect preterm labor risk with high sensitivity. Regardless of the method used, the preterm labor group may not be representative of the total population, so that future studies are still needed to further corroborate the generalization capability of these labor prediction systems. Increasing the number of observations of preterm and term labor would help to establish a more realistic boundary between the preterm and term classes. However, regardless of the sample size, there may be a theoretical upper limit to a prediction systems' maximum performance [84, 85, 86] due to the inherent overlapping of the preterm and term classes, which makes the classification task even more difficult.

Besides the resampling method used to deal with the imbalanced data problem, cost sensitive learning has also been proposed to predict preterm labor [57] although they use a data resampling-partitioning scheme that may overestimate the generalization capacity of the prediction model [51, 57]. Future work should be directed toward cost-sensitive learning with data the partitioning-resampling pipeline to estimate the "unbiased" generalization capability for predicting preterm labor.

Most preterm labor prediction systems are based on the classical approach, which defines EHG features with physical interpretation by specialists in the field. Convolutional neural networks (CNN) with automatic feature extraction capability could be another disruptive technique for predicting preterm labor. This approach has already been used to classify basal and contraction segments in EHG signals [87, 88]. However, possibly due to the need for a large amount of data to develop a consistent CNN [89], we have not found any delivery prediction system based on automatic feature extraction from EHG signals. In this regard, future work should be directed towards estimating the prediction model's performance of CNN-based systems by the data partitioning-resampling pipeline. It should also include transfer learning from pretrained models. Transfer learning can enhance the performance of CNN-based prediction models using the knowledge and representations learned from large-scale datasets, even with limited labeled data [90, 91]. This makes the models able to capture relevant features and patterns associated with preterm labor and improves their overall predictive capabilities.

In the present thesis we only analyzed demographic data for the detection of the risk of preterm labor due to the fact that EHG signals are typically recorded during routine check-ups in asymptomatic women with no preterm labor risk. In the literature, cervix measurement and chemical biomarkers such as cervical length, fetal fibronectin, and interleukin 6 are known to be effective predictors of preterm birth [4, 8, 81]. Including these clinical data could potentially enhance the performance of the preterm birth prediction models [16].

The present thesis has focused only on forecasting preterm labor from EHG recordings of singleton women who attended routine check-ups and therefore had no early labor onset symptoms. In clinical practice, tocolytic drugs are usually administered at the first signs of Threatened Preterm Labor (TPL). Previous studies showed that the EHG characteristics of undergoing alterations across various stages of tocolytic therapy [4, 6, 18] make imminent/preterm labor prediction based on EHG more challenging. Regardless of the tocolytic treatment phase at the EHG recording, there is evidence indicating that EHG can offer meaningful data for predicting imminent labor (time to delivery $< 7/14$ days) in single-gestation women with TPL using data resampling-partitioning, achieving an AUC of 91% [92]. Future research should be carried out on developing a prediction model for imminent/preterm labor in women with TPL using data partitioning-resampling and/or cost-sensitive learning. It will also be necessary to consider the phase of the tocolytic drug therapy at the time of recording.

Another limitation of the current study is that we only validated the prediction model in singleton gestations, while multiple gestation pregnancies have a higher incidence of preterm birth (up to 60%), with the risk increasing with the number of fetuses [93]. Multiple gestations are also associated with a higher very premature risk (19.19% vs 2.11%) [94]. The conventional treatments for single gestation, including tocolytic therapy, progesterone, and cervical cerclage, have shown limited effectiveness in multiple gestation pregnancies [95].In other words, multiple gestation pregnancies may involve specific pathways that contribute to increased uterine contractility, which in turn can lead to preterm birth [96]. Future work should be in this direction and would provide healthcare clinicians with a more reliable tool for preventing preterm labor in multiple gestation pregnancies, leading to improved clinical maternal-fetal outcomes.

However, EHG recordings are not the only promising techniques for developing effective preterm labor predictors for patients who have routine examinations. Grigorescu et al. recently used 157 neonatal T2-weighted magnetic resonance imaging records (MRI) of infants born between 23–42 weeks of gestation to predict preterm labor, using a 3D convolutional neural network (CNN) with layer-wise relevance propagation, which provides visual interpretation of the network's decisions. The outcomes were very encouraging, with an accuracy of 94%, a true positive rate of 100%, and a true negative rate of 86%. Their results showed that the most prominent feature to distinguish between preterm and term pregnancies was the infant's cerebrospinal fluid [97]. However, despite the promising results, MRI has major drawbacks for clinical practice. MRI machines are expensive to purchase and maintain and can take a long time to perform, sometimes up to an hour or more. This can be difficult for patients who are unable to remain still for that amount of time, as in pregnant women. These machines are not as widely available as other imaging technologies, especially in rural or low-income areas [98].

Although our study has its limitations, we believe that our results accurately reflect the models' generalization capacity for new data, achieving recall rates ranging from 79.6% to 89.8%, which is considerably higher than the methods currently used in clinical practice [99, 100, 101, 102]. These results contribute to improving prediction and prevention of preterm labor, which is important in clinical settings. Accurate prediction of preterm labor could eliminate false preterm labor cases, with estimated savings of $32,325 per patient [103], a significant cost saving for public healthcare systems. It would also enable clinicians to provide better and more personalized care for real preterm labor cases, potentially improving survival rates for extremely premature infants by prolonging pregnancy and reducing long-term morbidity and lifelong disabilities in the survivors.

6.5 References

[1] R. E. Garfield, L. Murphy, K. Gray, and B. Towe, Review and Study of Uterine Bioelectrical Waveforms and Vector Analysis to Identify Electrical and Mechanosensitive Transduction Control Mechanisms During Labor in Pregnant Patients, *Reproductive Sciences*, vol. 28, no. 3, pp. 838–856, 2021, ISSN: 19337205. DOI: 10.1007/s43032-020-00358-5.

[2] J. M. MARSHALL, Regulation of activity in uterine smooth muscle. *Physiological reviews. Supplement*, vol. 5, pp. 213–27, 1962, ISSN: 0554-1395.

[3] 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, ISSN: 00029378. DOI: 10.1016/0002-9378(93)90456-S.

[4] 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, 02TR01, 2018, ISSN: 13616579. DOI: 10.1088/1361-6579/aaad56.

[5] J. Mas-Cabo *et al.*, Robust Characterization of the Uterine Myoelectrical Activity in Different Obstetric Scenarios, *Entropy*, vol. 22, no. 7, p. 743, 2020, ISSN: 10994300. DOI: 10.3390/e22070743.

[6] 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 and Biological Engineering and Computing*, vol. 57, no. 2, pp. 401–411, 2019, ISSN: 17410444. DOI: 10.1007/s11517-018-1888-y.

[7] 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, 378.e1–378.e6, 2008, ISSN: 00029378. DOI: 10.1016/j.ajog.2008.08.003.

[8] 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 pre-term delivery groups, *Medical and Biological Engineering and Computing*, vol. 46, no. 9, pp. 911–922, 2008, ISSN: 01400118. DOI: 10.1007/s11517-008-0350-y.

[9] 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, ISSN: 02085216. DOI: 10.1016/j.bbe.2016.06.004.

[10] F. Nieto-del-Amor *et al.*, Optimized Feature Subset Selection Using Genetic Algorithm for Preterm Labor Prediction Based on Electrohysterography, *Sensors*, vol. 21, no. 10, p. 3350, 2021. DOI: 10.3390/s21103350.

[11] J. Peng *et al.*, Evaluation of electrohysterogram measured from different gestational weeks for recognizing preterm delivery: a preliminary study using random Forest, *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 352–362, 2020, ISSN: 02085216. DOI: 10.1016/j.bbe.2019.12.003.

[12] 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, ISSN: 13616579. DOI: 10.1088/0967-3334/36/2/341.

[13] R. E. Garfield, W. L. Maner, H. Maul, and G. R. Saade, Use of uterine EMG and cervical LIF in monitoring pregnant patients, *BJOG: An International Journal of Obstetrics and Gynaecology*, vol. 112, no. SUPPL. 1, pp. 103–108, 2005, ISSN: 14700328. DOI: 10.1111/j.1471-0528.2005.00596.x.

[14] J. Vrhovec and A. Macek, An Uterine Electromyographic Activity as a Measure of Labor Progression, in Applications of EMG in Clinical and Sports Medicine, Jan. 2012. DOI: 10.5772/25526.

[15] A. Lemancewicz *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, ISSN: 02085216. DOI: 10.1016/j.bbe.2015.11.005.

[16] J. Mas-Cabo *et al.*, Electrohysterogram for ANN-Based Prediction of Imminent Labor in Women with Threatened Preterm Labor Undergoing Tocolytic Therapy, *Sensors*, vol. 20, no. 9, p. 2681, 2020, ISSN: 14248220. DOI: 10.3390/s20092681.

[17] D. Schlembach, W. L. Maner, R. E. Garfield, and H. Maul, Monitoring the progress of pregnancy and labor using electromyography, *European Journal of Obstetrics and Gynecology and Reproductive Biology*, vol. 144, no. SUPPL 1, pp. 2–8, 2009, ISSN: 18727654. DOI: 10.1016/j.ejogrb.2009.02.016.

[18] J. Mas-Cabo, G. Prats-Boluda, Y. Ye-Lin, J. Alberola-Rubio, A. Perales, and J. Garcia-Casado, Characterization of the effects of Atosiban on uterine electromyograms recorded in women with threatened preterm labor, *Biomedical Signal Processing and Control*, vol. 52, pp. 198–205, 2019, ISSN: 17468108. DOI: 10.1016/j.bspc.2019.04.001.

[19] F. Nieto-del-amor, R. Beskhani, Y. Ye-lin, J. Garcia-casado, and A. Diaz-martinez, Assessment of Dispersion and Bubble Entropy Measures for Enhancing Preterm Birth Prediction Based on Electrohysterographic Signals, *Sensors*, vol. 21, no. 18, 2021. DOI: 10.3390/s21186071.

[20] 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 and Physics*, vol. 36, no. 6, pp. 761–767, 2014, ISSN: 18734030. DOI: 10.1016/j.medengphy.2014.01.009.

[21] M. Mischi *et al.*, Dedicated Entropy Measures for Early Assessment of Pregnancy Progression From Single-Channel Electrohysterography, *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 4, pp. 875–884, 2018, ISSN: 0018-9294. DOI: 10.1109/TBME.2017.2723933.

[22] H. Azami, M. Rostaghi, A. Fernandez, and J. Escudero, Dispersion entropy for the analysis of resting-state MEG regularity in Alzheimer's disease, in 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2016, pp. 6417–6420, ISBN: 9781457702204. DOI: 10.1109/EMBC.2016.7592197.

[23] H. Azami and J. Escudero, Amplitude- and fluctuation-based dispersion entropy, *Entropy*, vol. 20, no. 3, p. 210, 2018, ISSN: 10994300. DOI: 10.3390/e20030210.

[24] H. Romero-Morales, J. N. Muñoz-Montes de Oca, R. Mora-Martínez, Y. Mina-Paz, and J. J. Reyes-Lagos, Enhancing classification of preterm-term birth using continuous wavelet transform and entropy-based methods of electrohysterogram signals, *Frontiers in Endocrinology*, vol. 13, no. January, pp. 1–11, 2023, ISSN: 16642392. DOI: 10.3389/fendo.2022.1035615.

[25] 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, ISSN: 00029378. DOI: 10.1016/0002-9378(84)90677-X.

[26] H. Leitich, M. Brunbauer, A. Kaider, C. Egarter, and P. Husslein, Cervical length and dilatation of the internal cervical os detected by vaginal ultrasonography as markers for preterm delivery: A systematic review, *American Journal of Obstetrics and Gynecology*, vol. 181, no. 6, pp. 1465–1472, 1999, ISSN: 00029378. DOI: 10.1016/S0002-9378(99)70407-2.

[27] N. Nader, M. Hassan, W. Falou, M. Khalil, B. Karlsson, and C. Marque, Uterine muscle networks: Connectivity analysis of the ehg during pregnancy and labor, *arXiv: Quantitative Methods*, 2019. DOI: https://doi.org/10.48550/arXiv.1904.05021.

[28] R. C. Young, Myocytes, myometrium, and uterine contractions, *Annals of the New York Academy of Sciences*, vol. 1101, pp. 72–84, 2007, ISSN: 17496632. DOI: 10.1196/annals.1389.038.

[29] A. Diab, S. Boudaoud, B. Karlsson, and C. Marque, Performance comparison of coupling-evaluation methods in discriminating between pregnancy and labor EHG signals, *Computers in Biology and Medicine*, vol. 132, no. August 2020, p. 104 308, 2021, ISSN: 18790534. DOI: 10.1016/j.compbiomed.2021.104308.

[30] R. C. Young and P. Barendse, Linking Myometrial Physiology to Intrauterine Pressure; How Tissue-Level Contractions Create Uterine Contractions of Labor, *PLoS Computational Biology*, vol. 10, no. 10, 2014, ISSN: 15537358. DOI: 10.1371/journal.pcbi.1003850.

[31] F. Nieto-del-Amor *et al.*, Combination of Feature Selection and Resampling Methods to Predict Preterm Birth Based on Electrohysterographic Signals from Imbalance Data, *Sensors*, vol. 22, no. 14, p. 5098, 2022, ISSN: 1424-8220. DOI: 10.3390/s22145098.

[32] L. O. Jimenez and D. A. Landgrebe, Supervised classification in high-dimensional space: Geometrical, statistical, and asymptotical properties of multivariate data, *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 28, no. 1, pp. 39–54, 1998, ISSN: 10946977. DOI: 10.1109/5326.661089.

[33] G. Doquire and M. Verleysen, A comparison of multivariate mutual information estimators for feature selection, in ICPRAM 2012 - Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods, vol. 1, 2012, pp. 176–185, ISBN: 9789898425980. DOI: 10.5220/0003726101760185.

[34] M. Pett, N. Lackey, and J. Sullivan, *Making sense of factor analysis: The use of factor analysis for instrument development in health care research.* 2003. DOI: 10.4135/9781412984898.

[35] J. Xu *et al.*, Network Theory Based EHG Signal Analysis and its Application in Preterm Prediction, *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 7, pp. 2876–2887, 2022, ISSN: 21682208. DOI: 10.1109/JBHI.2022.3140427.

[36] D. Alamedine, M. Khalil, and C. Marque, Comparison of different EHG feature selection methods for the detection of preterm labor, *Computational and Mathematical Methods in Medicine*, vol. 2013, 2013, ISSN: 17486718. DOI: 10.1155/2013/485684.

[37] W. Bouaguel, A New Approach for Wrapper Feature Selection Using Genetic Algorithm for Big Data, in Intelligent and Evolutionary Systems, 2016, pp. 75–83. DOI: 10.1007/978-3-319-27000-5_6.

[38] N. A. Nnamoko, F. N. Arshad, D. England, J. Vora, and J. Norman, Evaluation of Filter and Wrapper Methods for Feature Selection in Supervised Machine Learning, *34th International Conference on Machine Learning, ICML 2017*, no. September, 2014.

[39] A. Cheng *et al.*, Novel Multichannel Entropy Features and Machine Learning for Early Assessment of Pregnancy Progression Using Electrohysterography, *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 12, pp. 3728–3738, 2022, ISSN: 15582531. DOI: 10.1109/TBME.2022.3176668.

[40] B. Remeseiro and V. Bolon-Canedo, A review of feature selection methods in medical applications, *Computers in Biology and Medicine*, vol. 112, no. July, p. 103 375, 2019, ISSN: 18790534. DOI: 10.1016/j.compbiomed.2019.103375.

[41] Voratas Kachitvichyanukul, Comparison of Three Evolutionary Algorithms: GA, PSO, and DE, *Industrial Engineering & Management Systems*, vol. 11, no. 3, pp. 215–223, 2012.

[42] C. Benalcazar-Parra *et al.*, Prediction of Labor Induction Success from the Uterine Electrohysterogram, *Journal of Sensors*, vol. 2019, pp. 1–12, 2019, ISSN: 1687-725X. DOI: 10.1155/2019/6916251.

[43] D. Alamedine, M. Khalil, and C. Marque, Comparison of Feature selection for Monopolar and Bipolar EHG signal, in Journees Recherche en Imagerie et Technologies pour la Santé (RITS 2015), 2015, pp. 100–101.

[44] U. M. Khaire and R. Dhanalakshmi, Stability of feature selection algorithm: A review, *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4, pp. 1060–1073, 2022, ISSN: 22131248. DOI: 10.1016/j.jksuci.2019.06.012.

[45] S. B. Kotsiantis, Feature selection for machine learning classification problems: A recent overview, *Artificial Intelligence Review*, vol. 42, no. 1, p. 157, 2014, ISSN: 02692821. DOI: 10.1007/s10462-011-9230-1.

[46] G. De Lannoy, G. Doquire, D. François, and M. Verleysen, Feature selection for interpatient supervised heart beat classification, *Computational Intelligence and Neuroscience*, vol. 2011, 2011, ISSN: 16875265. DOI: 10.1155/2011/643816.

[47] H. Léman, C. Marque, and J. Gondry, Use of the electrohysterogram signal for characterization of contractions during pregnancy, *IEEE Transactions on Biomedical Engineering*, vol. 46, no. 10, pp. 1222–1229, 1999, ISSN: 00189294. DOI: 10.1109/10.790499.

[48] 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, ISSN: 00297844. DOI: 10.1016/S0029-7844(03)00341-7.

[49] H. A. Frey and M. A. Klebanoff, The epidemiology, etiology, and costs of preterm birth, *Seminars in Fetal and Neonatal Medicine*, vol. 21, no. 2, pp. 68–73, 2016, ISSN: 18780946. DOI: 10.1016/j.siny.2015.12.011.

[50] M. A. Khaldy, Resampling Imbalanced Class and the Effectiveness of Feature Selection Methods for Heart Failure Dataset, *International Robotics & Automation Journal*, vol. 4, no. 1, 2018. DOI: 10.15406/iratj.2018.04.00090.

[51] G. Vandewiele *et al.*, Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling, *Artificial Intelligence in Medicine*, vol. 111, p. 101 987, 2021, ISSN: 18732860. DOI: 10.1016/j.artmed.2020.101987. arXiv: 2001.06296.

[52] Y. Liu, Y. Wang, X. Ren, H. Zhou, and X. Diao, A Classification Method Based on Feature Selection for Imbalanced Data, *IEEE Access*, vol. 7, pp. 81 794–81 807, 2019, ISSN: 21693536. DOI: 10.1109/ACCESS.2019.2923846.

[53] F. Namous, H. Faris, A. A. Heidari, M. Khalafat, R. S. Alkhawaldeh, and N. Ghatasheh, *Evolutionary and Swarm-Based Feature Selection for Imbalanced Data Classification*. 2020, pp. 231–250, ISBN: 9789813299900. DOI: 10.1007/978-981-32-9990-0_11.

[54] R. Martín-Félez and R. A. Mollineda, On the suitability of combining feature selection and resampling to manage data complexity, in Conference of the Spanish Association for Artificial Intelligence, vol. 5988 LNAI, 2009, pp. 141–150, ISBN: 364214263X. DOI: 10.1007/978-3-642-14264-2_15.

[55] A. Jain, S. Ratnoo, and D. Kumar, Addressing class imbalance problem in medical diagnosis: A genetic algorithm approach, in 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC), 2018, pp. 1–8, ISBN: 9781509063130. DOI: 10.1109/ICOMICON.2017.8279150.

[56] W. Pei, B. Xue, L. Shang, and M. Zhang, New Fitness Functions in Genetic Programming for Classification with High-dimensional Unbalanced Data, *2019 IEEE Congress on Evolutionary Computation, CEC 2019 - Proceedings*, vol. 42, no. 2, pp. 2779–2786, 2019. DOI: 10.1109/CEC.2019.8789974.

[57] T. Włodarczyk *et al.*, Machine learning methods for preterm birth prediction: A review, *Electronics (Switzerland)*, vol. 10, no. 5, pp. 1–24, 2021, ISSN: 20799292. DOI: 10.3390/electronics10050586.

[58] M. U. Khan, S. Aziz, S. Ibraheem, A. Butt, and H. Shahid, Characterization of Term and Preterm Deliveries using Electrohysterograms Signatures, *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2019*, pp. 899–905, 2019. DOI: 10.1109/IEMCON.2019.8936292.

[59] J. Mas-Cabo, G. Prats-Boluda, J. Garcia-Casado, J. Alberola-Rubio, A. Perales, and Y. Ye-Lin, Design and Assessment of a Robust and Generalizable ANN-Based Classifier for the Prediction of Premature Birth by means of Multichannel Electrohysterographic Records, *Journal of Sensors*, vol. 2019, pp. 1–13, 2019, ISSN: 16877268. DOI: 10.1155/2019/5373810.

[60] J. Xu, Z. Chen, J. Zhang, Y. Lu, X. Yang, and A. Pumir, Realistic preterm prediction based on optimized synthetic sampling of EHG signal, *Computers in Biology and Medicine*, vol. 136, no. April, p. 104 644, 2021, ISSN: 18790534. DOI: 10.1016/j.compbiomed.2021.104644.

[61] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, *Electronics (Switzerland)*, vol. 8, no. 8, 2019, ISSN: 20799292. DOI: 10.3390/electronics8080832.

[62] K. P. Murphy, *Machine Learning: A Probabilistic Perspective.* 2012, pp. 1–100, ISBN: 978026208029.

[63] Dietterich Thomas G., An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization, *Machine Learning*, vol. 40, no. 2, pp. 139–157, 2000.

[64] I. K. Ludmila and J. W. Christopher, Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy, *Machine Learning*, vol. 51, no. 2, pp. 181–207, 2003.

[65] K. M. Ting and I. H. Witten, Issues in stacked generalization, *Journal of Artificial Intelligence Research*, vol. 10, pp. 271–289, 1999, ISSN: 10769757. DOI: 10.1613/jair.594.

[66] 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, e0132116, 2015, ISSN: 19326203. DOI: 10.1371/journal.pone.0132116.

[67] I. O. Idowu *et al.*, Artificial intelligence for detecting preterm uterine activity in gynacology and obstertric care, *Proceedings - 15th IEEE International Conference on Computer and Information Technology, CIT 2015, 14th IEEE International Conference on Ubiquitous Computing and Communications, IUCC 2015, 13th IEEE International Conference on Dependable, Autonomic and Se*, pp. 215–220, 2015. DOI: 10.1109/CIT/IUCC/DASC/PICOM.2015.31.

[68] S. Dudoit and J. Fridlyand, Bagging to improve the accuracy of a clustering procedure, *Bioinformatics*, vol. 19, no. 9, pp. 1090–1099, 2003, ISSN: 13674811. DOI: 10.1093/bioinformatics/btg038.

[69] R. Ghaemi, N. Sulaiman, H. Ibrahim, and N. Mustapha, A survey: Clustering ensembles techniques, *World Academy of Science, Engineering and Technology*, vol. 38, no. February, pp. 644–653, 2009, ISSN: 2010376X.

[70] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, and J. Santos, Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches, *IEEE Computational Intelligence Magazine*, vol. 13, no. 4, pp. 59–76, 2018, ISSN: 15566048. DOI: 10.1109/MCI.2018.2866730.

[71] U. Goldsztejn and A. Nehorai, Predicting preterm births from electrohysterogram recordings via deep learning, pp. 1–18, 2022.

[72] V. Selvaraju, P. A. Karthick, and R. Swaminathan, Analysis of frequency bands of uterine electromyography signals for the detection of preterm birth, *Public Health and Informatics: Proceedings of MIE 2021*, vol. 0, pp. 283–287, 2021. DOI: 10.3233/SHTI210165.

[73] H. Lou, H. Liu, Z. Chen, Z. Zhen, B. Dong, and J. Xu, Bio-process inspired characterization of pregnancy evolution using entropy and its application in preterm birth detection, *Biomedical Signal Processing and Control*, vol. 75, no. January, p. 103 587, 2022, ISSN: 17468108. DOI: 10.1016/j.bspc.2022.103587.

[74] L. A. Jeni, J. F. Cohn, and F. De La Torre, Facing imbalanced data - Recommendations for the use of performance metrics, in 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, 2013, pp. 245–251, ISBN: 9780769550480. DOI: 10.1109/ACII.2013.47.

[75] T. Saito and M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, *PLoS ONE*, vol. 10, no. 3, pp. 1–21, 2015, ISSN: 19326203. DOI: 10.1371/journal.pone.0118432.

[76] S. Vluymans, Learning from imbalanced data, *Studies in Computational Intelligence*, vol. 807, no. 9, pp. 81–110, 2019, ISSN: 1860949X. DOI: 10.1007/978-3-030-04663-7_4.

[77] Y. Yuan, W. Su, and M. Zhu, Threshold-Free Measures for Assessing the Performance of Medical Screening Tests, *Frontiers in Public Health*, vol. 3, no. April, 2015, ISSN: 2296-2565. DOI: 10.3389/fpubh.2015.00057.

[78] J. Davis and M. Goadrich, The relationship between precision-recall and ROC curves, *ACM International Conference Proceeding Series*, vol. 148, pp. 233–240, 2006. DOI: 10.1145/1143844.1143874.

[79] H. R. Sofaer, J. A. Hoeting, and C. S. Jarnevich, The area under the precision-recall curve as a performance metric for rare binary events, *Methods in Ecology and Evolution*, vol. 10, no. 4, pp. 565–577, 2019, ISSN: 2041210X. DOI: 10.1111/2041-210X.13140.

[80] Y. Sun, A. K. Wong, and M. S. Kamel, Classification of imbalanced data: A review, *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 4, pp. 687–719, 2009, ISSN: 02180014. DOI: 10.1142/S0218001409007326.

[81] F. Jager, S. Libenšek, and K. Geršak, Characterization and automatic classification of preterm and term uterine records, *PLoS ONE*, vol. 13, no. 8, O. Uthman, Ed., e0202125, 2018, ISSN: 19326203. DOI: 10.1371/journal.pone.0202125.

[82] K. Zhang *et al.*, An interpretable and expandable deep learning diagnostic system for multiple ocular diseases: Qualitative study, *Journal of Medical Internet Research*, vol. 20, no. 11, pp. 49–50, 2018, ISSN: 14388871. DOI: 10.2196/11144.

[83] P. Tewari, Artificial Intelligence and the Internet of Things in Water Management, *Advanced Water Technologies*, vol. 8, no. January, pp. 235–239, 2020. DOI: 10.1201/9781315101514-10.

[84] R. Aggarwal *et al.*, Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis, *npj Digital Medicine*, vol. 4, no. 1, 2021, ISSN: 23986352. DOI: 10.1038/s41746-021-00438-z.

[85] N. Artrith *et al.*, Best practices in machine learning for chemistry, *Nature Chemistry*, vol. 13, no. 6, pp. 505–508, 2021, ISSN: 17554349. DOI: 10.1038/s41557-021-00716-z.

[86] J. Canny, H. Zhao, B. Jaros, Y. Chen, and J. Mao, Machine learning at the limit, *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, pp. 233–242, 2015. DOI: 10.1109/BigData.2015.7363760.

[87] H. Allahem and S. Sampalli, Automated labour detection framework to monitor pregnant women with a high risk of premature labour using machine learning and deep learning, *Informatics in Medicine Unlocked*, vol. 28, 2022, ISSN: 23529148. DOI: 10.1016/j.imu.2021.100771.

[88] D. Hao, J. Peng, Y. Wang, J. Liu, X. Zhou, and D. Zheng, Evaluation of convolutional neural network for recognizing uterine contractions with electrohysterogram, *Computers in Biology and Medicine*, vol. 113, no. August, p. 103 394, 2019, ISSN: 18790534. DOI: 10.1016/j.compbiomed.2019.103394.

[89] A. Bansal, R. Sharma, and M. Kathuria, A Systematic Review on Data Scarcity Problem in Deep Learning: Solution and Applications, *ACM Computing Surveys*, vol. 54, no. 10, 2022, ISSN: 15577341. DOI: 10.1145/3502287.

[90] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, Transfer learning for medical image classification: a literature review, 2022. DOI: 10.1186/s12880-022-00793-7.

[91] G. Pinto, Z. Wang, A. Roy, T. Hong, and A. Capozzoli, Transfer learning for smart buildings: A critical review of algorithms, applications, and future perspectives, *Advances in Applied Energy*, vol. 5, no. November 2021, p. 100 084, 2022, ISSN: 26667924. DOI: 10.1016/j.adapen.2022.100084.

[92] G. Prats-Boluda *et al.*, Optimization of imminent labor prediction systems in women with threatened preterm labor based on electrohysterography, *Sensors*, vol. 21, no. 7, pp. 1–18, 2021, ISSN: 14248220. DOI: 10.3390/s21072496.

[93] F. Fuchs and M. V. Senat, Multiple gestations and preterm birth, *Seminars in Fetal and Neonatal Medicine*, vol. 21, no. 2, pp. 113–120, 2016, ISSN: 18780946. DOI: 10.1016/j.siny.2015.12.010.

[94] E. Jurado-garcía, A. Botello-hermosa, F. J. Fernández-carrasco, J. Gómez-salgado, N. Navas-rojano, and R. Casado-mejía, Multiple gestations and assisted reproductive technologies: Qualitative study of the discourse of health professionals in spain, *International Journal of Environmental Research and Public Health*, vol. 18, no. 11, pp. 1–15, 2021, ISSN: 16604601. DOI: 10.3390/ijerph18116031.

[95] S. Stock and J. Norman, Preterm and term labour in multiple pregnancies, *Seminars in Fetal and Neonatal Medicine*, vol. 15, no. 6, pp. 336–341, 2010, ISSN: 1744165X. DOI: 10.1016/j.siny.2010.06.006.

[96] F. Lyall, S. J. Lye, T. G. Teoh, F. Cousins, G. Milligan, and S. C. Robson, Expression of Gsα, connexin-43, connexin-26, and EP1, 3, and 4 receptors in myometrium of prelabor singleton versus multiple gestations and the effects of mechanical stretch and steroids on Gsα, *Journal of the Society for Gynecologic Investigation*, vol. 9, no. 5, pp. 299–307, 2002, ISSN: 10715576. DOI: 10.1016/S1071-5576(02)00175-2.

[97] I. Grigorescu, L. Cordero-Grande, A. D. Edwards, J. Hajnal, M. Modat, and M. Deprez, Interpretable Convolutional Neural Networks for Preterm Birth Classification, pp. 1–4, 2019. arXiv: 1910.00071.

[98] A. S. Moerdijk *et al.*, Fetal MRI of the heart and brain in congenital heart disease, *The Lancet Child and Adolescent Health*, vol. 7, no. 1, pp. 59–68, 2023, ISSN: 23524642. DOI: 10.1016/S2352-4642(22)00249-8.

[99] V. Berghella, E. Hayes, J. Visintine, and J. K. Baxter, Fetal fibronectin testing for reducing the risk of preterm birth, *Cochrane Database of Systematic Reviews*, no. 4, 2008, ISSN: 1469493X. DOI: 10.1002/14651858.CD006843.pub2.

[100] R. E. Garfield and W. L. Maner, Physiology and electrical activity of uterine contractions, *Seminars in Cell and Developmental Biology*, vol. 18, no. 3, pp. 289–295, 2007, ISSN: 10849521. DOI: 10.1016/j.semcdb.2007.05.004.

[101] M. Pandey, M. Chauhan, and S. Awasthi, Interplay of cytokines in preterm birth, *Indian Journal of Medical Research*, vol. 146, no. September, pp. 316–327, 2017, ISSN: 09715916. DOI: 10.4103/ijmr.IJMR_1624_14.

[102] M. Sean Esplin *et al.*, Predictive accuracy of serial transvaginal cervical lengths and quantitative vaginal fetal fibronectin levels for spontaneous preterm birth among nulliparous women, *Obstetrical and Gynecological Survey*, vol. 72, no. 7, pp. 397–399, 2017, ISSN: 15339866. DOI: 10.1097/OGX.0000000000000455.

[103] C. Leung, Born too soon, *Neuroendocrinology Letters*, vol. 25, no. SUPPL. 1, J. L. CP Howson, MV Kinney, Ed., pp. 133–136, 2004, ISSN: 0172780X. DOI: 10.2307/3965140.

Chapter 7

Conclusions

This chapter discusses the conclusions reached and evaluates whether they represent a novel contribution, in accordance with the specific objectives set out in Chapter 2.

Objective 1. To extract relevant features from EHG signals to discriminate the preterm versus term labor in women undergoing regular prenatal check-ups

We assessed the ability of temporal, spectral, and non-linear features to discriminate between preterm and term labor cases. Our results support previous results in the literature regarding the fact that spectral features are more reliable than amplitude-related features in distinguishing between these two groups. We also confirmed the suitability of classical non-linear features (sample entropy, fuzzy entropy, spectral entropy, Lempel Ziv, spectral entropy, Poincaré ellipse plot derivates, time reversibility and Kazt's fractal dimension) for differentiating between preterm and term deliveries.

For the first time, we used dispersion and bubble entropy to characterize EHG signals. Both novel entropy measures effectively distinguished between term and preterm cases and outperformed even the classical entropy metrics. Also, bubble entropy surpassed dispersion entropy and is seen as a promising feature for identifying preterm labor.

Objective 2. To determine the complementary, redundant and noisy information of EHG features to optimize the feature subspace for predicting preterm delivery

We have demonstrated the ability of the genetic algorithm to obtain an optimum feature subspace while maximizing complementary information and rejecting redundant or noisy features.

The optimal feature subset mainly depends on the classification algorithm, giving rise to different prediction models with a similar performance, but inferior to that of ensemble method, such as the majority voting strategy.

We also confirmed that resampling before or after partitioning has a large impact on the complexity of the classification task. The resampling-partitioning approach facilitates classification due to the information leakage between the training and test datasets, leading to a lower-dimensional optimal feature subset. On the other hand, the partitioning-resampling pipeline presents a more complex challenge, as reflected in a high dimensional optimized feature subset.

Objective 3. To assess different imbalanced data learning strategies to achieve a robust and generalizable preterm birth prediction system

In this thesis, our findings further support the claim that balancing the database by generating synthetic samples of preterm delivery records before splitting the database into training and test partitions can result in unrealistic outcomes. The resampling-partitioning scheme reduces the classification task complexity and overestimates the generalization capability. It is essential to use partitioning-resampling to maintain completely "new" and "unseen" testing samples for accurate model performance evaluation.

We have also confirmed the feasibility of combining different resampling methods in feature selection and LDA as a simple and easy-to-interpret classifier algorithm to deal with the imbalanced data problem in preterm labor prediction. We found the best strategy to mitigate imbalanced data consisted of balancing the validation dataset to 1:1 by using undersampling methods.

This predictive model surpasses the results of studies in the literature that used a partitioning-resampling pipeline almost comparable with those obtained by the resampling-partitioning approach, reaching an AUC of 94% and AP of 84%. The maximum F1-score is about 80% and the recall is about 80%. When we maximized the mean of G-mean, the recall of the best model was about 90% and the F1-score was about 72%.

This approach paves the way for the development of an integral preterm birth prediction system based on the EHG technique for use in clinics that would allow patient-oriented strategies to be designed for better preterm labor prevention and improve maternal-fetal well-being, besides the optimal management of hospital resources.