Integrating Machine Learning Techniques and Physiology Based Heart Rate Features for Antepartum Fetal Monitoring

Maria G. Signorini[1,*], Nicolò Pini[1], Alberto Malovini[2], Riccardo Bellazzi[3], and Giovanni Magenes[3]

mariagabriella.signorini@polimi.it, nicolo.pini@polimi.it, alberto.malovini@unipv.it, riccardo.bellazzi@unipv.it, giovanni.magenes@unipv.it

* Corresponding author, corresponding author email: mariagabriella.signorini@polimi.it

[1] Department of Electronics, Information and Bioengineering (DEIB), Politecnico Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy

[2] IRCCS Fondazione S. Maugeri, Via Maugeri 10, 27100, Pavia, Italy

[3] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Via Ferrata 5, 27100, Pavia, Italy

ABSTRACT

Background and Objectives: Intrauterine Growth Restriction (IUGR) is a fetal condition defined as the abnormal rate of fetal growth. The pathology is a documented cause of fetal and neonatal morbidity and mortality. In clinical practice, diagnosis is assessed at birth and may only be suspected during pregnancy. Therefore, designing an accurate model for the early and prompt identification of pathology in the antepartum period is crucial in view of pregnancy management.

Methods: We tested the performance of 15 machine learning techniques in discriminating healthy versus IUGR fetuses. The various models were trained with a set of 12 physiology based heart rate features extracted from a single antepartum CardioTocographic (CTG) recording. The reason for the utilization of time, frequency, and nonlinear indices is based on their standalone documented ability to describe several physiological and pathological fetal conditions.

Results: We validated our approach on a database of 60 healthy and 60 IUGR fetuses. The machine learning methodology achieving the best performance was Random Forests. Specifically, we obtained a mean classification accuracy of 0.911 [0.860, 0.961 (0.95 confidence interval)] averaged over 10 test sets (10 Fold Cross Validation). Similar results were provided by Classification Trees, Logistic Regression, and Support Vector Machines. A features ranking procedure highlighted that nonlinear indices showed the highest capability to discriminate between the considered fetal conditions. Nevertheless, is the combination of features investigating CTG signal in different domains, that contributes to an increase in classification accuracy.

Conclusions: We provided validation of an accurate artificially intelligence framework for the diagnosis of IUGR condition in the antepartum period. The employed physiology based heart rate features constitute an interpretable link between the machine learning results and the quantitative estimators of fetal wellbeing.

***Keywords*— Machine learning and statistical models; Fetal Heart Rate monitoring; Predictive Analytics; Physiology-based features; Multivariate analysis;**

1. INTRODUCTION

Nowadays, antepartum fetal monitoring is a routine methodology adopted in clinical practice to assess fetal wellbeing throughout pregnancy, mainly in the context of pathological fetal state identification [1,2]. The most used technique consists in recording the Fetal Heart Rate (FHR) by means of the CardioTocography (CTG) [3]. The rationale for its utilization relies on the fact that it has been extensively shown how FHR changes can anticipate and/or even predict fetal distress as well as adverse conditions before the insurgence of any other symptom [4].

CTG analysis has been progressively shifting from pure visual observation of the traces to its computerized version [5], which consists of extracting various quantitative parameters associated with fetal conditions [6,7]. Morphological [8], frequency [7,9], and nonlinear/complexity indices [10–13] are usually thought to summarize the various pathophysiological aspects of FHR.

Despite the large availability of FHR quantitative indicators, a very limited portion of fetal-related literature addresses the investigation of fetal surveillance by means of multivariate approaches. If this latter consideration was to be attributed to scarce data availability in the past, recent years have seen the endless growth of data generated during patients' care path [14]. Additionally, the technological advancements in parallel with novel parameters contributed to an increase in the amount of available data related to fetal monitoring [15].

As a result, if adding more measurements could hopefully contribute to better insights into pathophysiological systems, inevitability it increases the complexity of data analysis as well as the interpretation of the extracted results. Machine learning methodologies appear as a possible solution to this issue, as they can face large and complex datasets [15,16]. However, it is also to be underlined that when a subset of features is automatically extracted from a large amount of data, the interpretation of the results is usually difficult to be linked to the a priori knowledge of the underlying physiological mechanisms.

In the presented study, we designed a two-step methodology for the early identification of a pathological fetal state, namely: Intrauterine Growth Restriction (IUGR). The implementation was achieved by deriving features from a single antepartum CTG trace by means of advanced signal analytics. Subsequently, various machine learning techniques were trained with the extracted FHR features. The rationale for employing such physiology based heart rate features aimed to realize a tool capable of providing an interpretable link between the machine learning results and the physiological mechanisms of fetal regulation. Moreover, the specification of early identification is achieved by removing the influence of gestational age (GA) at which

the available traces were acquired, thus providing a reliable and effective set of tools for the antenatal IUGR discrimination.

As a proof of concept of an impactful and clinically relevant application of artificial intelligence in the field of fetal monitoring, in this paper we compared the validity and performances of several machine learning techniques for the classification of healthy fetuses versus fetuses affected by IUGR. The former pathology along with small for gestational age (SGA) represent the second cause of perinatal mortalities, contributing to 52% of stillbirths [17]. Moreover, the IUGR condition has been extensively reported as affecting perinatal and postnatal development under several different aspects [18].

As reported in [19], the key point in IUGR management is the early identification of the pathology to the aim of improving both the time setting and the management of delivery. Unfortunately, methodologies towards a reliable and timely detection of IUGR condition are still pending, to the point where the assessment can only be performed at birth [19]. As a consequence, the overall outcome of IUGR babies has not changed much over time [19]. The crucial challenge which is yet to be addressed is aimed to develop reliable tools which ideally would be able to provide antenatal identification of IUGR condition, starting from the available and clinically recorded data.

2. Materials and Methods

2.1 Data collection and subject selection

In a collaboration framework among the Ob-Gyn Clinics at the Azienda Ospedaliera Universitaria Federico II, Napoli, Italy, Biomedical Engineering Labs of Politecnico di Milano, Italy, and Università di Pavia, Italy, FHR traces were collected in a large population of pregnant women.

Among the available CTG recordings, we asked clinicians to select 120 CTG recordings: 60 Healthy and 60 IUGR fetuses. The left-hand side of Fig. 1 displays a 30-minutes segment for a healthy (top) and an IUGR (bottom) CTG traces. The prenatal fetal condition for each subject was verified after delivery to confirm group membership previously assessed at the CTG timepoint. Healthy fetuses at birth presented the following characteristics: weight and abdominal circumference ±10% with respect to the normative ranges and Apgar score = 10. IUGRs were identified based on anamnesis and showed weight below the 10^{th} percentile for the corresponding GA, abdominal circumference below the 10^{th} percentile, and Apgar score < 8. The average duration of the FHR tracing was > 30 minutes for both healthy fetuses and IUGRs to contain both activity and quiet periods of the fetus. FHR recordings were collected in a controlled clinical environment, with the pregnant woman lying on a bed during the standard protocol of non-stress test. The average GA at CTG measurement for healthy fetuses was 34.78±0.53 weeks (Inter Quartile Range (IQR) = 34-35) whereas for IUGR fetuses was 32.27±2.79 weeks (IQR = 30-34). The reason for a nonoverlap in terms of GA between

Figure 1

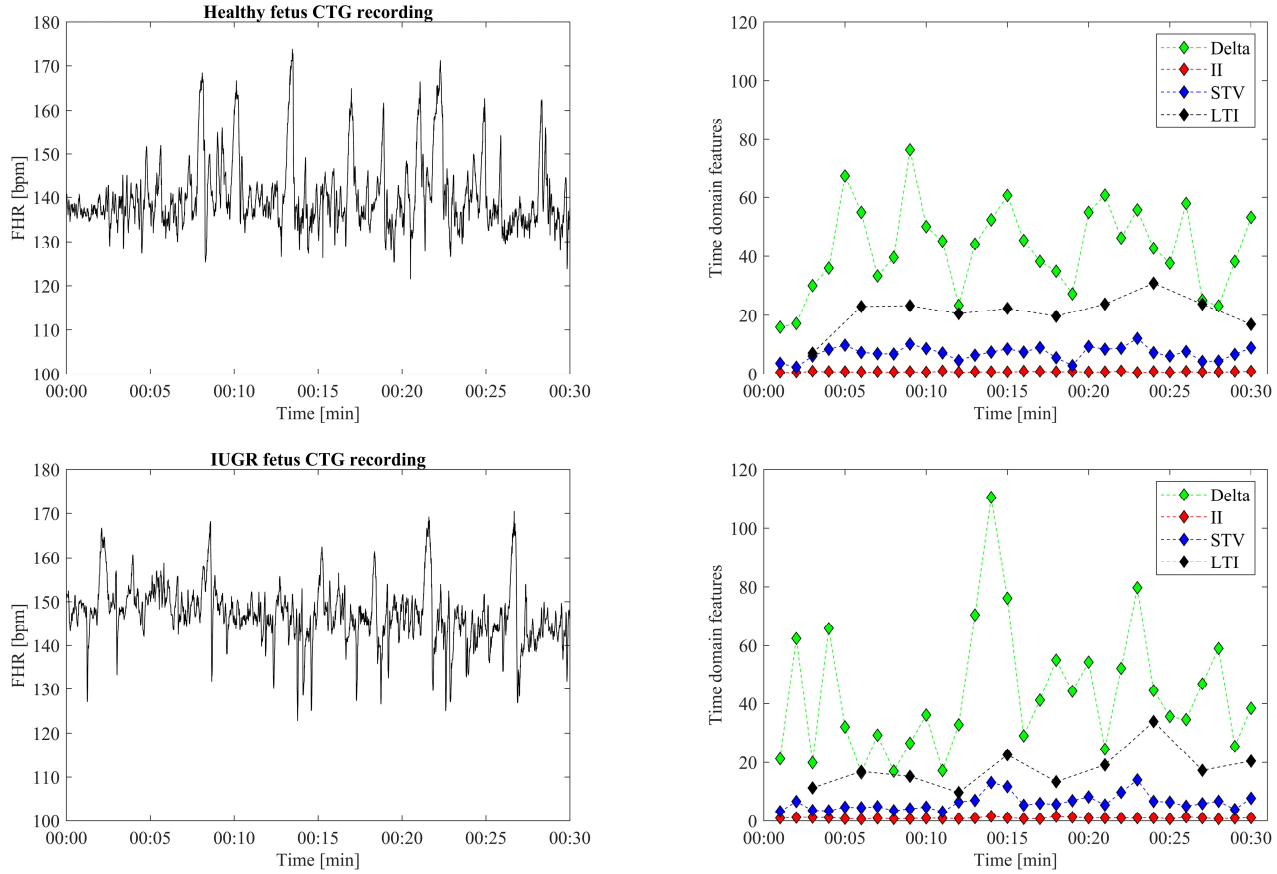

Fig. 1. Left panel display two 30-minutes CTG recordings of a healthy (top) and an IUGR (bottom) subjects respectively. On the right-hand side, the corresponding time series of time domain features are shown. Delta, II, and STV are computed by considering 1-minute window, thus resulting in 30 estimates throughout the reported recording. On the other hand, a 3-minutes window is employed in STV computations, thus 10 values are obtained.

the two groups relies on the fact that in clinical practice normal pregnancies are usually monitored only after the 33rd week of gestation whereas earlier assessments are usually available when considering suspected IUGR cases.

2.2 Selection and statistical preprocessing of features

In previous works, we approached the identification of IUGR fetuses by means of various FHR-based encompassing time domain, frequency, and nonlinear domains.

Time domain indices were computed as suggested by Arduini et al. [8]: Delta, Interval Index (II), Short Term Variability (STV), and Long Term Irregularity (LTI). The former three parameters were computed dividing the signal in windows of length equal to 60 s, LTI in windows of length equal to 180 s.

The frequency content of the FHR signal was analyzed by means of Power Spectral Density (PSD) [20]. This technique provide the power associated with specific frequency components of FHR, as described in detail in [7]. In this study, the power associated to Low Frequency band (LF_pow) is computed in the

frequency range (0.03-0.15 Hz), Movement Frequency power (MF_pow) in the frequency range (0.15-0.5 Hz), High Frequency power (HF_pow) in the frequency range (0.5-1 Hz). Powers in the different bands have also been combined to extract the ratio LF/(MF+HF).

Regarding nonlinear features of the FHR signal, we estimated Approximate Entropy (ApEn) [10], Lempel Ziv Complexity (LZC) with binary alphabet [21], and Phase Rectified Signal Average (PRSA) features [22]. Nonlinear FHR measures were computed considering nonoverlapping windows of length equal to 180 s, with the exception of PRSA features, namely APRS and DPRS which consist of a single estimate since they are computed based on the entire recorded CTG trace. A detailed and more extensive description of the computed indices is reported in the Supplementary Materials and in the Data Brief Article.

The complete procedure of parameter extraction produced N=12 indices, 10 of which are extracted by averaging the corresponding time series (extracted by subdividing the FHR recording in windows), and 2 of them are global parameters computed considering the whole recording at once. The right-hand side of Fig. 1 displays an example of time domain parameters. The length of each series is equal to the number of available windows (number of acceptable intervals) in the original CTG trace after performing the quality assessment. As reported in [7], the majority of FHR parameters can noticeably vary depending on the fetal state (quiet or activity). In order to reduce such intrasubject source of variability and considering that fetal state annotation cannot be performed routinely in clinical practice, the average of the parameters of each time series was calculated. This approach is justified by the fact that our database contains recordings with both activity and quiet periods.

The reason behind the selection of this restricted subset of features relies on their individual peculiar ability in discriminating IUGRs and normal fetuses as described in the following. To summarize: the a priori knowledge parameters to be employed as the starting set of features for further analysis are: Delta, II, STV, LTI (time domain) [8]; LF_pow, MF_pow, HF_pow, LF/(MF+HF) (frequency domain) [7]; ApEn(1, 0.1) [10], LZC(2, 0) , Acceleration Phase Rectified Slope (APRS) and Deceleration Phase Rectified Slope (DPRS) (nonlinear domain) [23].

Parameters employed in the computation of ApEn were m=1 and r=0.1 thus resulting in the feature ApEn(1, 0.1). LZC was computed within a binary approach, having the factor value (p) set to zero, the computed quantity is reported as LZC(2, 0). Additional information regarding LZC applied to FHR analysis may be found in [24]. The last nonlinear technique employed to investigate FHR was the so-called PRSA method, introduced by Bauer [22]. In this context, Acceleration Phase Rectified Slope (APRS) and Deceleration Phase Rectified Slope (DPRS) were computed as reported in [25].

All preprocessing operations on the extracted features were performed by R, a free software environment for statistical computing [26]. A very limited portion of the total number of subjects (8 out of 120) presented some missing features (the percentage of missing features is equal to 1.5% of the total number of features). In order to account for features missingness, we employed the R package missForest [27]. It is suitable to be used in the case of mixed-type data. The imputation procedure is based on the training of a random forest which is capable of predicting the missingness based on the observed and available data [27].

The majority of extracted features showed evidence of intermediate correlation accordingly to the definition by Cohen [28] ($0.30 < |\text{Spearman's Rank Correlation coefficient } (\rho)| < 0.50$) with the GA at which trace was acquired (GA_CTG). This assumption stands considering both the whole cohort but even when limiting the analysis to the considered IUGR population as reported in Table I.

TABLE I
SPEARMAN'S RANK CORRELATION COEFFICIENT BETWEEN EXTRACTED PARAMETERS AND GA_CTG — UNADJUSTED (U) AND ADJUSTED (A) DISTRIBUTIONS
* indicates statistically significant correlation $p < 0.05$

	Correlation Coefficient ρ					
	Unadjusted (U)			**Adjusted (A)**		
Parameter	**Overall**	**Healthy**	**IUGR**	**Overall**	**Healthy**	**IUGR**
Delta	0.3295 *	-0.1000	0.2128	0.0117	-0.0184	-0.0339
II	0.0024	0.0286	-0.1455	0.0174	0.0061	-0.0412
STV	0.4170 *	-0.0388	0.3505 *	0.0178	-0.0388	-0.0399
LTI	0.3030 *	-0.1266	0.0937	0.0103	-0.0429	-0.0101
LF_pow	0.1684	-0.0265	-0.1254	0.0393	0.0143	0.0052
MF_pow	-0.0286	-0.0122	0.1120	-0.0017	0.0347	-0.0124
HF_pow	-0.2407 *	-0.0041	-0.0732	-0.0544	-0.0408	-0.0049
LF/(MF+HF)	0.1684	-0.0265	-0.1254	0.0498	-0.0653	-0.0092
ApEn(1, 0.1)	0.2313 *	0.0551	0.1527	0.0083	0.0245	-0.0288
LZC(2, 0)	0.2310 *	-0.2799 *	0.0204	0.0839	-0.1556	0.0429
APRS	0.4112 *	0.0408	0.2502	0.0147	0.0184	-0.0283
DPRS	-0.4896 *	-0.0408	-0.3478	-0.0152	0.0163	0.0087

Therefore, to address the dependence of measures with respect to time of the assessment, all variables were adjusted using a Robust Linear Regression (RLR) and the derived residuals were employed in the further analyses to provide machine learning classifier with features independent from the GA at which the CTG traces were recorded. Results of correlation between parameters (unadjusted and adjusted) and GA_CTG are reported in Table I. As an example, distribution of Unadjusted (U) and Adjusted (A) of four of the further employed covariates are shown in Fig. 2 along with the corresponding regression lines.

Figure 2

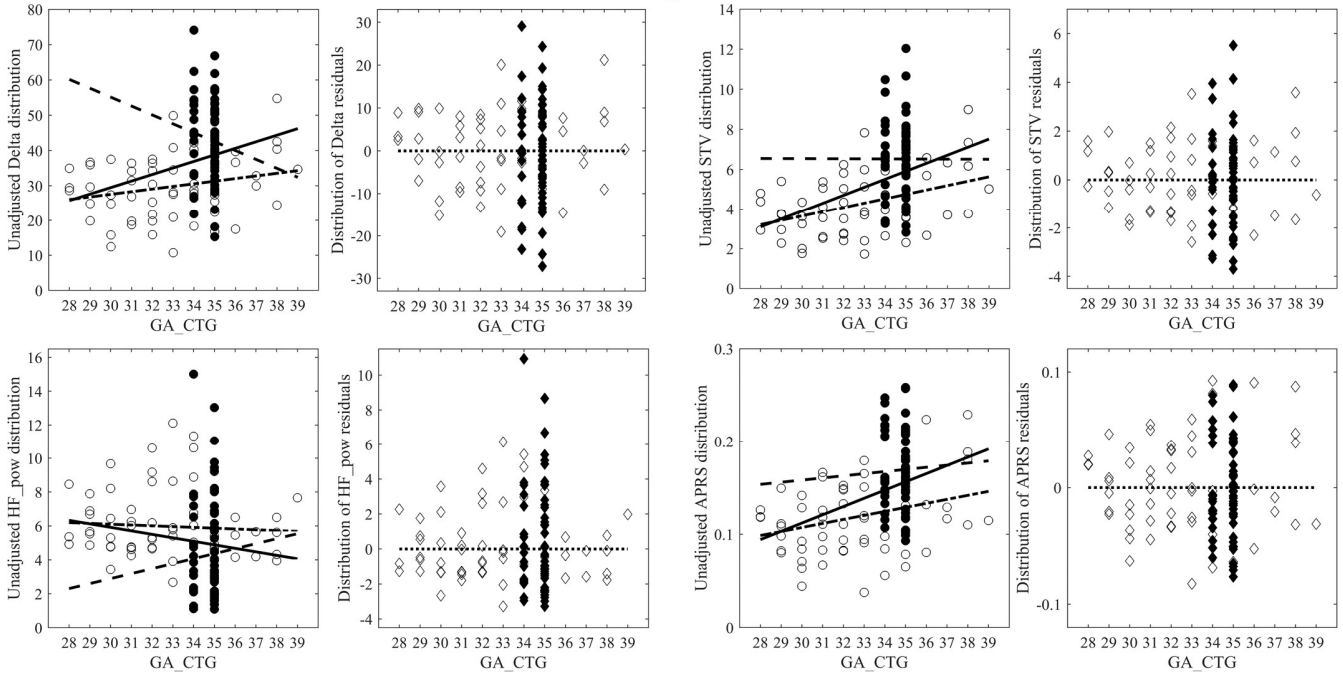

Fig. 2. Scatterplots showing Delta, STV, HF_pow, APRS distributions as a function of GA_CTG. The unadjusted-distribution graphs show the regression lines for Healthy (dashed), IUGR (twodash), and overall population (solid) for the U set. The adjusted-distribution graphs display the A set distribution and the derived regression lines (dotted). In the latter case, the absence of any trend (after performing RLR) is reflected in a single regression line (dotted) for the three distributions.

2.3 Multivariate Analysis

In this investigation, we deepened the preliminary results obtained by analyzing the same database analyzed in [29]. While our previous work was mainly focused on the comparison between the performances of univariate versus multivariate classifiers, in this paper we investigate within a more detailed approach the possible influence of GA over the performances of machine learning methodologies and their feature robustness and insensitivity to GA_CTG. Moreover, the more precisely conducted analysis on feature space will provide validation for the utilization of physiology based heart rate features for the early identification of IUGR pathology. As a general consideration, multivariate analysis was designed to search for an optimal decision rule in the multidimensional space of the parameters to predict the class of interest, namely healthy versus IUGR. A complete roadmap from CTG signal to binary classification is depicted in Fig. 3.

Figure 3

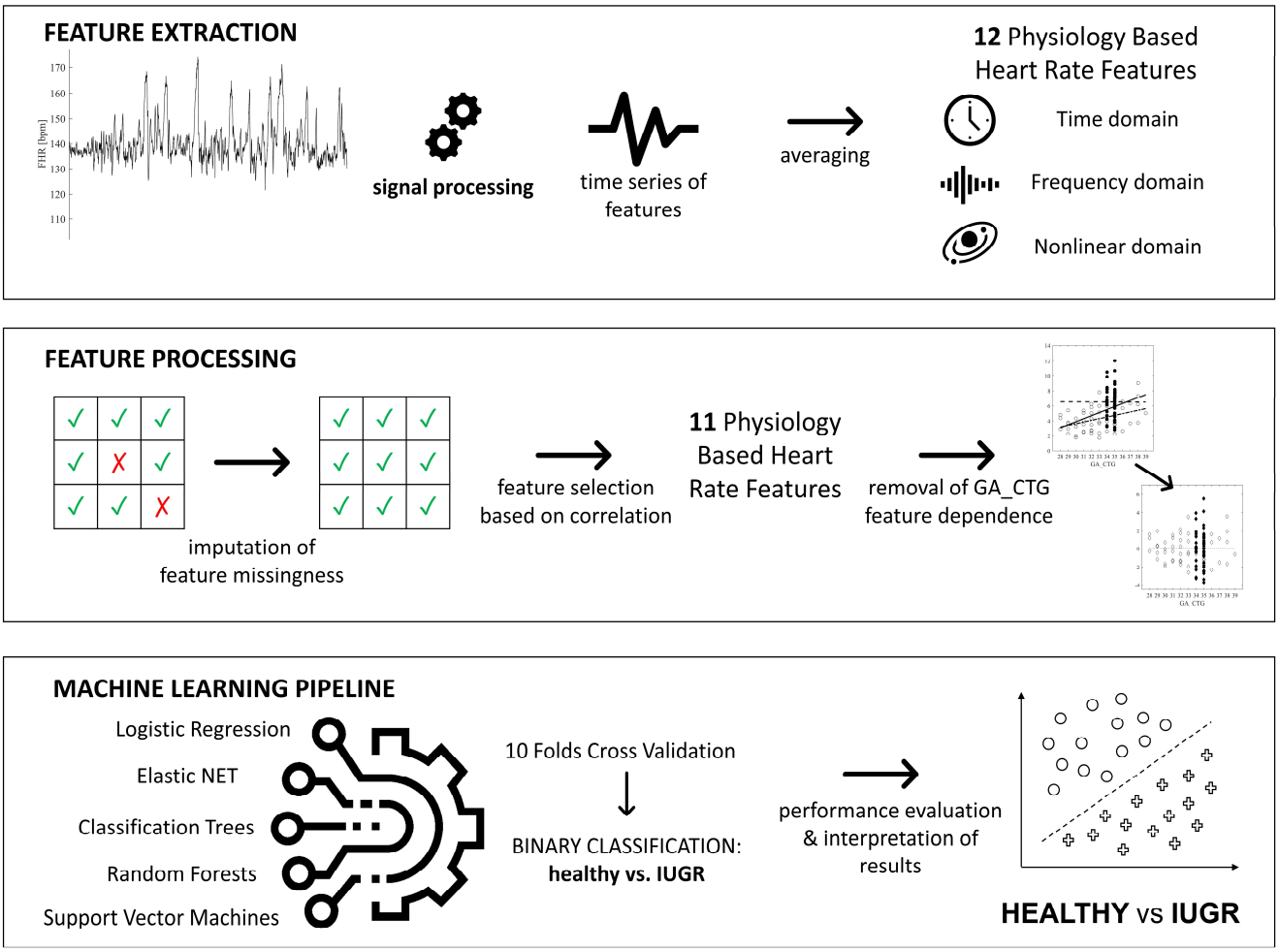

Fig. 3. Machine learning approach to the classification of antepartum fetal heart rate signal. Top panel: from signal to feature extraction; middle panel: feature processing, imputation of missing parameters, feature space reduction based on physiological knowledge; bottom panel: machine learning techniques and validation of performances.

Several multivariate models were employed towards to aim of identifying the most reliable technique for predicting IUGR condition. The employed machine learning techniques and the corresponding employed R packages are reported in Table II.

The following algorithms were applied:

Logistic Regression (LR): is a regression model where the probability of a class of interest is obtained as the results of a logistic function provided with a linear combination of the features. The general formulation for Logistic Regression is expressed in Equation 1:

$$P(y|\boldsymbol{x}) = \frac{e^{\alpha+\sum_{i=1}^{N}\beta_i x_i}}{1+e^{\alpha+\sum_{i=1}^{N}\beta_i x_i}} \quad (1)$$

TABLE II
TESTED MACHINE LEARNING TECHNIQUES, THEIR CORRESPONDING ACRONYMS, AND EMPLOYED R PACKAGES

Machine Learning Technique	Acronym	R package
Logistic Regression including all covariates	LR	stats
Logistic Regression, stepwise feature selection and pairwise interactions between features	LR-SW-INT	stats
Logistic Regression, stepwise feature selection and without pairwise interactions between features	LR-SW	stats
RIDGE regression	RIDGE	glmnet
Elastic NET, alpha = 0.25	ENET 0.25	glmnet
Elastic NET, alpha = 0.50	ENET 0.5	glmnet
Elastic NET, alpha = 0.75	ENET 0.75	glmnet
Least Absolute Selection and Shrinkage Operator	LASSO	glmnet
Naïve Bayes	NB	e1071
Classification Trees	CT	rpart
Random Forests	RF	randomForest
Support Vector Machines, linear kernel	SVM-LIN	e1071
Support Vector Machines, polynomial kernel	SVM-POLY	e1071
Support Vector Machines, radial kernel	SVM-RAD	e1071
Support Vector Machines, sigmoid kernel	SVM-SIGM	e1071

where *y* is a target class (Healthy versus IUGR), x_i are the available features, α and βs are the regression coefficients estimated by the algorithm. The method, as formulated in the previous Equation 1 generates a linear decision boundary, i.e. a hyperplane in the multidimensional space.

LR can be utilized including all covariates (LR), namely the whole set of previously extracted parameters or coupled with a features selection algorithm called stepwise selection of informative features (step function) [30], allowing (LR-SW-INT) or not (LR-SW) for pairwise interactions between features [31];

Within the family of approaches based on regression we also employed *RIDGE regression* for binary outcomes *(RIDGE)* [32]; *elastic net regression* for binary outcomes with different alpha settings (alpha = 0.25 *(ENET 0.25)*, alpha = 0.50 *(ENET 0.5)*, alpha = 0.75 *(ENET 0.75)* [33]; and *Least Absolute Selection and Shrinkage Operator* for binary outcomes *(LASSO)* [34]. For *RIDGE*, *ENET* and *LASSO* regressions, the optimal lambda parameter was computed by considering each training set separately (as described in the following) and the same seed was imposed for each analysis.

Naïve Bayes (NB): is a classification algorithm based on the Bayes theorem. NB assumes that the attributes x_i are conditionally independent given the class *y*, as formulated in Equation 2:

$$P(y|\boldsymbol{x}) \propto P(y)\prod_{i=1}^{N} P(x_i|y) \tag{2}$$

Despite these rather simplistic assumptions, NB often outperforms more sophisticated machine learning

algorithms. This is due to the fact that although the individual class density estimates may be biased, the assumption of feature independence (given the class variable) is not effectively affecting the posterior probabilities [35] of belonging to a specific class.

Classification Trees (CT): are widely used in the machine learning field and they consist of a set of rules that defines a tree-like structure, in which branches represent different decisional paths and terminal nodes (leaves) corresponds to the assignment to a target class. CT generates a set of nonlinear decision boundaries through piecewise constant functions in the multidimensional space. In this work, the information gain was employed as the splitting criterion for CT [36].

Random Forests (RF): are ensemble classifiers that consist of a variable number of CTs grown based on a set of attributes selected randomly from the complete set of parameters; each CT contributes with its own classification of the analyzed examples. As a result, the final classification is provided by a voting approach, which considers the complete set of CTs. Thanks to their scalability and generalization performance, RFs are increasingly exploited in clinical research [37].

Support Vector Machines (SVM): are a family of classifier capable of mapping the training samples into high-dimensional attributes space, to the aim of defining a hyperplane that maximizes the distance between observations belonging to the different classes. If the training set cannot be separated by a linear boundary, the optimal hyperplane that best discriminates between/among examples of different class labels is identified resorting to a suitable space transformation through kernel functions. In this work, we employed SVMs with linear kernel *(SVM-LIN)*, polynomial kernel *(SVM-POLY)*, radial kernel (SVM-RAD), and sigmoid kernel *(SVM-SIGM)* [38].

From an implementation point of view, models were learned on the training sets using the default settings. The available data were split into training and testing sets according to a 10 Folds Cross Validation (CV). The training sets were employed to the aim of evaluating the performances of classification algorithms and different feature selection while the corresponding test sets were used to test the relative discriminative performances. The above-described machine learning methods were tested on either the U and A set of features towards to aim of comparing the two approaches and identifying if GA_CTG had a significant effect on IUGR classification.

2.4 Multivariate model evaluation

In this work, an IUGR subject correctly classified as such is counted as a true positive (TP), and a healthy subject correctly classified is counted as a true negative (TN). On the contrary, an IUGR subject erroneously classified as healthy is counted as a false negative (FN), and a healthy subject erroneously classified as IUGR is counted as a false positive (FP).

The performances of each model are reported in terms of four different figures of merit, namely Classification Accuracy: $CA = (TP + TN)/(TP + TN + FP + FN)$, sensitivity: $sensitivity = TP/(TP + FN)$, specificity: $specificity = TN/(TN + FP)$, positive predictive value: $PPV = TP/(TP + FP)$, and negative predictive value: $NPV = TN/(TN + FN)$. The Area Under the Receiver Operating Characteristic (AUROC) was estimated by averaging the results obtained by providing the model with different test sets, namely the ones obtained using 10 Folds CV procedure. Since the healthy/IUGR ratio was 1, model ranking was performed based on CA.

3. RESULTS

Multivariate analysis has been performed considering alternatively the U and A set of features. Prior to multivariate testing, a preliminary analysis of the correlation between covariates has been performed. As a general consideration, features to be provided to any machine learning algorithm should be highly correlated with the classes to be distinguished but not be highly correlated with one another [39]. By way of example, values of correlation for the A set of covariates are reported in Table III.

TABLE III
SPEARMAN'S RANK CORRELATION COEFFICIENT COMPUTED ON THE ADJUSTED SET OF COVARIATES

	Delta	II	STV	LTI	LF_pow	MF_pow	HF_pow	LF/(MF+HF)	ApEn(1, 0.1)	LZC(2, 0)	APRS
DPRS	-0.60	-0.03	-0.62	-0.42	-0.37	0.22	0.29	-0.37	-0.03	-0.26	-0.84
Delta		0.03	0.93	0.43	0.4	-0.17	-0.45	0.40	0.06	0.33	0.57
II			-0.10	0.03	-0.11	0.15	0.01	-0.11	0.01	0.13	-0.07
STV				0.38	0.36	-0.13	-0.41	0.36	0.02	0.31	0.57
LTI					0.29	-0.18	-0.21	0.28	0.08	0.17	0.37
LF_pow						-0.72	-0.75	0.99	-0.27	0.24	0.33
MF_pow							0.15	-0.72	0.20	0.07	-0.27
HF_pow								-0.75	0.32	-0.39	-0.18
LF/(MF+HF)									-0.27	0.26	0.33
ApEn(1, 0.1)										0.14	0.02
LZC(2, 0)											0.19

The correlation coefficient values in each domain are on average higher than comparing feature correlation in the same area. The former result is related to the fact that the proposed features have the ability to grasp different characteristics of FHR, thus their information content is different, resulting in a low value

of correlation. Regarding high values of correlation among indexes of the same domain, a clear example is the parameter LF/(MF+HF) which is highly correlated to the other frequency extracted indexes (LF_pow, MF_pow, and HF_pow). Based on this criterion, the ratio LF/(MF+HF) was excluded from the set of employed parameters, resulting in a reduced parameter space of 11 features: Delta, II, STV, LTI (time domain); LF_pow, MF_pow, HF_pow (frequency domain); ApEn(1, 0.1), LZC(2, 0), APRS, DPRS (nonlinear domain).

Table IV and Table V report the mean discriminative performances of the top five machine learning techniques in classifying the test sets (10 Folds CV) for U and A set of covariates respectively. The average performances in the first case are: CA=0.8812, Sensitivity=0.8912, Specificity=0.8704, PPV=0.8908, NPV=0.8988, whereas on the second one: CA=0.8296, Sensitivity=0.8544, Specificity=0.8048, PPV=0.8320, NPV=0.8666.

The two machine learning techniques which outperformed, showing the best discriminative performances were: RF_U: mean CA=0.911 and CT_U: CA = 0.911 when considering the model learned on U covariates. In the case of adjusted covariates by GA_CTG, RF_A: mean CA=0.855, and LR-SW_A: mean CA=0.833, showed the best CA among the proposed machine learning models.

Focusing the attention on the comparison of AUROC for RF_U, CT_U, RF_A, and LR-SW_A, no statistically significant difference was observed after performing post-hoc tests between models' AUROC. The values of AUROC (averaged over the 10 test sets) for RF_U: AUROC=0.974, CT_U: AUROC=0.892, RF_A: AUROC=0.935, and LR-SW_A: AUROC=0.933 are reported in Fig. 4.

TABLE IV
MEDIAN (25th, 75th PERCENTILES) OF CLASSIFICATION ACCURACY (CA), SENSITIVITY, SPECIFICITY, POSITIVE AND NEGATIVE PREDICTIVE VALUES (PPV AND NPV) FOR ADOPTED MACHINE LEARNING TECHNIQUES LEARNED ON THE UNADJUSTED SET OF COVARIATES. MACHINE LEARNING TECHINIQUES ARE SORTED IN DESCENDING ORDER OF CA

Model	CA	Sensitivity	Specificity	PPV	NPV
RF	0.911 (0.860, 0.961)	0.902 (0.820, 0.985)	0.919 (0.819, 1.019)	0.936 (0.859, 1.013)	0.918 (0.852, 0.984)
CT	0.911 (0.846, 0.975)	0.871 (0.766, 0.976)	0.950 (0.892, 1.008)	0.949 (0.890, 1.009)	0.893 (0.808, 0.978)
LR-SW	0.867 (0.797, 0.937)	0.900 (0.817, 0.983)	0.833 (0.721, 0.946)	0.859 (0.774, 0.944)	0.900 (0.822, 0.978)
SVM-RAD	0.867 (0.781, 0.952)	0.850 (0.762, 0.938)	0.883 (0.770, 0.996)	0.893 (0.790, 0.996)	0.856 (0.775, 0.938)
SVM-POLY	0.850 (0.762, 0.938)	0.933 (0.850, 1.017)	0.767 (0.627, 0.907)	0.817 (0.712, 0.922)	0.927 (0.838, 1.017)

TABLE V
MEDIAN (25th, 75th PERCENTILES) OF CLASSIFICATION ACCURACY (CA), SENSITIVITY, SPECIFICITY, POSITIVE AND NEGATIVE PREDICTIVE VALUES (PPV AND NPV) FOR ADOPTED MACHINE LEARNING TECHNIQUES LEARNED ON THE ADJUSTED SET OF COVARIATES. MACHINE LEARNING TECHINIQUES ARE SORTED IN DESCENDING ORDER OF CA

Model	CA	Sensitivity	Specificity	PPV	NPV
RF	0.855 (0.794, 0.916)	0.838 (0.729, 0.947)	0.871 (0.766, 0.976)	0.889 (0.799, 0.980)	0.862 (0.773, 0.951)
LR-SW	0.833 (0.759, 0.908)	0.867 (0.773, 0.961)	0.800 (0.665, 0.935)	0.835 (0.737, 0.934)	0.870 (0.785, 0.955)
LR	0.825 (0.743, 0.907)	0.850 (0.731, 0.969)	0.800 (0.665, 0.935)	0.830 (0.730, 0.931)	0.862 (0.766, 0.958)
SVM-RAD	0.818 (0.738, 0.897)	0.850 (0.719, 0.981)	0.786 (0.687, 0.885)	0.806 (0.723, 0.888)	0.866 (0.756, 0.977)
LASSO	0.817 (0.716, 0.917)	0.867 (0.744, 0.990)	0.767 (0.627, 0.907)	0.800 (0.687, 0.914)	0.873 (0.761, 0.985)

Figure 4

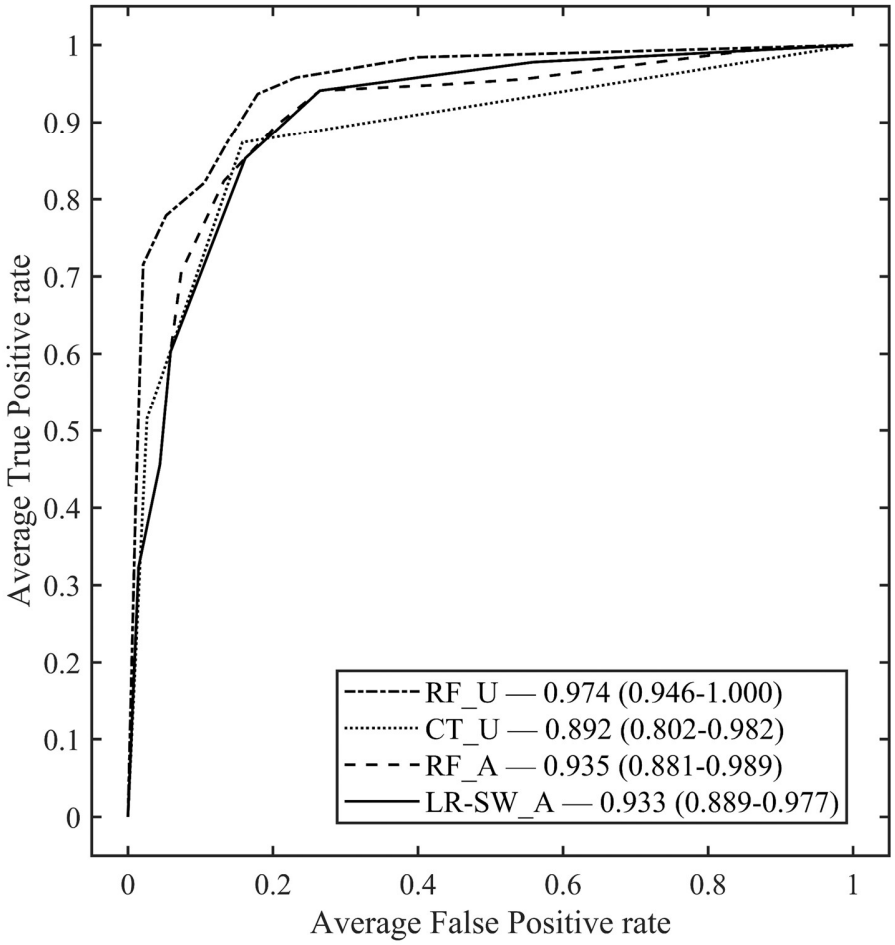

Fig. 4. Mean AUROC (95% CI) for the multivariate analysis: RF_U, CT_U, RF_A, and LR-SW_A. False positive rate is defined as FP/(FP+TN) and true positive rate corresponds to sensitivity.

In order to define the final models, RF and LR-SW were learned on the whole set of computed features. Nevertheless, features selection procedures for both RF and LR-SW were performed. It is crucial to pinpoint

that investigating the performances of the former machine learning techniques on a reduced feature space may be helpful in reducing the amount of FHR extracted parameters within achieving the same level of prediction accuracy.

Regarding RF_U and RF_A, the relative importance of each feature is shown in Fig. 5. Results indicate that for the U set of covariates LZC(2, 0), ApEn(1, 0.1), HF_pow, LTI and DPRS caused the greatest decrease in terms of CA if removed from the model. Similarly, when considering RF_A, LZC(2, 0), HF_pow, ApEn(1, 0.1), LTI and LF_pow were identified as most explanatory variables for the model. It is crucial to highlight

Figure 5

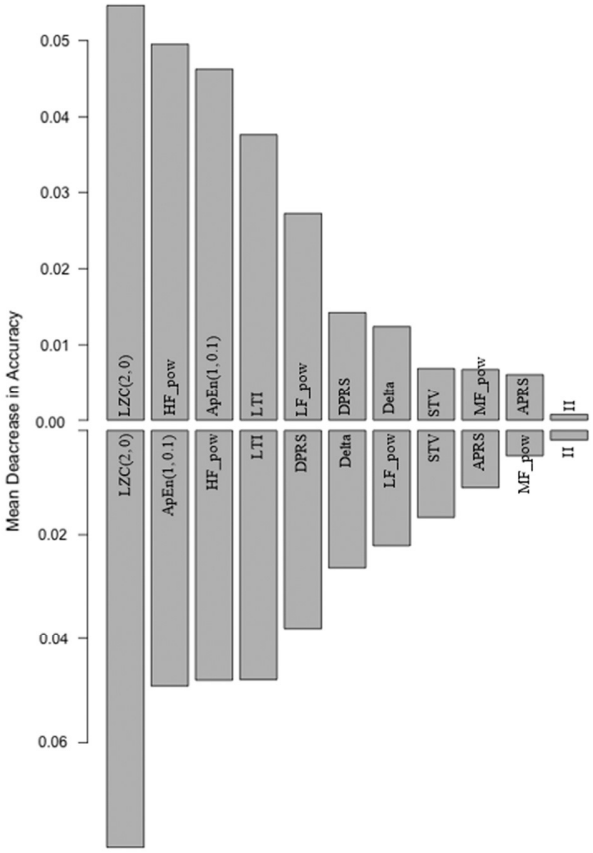

Fig. 5. Variables importance according to the RF classifier on the whole cohort and associated decrease in model CA when progressively excluding features. Top bar graph is relative to RF_A and bottom one to RF_U.

that the most explanatory parameters are encompassing all the investigated domain: time, frequency and nonlinear. The latter assumption is valid for both RF_U and RF_A enforcing the idea that combining FHR features belonging to different domains provides a more comprehensive and extensive snapshot of the interacting mechanism leading to the IUGR condition.

Coming to the second-best performing machine learning techniques, CT_A identifies the same covariates previously found for RF_U as most explanatory. In the case of LR-SW_U, the covariates producing the highest decrease in accuracy if excluded are: LTI, LZC(2, 0), STV, LF_pow, HF_pow, ApEn(1, 0.1), MF_pow. Consistently with RF, despite the different ranking of importance, CT_A and LR-SW_U select a reduced set of variables encompassing the three different domains. Fig. 5 displays variables ranked in descending order of mean decrease in accuracy. It is crucial to pinpoint that RF methodology appears stable and quite insensitive to covariance dependence upon GA_CTG. LZC(2, 0) is identified as producing the most impactful decrease in accuracy by both RF_U and RF_A. The remaining covariates are on average ranked in a similar fashion by the two models, strengthening the limited impact of GA_CTG on classification accuracy.

On the opposite, it is peculiar to observe the dramatic drop of performances when comparing CT_U and CT_A. In particular, CT_U is ranked as the second-best performing machine learning technique while CT_A is the least performing one (CA=0.771, Sensitivity=0.757, Specificity=0.786, PPV=0.805, NPV=0.802) followed by SVM-SIGM only (CA=0.702, Sensitivity=0.705, Specificity=0.700, PPV=0.709, NPV=0.710). In the latter case, GA_CTG plays a fundamental role and it becomes evident that the correction for such dependence is mandatory to provide accurate discrimination between healthy and IUGR fetuses. Regarding the remaining machine learning techniques, namely SVM and ENET, yet not giving the best performances, they appear less dependent upon GA_CTG providing comparable results in terms of CA when employing either U or A set of covariates.

4. DISCUSSION

The presented investigation provides evidence of a feasible application of machine learning techniques for the early identification of IUGR condition in the antepartum period. Such design appears as radically different with respect to the up-to-date clinical practice where IUGR condition is assessed at birth and only suspected in the antepartum period. The rationale for the utilization of the presented physiology based heart rate features relies on the fact that these features as standalone parameters had shown enhanced discrimination power in classifying healthy versus IUGR fetuses [21,25,40].

However, throughout the years it has become clear that a single index cannot be descriptive of all pathophysiological processes taking place in the pregnancy period thus the need for multivariate analysis of FHR emerged as evident. These are the main reasons contributing to the choice of the use of physiology based heart rate features in this investigation. Moreover, our approach demonstrated the independence of different machine learning methodologies to the time at which CTG recordings were acquired.

Coming to the discussion of the results section, it appears evident that both RF_U and RF_A achieved adequate performances, thus proposing as a possible candidate as a tool for early discrimination in the context of the presented investigation. Random Forest is becoming a popular machine learning technique and it has been claimed as particularly accurate and interpretable by several authors [41], [42]. A clear example of interpretability of the results is the feature ranking results reported in Fig. 5. Consistently with previous findings [40], LZC(2, 0) is associated with the most considerable mean decrease in accuracy. On average, IUGR fetuses have been reported as characterized by lower values of LZC with respect to healthy ones [25], as this is also verified in this analysis. The reported difference is to be attributed to lower complexity of FHR for pathological subjects, thus supporting the hypothesis of an unbalance in the autonomic nervous system mechanisms in IUGR condition. Similarly, values of ApEn(1, 0.1) are greater in healthy versus IUGR. Nevertheless, this entropy index resulted in a lower mean decrease in accuracy accordingly to the reported lower discriminative power with respect to LZC measures [40]. Moreover, the corresponding time domain

index (LTI) which quantifies FHR variability considering windows of analogous time duration was found among the top informative features. As for both LZC(2, 0) and ApEn(1, 0.1), LTI values in healthy are greater to the ones for IUGR subjects as previously found in [25]. This latter finding contributes to the hypothesis of an impaired ANS regulation in the pathological conditions. Lastly, PRSA-derived index DPRS was found significantly greater in IUGR versus healthy. Despite not providing an analogous definition of acceleration and deceleration as ones found in the clinical context, the PRSA slope is dependent upon both the amplitude and duration of the ANS-related events modulating the FHR [9].

To summarize, the reported results reinforce the idea that several controlling mechanisms affect HRV, acting linearly and nonlinearly. This specifically happens when a pathological condition arises, and the analytic frameworks need to merge and combine information coming from different domains to obtain an exhaustive and comprehensive description of FHR dynamics. The latter consideration is reflected in the obtained findings considering feature ranking in RF, reporting the first five features encompassing the three domains of investigation, namely time, frequency, and nonlinear.

5. CONCLUSIONS

Findings reported in this investigation confirm the importance of a multivariate approach to investigate the variety of implications resulting from a pathological condition such as IUGR. The advantages resulted by the application of several machine learning techniques rely on: i) easy-to-use model capable of providing an early and interpretable antenatal diagnosis of IUGR condition; ii) parameters extracted from routinely CTG examination can be fed into the model regardless the considered GA_CTG. The latter novelty is of primary importance given that, in nowadays clinical practice, IUGR fetuses are usually monitored far in advance with respect to healthy ones so that the proposed model may see its direct translation in the clinical field. This opens to further and extensive validation of multi-feature model presented in this work on a large already recorded and available dataset. The cutting-edge frontier for the methods described in this work would be focusing on tracking the evolution from health condition to pathological state in a patient-specific way by integrating heterogeneous data which are dynamically evolving in time.

ACKNOWLEDGMENT

The authors are grateful to Dr. M. Campanile and her team, Dept. of Obstetrical–Gynaecological and Urological Science and Reproductive Medicine, Federico II University, Naples, Italy, for the collection of CTG tracings of healthy and IUGR fetuses.

REFERENCES

[1] M.B. Landon, S.G. Gabbe, Antepartum fetal surveillance in gestational diabetes mellitus, Diabetes. 34 (1985) 50–54.

[2] A.A. Baschat, R.M. Viscardi, B. Hussey-Gardner, N. Hashmi, C. Harman, Infant neurodevelopment following fetal growth restriction: Relationship with antepartum surveillance parameters, Ultrasound Obstet. Gynecol. 33 (2009) 44–50. doi:10.1002/uog.6286.

[3] J. de Haan, J.H. van Bemmel, B. Versteeg, A.F.L. Veth, L.A.M. Stolte, J. Janssens, T.K.A.B. Eskes, Quantitative evaluation of fetal heart rate patterns: I. Processing methods, Eur. J. Obstet. Gynecol. 1 (1971) 95–102. doi:10.1016/0028-2243(71)90056-6.

[4] D. Hoyer, J. Zebrowski, D. Cysarz, H. Goncalves, A. Pytlik, C. Amorim-Costa, J. Bernardes, D. Ayres-De-Campos, O.W. Witte, E. Schleußner, L. Stroux, C. Redman, A. Georgieva, S. Payne, G. Clifford, M.G. Signorini, G. Magenes, F. Andreotti, H. Malberg, S. Zaunseder, I. Lakhno, U. Schneider, Monitoring fetal maturation - Objectives, techniques and indices of autonomic function, Physiol. Meas. 38 (2017) R61–R88. doi:10.1088/1361-6579/aa5fca.

[5] T. Todros, C.U. Preve, C. Plazzotta, M. Biolcati, P. Lombardo, Fetal heart rate tracings: Observers versus computer assessment, 1996. doi:10.1016/0301-2115(96)02487-6.

[6] G. Magenes, M.G. Signorini, D. Arduini, S. Cerutti, Fetal Heart Rate Variability due to Vibroacoustic Stimulation: Linear and Nonlinear Contribution, Methods Inf. Med. 43 (2004) 47–51. doi:10.1267/METH04010047.

[7] M.G. Signorini, G. Magenes, S. Cerutti, D. Arduini, Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings, IEEE Trans. Biomed. Eng. 50 (2003) 365–374. doi:10.1109/TBME.2003.808824.

[8] D. Arduini, G. Rizzo, A. Piana, B.P. Brambilla, C. Romanini, Computerized analysis of fetal heart rate: I. description of the system (2ctg), J Matern Fetal Invest. 3 (1993) 159–164.

[9] M.G. Signorini, A. Fanelli, G. Magenes, Monitoring fetal heart rate during pregnancy: Contributions from advanced signal processing and wearable technology, Comput. Math. Methods Med. 2014 (2014). doi:10.1155/2014/707581.

[10] S. Pincus, Approximate entropy (ApEn) as a complexity measure, Chaos. 5 (1995) 110–117. doi:10.1063/1.166092.

[11] S.M. Pincus, R.R. Viscarello, Approximate entropy: a regularity measure for fetal heart rate analysis., Obstet. Gynecol. 79 (1992) 249–55.

[12] J.S. Richman, J.R. Moorman, Physiological time-series analysis using approximate entropy and sample entropy, Am. J. Physiol. Circ. Physiol. 278 (2000) H2039–H2049. doi:10.1152/ajpheart.2000.278.6.H2039.

[13] A. Lempel, J. Ziv, On the complexity of finite sequences, IEEE Trans. Inf. Theory. 22 (1976) 75–81. doi:10.1109/TIT.1976.1055501.

[14] W.W. Stead, Clinical implications and challenges of artificial intelligence and deep learning, JAMA - J. Am. Med. Assoc. 320 (2018) 1107–1108. doi:10.1001/jama.2018.11029.

[15] G. Hinton, Deep Learning—A Technology With the Potential to Transform Health Care, JAMA. 320 (2018) 1101. doi:10.1001/jama.2018.11100.

[16] C.D. Naylor, On the prospects for a (Deep) learning health care system, JAMA - J. Am. Med. Assoc. 320 (2018) 1099–1100. doi:10.1001/jama.2018.11103.

[17] J. Smith, M. Murphy, Y. Kandasamy, The IUGR infant: A case study and associated problems with IUGR infants, J. Neonatal Nurs. 19 (2013) 46–53. doi:10.1016/j.jnn.2012.12.005.

[18] A. Rosenberg, The IUGR Newborn, Semin. Perinatol. 32 (2008) 219–224. doi:10.1053/j.semperi.2007.11.003.

[19] D. Sharma, S. Shastri, P. Sharma, Intrauterine Growth Restriction: Antenatal and Postnatal Aspects, Clin. Med. Insights Pediatr. 10 (2016) CMPed.S40070. doi:10.4137/cmped.s40070.

[20] Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology, Heart rate variability. Standard of measurement, physiological interpretation and clinical use, Circulation. 93 (1996) 1043–1065.

[21] M. Ferrario, M.G. Signorini, G. Magenes, Comparison between fetal heart rate standard parameters and complexity indexes for the identification of severe intrauterine growth restriction, Methods Inf. Med. 46 (2007) 186–190. doi:07020186.

[22] A. Bauer, J.W. Kantelhardt, A. Bunde, P. Barthel, R. Schneider, M. Malik, G. Schmidt, Phase-rectified signal averaging detects quasi-periodicities in non-stationary data, Phys. A Stat. Mech. Its Appl. 364 (2006) 423–434. doi:10.1016/j.physa.2005.08.080.

[23] S.M. Lobmaier, E.A. Huhn, S. Pildner Von Steinburg, A. Müller, T. Schuster, J.U. Ortiz, G. Schmidt, K.T. Schneider, Phase-rectified signal averaging as a new method for surveillance of growth restricted fetuses, J. Matern. Neonatal Med. 25 (2012) 2523–2528. doi:10.3109/14767058.2012.696163.

[24] M. Ferrario, M.G. Signorini, G. Magenes, Comparison between fetal heart rate standard parameters and complexity indexes for the identification of severe intrauterine growth restriction, Methods Inf. Med. 46 (2007) 186–190. doi:07020186 [pii].

[25] A. Fanelli, G. Magenes, M. Campanile, M.G. Signorini, Quantitative assessment of fetal well-being through ctg recordings: A new parameter based on phase-rectified signal average, IEEE J. Biomed. Heal. Informatics. 17 (2013) 959–966. doi:10.1109/JBHI.2013.2268423.

[26] R Development Core Team, R: A language and environment for statistical computing, 2011.

[27] D.J. Stekhoven, P. Bühlmann, Missforest—non-parametric missing value imputation for mixed-type data, Bioinformatics. 28 (2012) 112–118. doi:10.1093/bioinformatics/btr597.

[28] J. Cohen, Statistical power analysis for the behavioral sciences, L. Erlbaum Associates, 1988.

[29] G. Magenes, R. Bellazzi, A. Malovini, M.G. Signorini, Comparison of data mining techniques applied to fetal heart rate parameters for the early identification of IUGR fetuses, in: 2016 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., IEEE, 2016: pp. 916–919. doi:10.1109/EMBC.2016.7590850.

[30] P. McCullagh, J.A. Nelder, Generalized linear models, Chapman and Hall, 1989.

[31] Z. Bursac, C.H. Gauss, D.K. Williams, D.W. Hosmer, Source Code for Biology and Medicine Purposeful selection of variables in logistic regression, (n.d.). doi:10.1186/1751-0473-3-17.

[32] A.E. Hoerl, R.W. Kennard, Ridge Regression: Biased Problems Nonorthogonal Estimation for, Technometrics. 42 (2000) 80–86. doi:10.1080/00401706.1970.10488634.

[33] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2005) 301–320. doi:10.1111/j.1467-9868.2005.00503.x.

[34] R. Tibshirani, Regression shrinkage and selection via the lasso: a retrospective, J. R. Stat. Soc. B. 73 (2011) 273–282.

[35] D.J. Hand, K. Yu, Idiot's Bayes - Not so stupid after all?, Int. Stat. Rev. 69 (2001) 385–398. doi:10.1111/j.1751-5823.2001.tb00465.x.

[36] S.B. Kotsiantis, Decision trees: a recent overview, Artif Intell Rev. 39 (2013) 261–283. doi:10.1007/s10462-011-9272-4.

[37] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and regression trees, Chapman & Hall/CRC, 1984.

[38] Y. Zhang, Support Vector Machine Classification Algorithm and Its Application, in: Springer, Berlin, Heidelberg, 2012: pp. 179–186. doi:10.1007/978-3-642-34041-3_27.

[39] M. Hall, Correlation-based Feature Selection for Machine Learning, Methodology. 21i195-i20 (1999) 1–5. doi:10.1.1.149.3848.

[40] M. Ferrario, M.G. Signorini, G. Magenes, Complexity analysis of the fetal heart rate for the identification of pathology in fetuses, Comput. Cardiol. 32 (2005) 989–992. doi:10.1109/CIC.2005.1588275.

[41] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32. doi:10.1023/A:1010933404324.

[42] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, 2nd ed., Springer, 2009. doi:10.1007/978-0-387-84858-7.