POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Predicting Fetal Weight Disorders in Diabetic Pregnancies: an explainable Machine Learning approach

Laurea Magistrale in Biomedical Engineering - Ingegneria biomedica

Author: Elena Novelli

Advisor: Prof. Maria Gabriella Signorini

Co-advisor: Giulio Steyde

Academic year: 2023-2024

1. Introduction

Pregnancy is a complex period that requires careful monitoring, expecially ante-partum, to prevent potential complications. Among these, diabetes represents one of the most significant risks, with long-term detrimental effects on various organs functionalities. Diabetes is defined as a "metabolic disorder characterized by chronic hyperglycemia resulting from defects in insulin secretion, insulin action, or both". With diabetes the risk of complications increases for both mother and fetus, underscoring the importance of a careful monitoring during pregnancy [1]. Cardiotocography (CTG), as a non-invasive examination, is a primary tool for such monitoring. By measuring fetal heart activity and uterine contractions, it allows for the prompt identification of signs of fetal distress and enables the implementation of necessary measures to ensure the well-being of the fetus and the mother. Its computerized version (cCTG) provides numerical parameters related to fetal conditions, playing a critical role in evaluating fetal heart rate (FHR) [2]. Among the risks associated with diabetes in pregnancy, significant are those related to fetal weight. Specifically, the condition of "small for gestational age" (SGA) indicates a fetal weight below the 10th percentile of the general population, considering gestational age, often associated with maternal metabolic alterations that limit nutrient exchange with the fetus. Conversely, "large for gestational age" (LGA) refers to a population with fetal weight above the 90th percentile of the general one, often correlated with high maternal glucose levels that promote excessive fetal growth.

1.1. Aim of the work

This study aims to explore the relationship between cCTG parameters and fetal weight at birth, specifically focusing on pregnant women affected by diabetes. The main objective is to train a machine learning classification algorithm to discriminate fetal weight, divided into three categories that are SGA, "normal for gestational age" (NGA), and LGA, using parameters calculated from the FHR signal, obtained through cCTG, and maternal information derived from her clinical history. This could serve as a primary tool for clinicians in identifying high-risk pregnancies in a delicate situation such as pregnancies with diabetes. Additionally, the study aims to evaluate, through an explainability method, which variables are most relevant for this prediction and especially the interaction of cCTG parameters with general physiological indices already accessible to healthcare professionals, such as anamnestic data of the pregnant woman and particularly the presence and type of diabetes. The importance of variables derived from maternal medical history is acknowledged; nonetheless, this study aims to determine if integrating these factors with a set of quantitative parameters from CTG analysis might represent a step towards improving indicative assessment of fetal health state. There is indeed a lack of literature on studies concerning variables capable of distinguishing problematic classes within an already pathological category such as diabetes.

2. Dataset and features selection

The data employed in this thesis were gathered at the ObGyn Department of the University Hospital Federico II in Naples, Italy. The derived dataset was obtained as a result of routine antepartum fetal monitoring examinations [3]. cCTG records, lasting between 20 to 60 minutes, were obtained utilizing Philips Avalon family monitors that generate FHR signal sampled at 2Hz. Each tracing was categorized by medical professionals. From this database, only the records belonging to diabetic pregnant women for whom information on fetal weight at birth was available were selected. The fetal weight categories SGA, NGA, and LGA have been defined using fetal growth charts based on estimated fetal weight developed by World Health Organization [4]. In particular, the week of delivery, which was provided in the notes made by clinicians, was utilized to establish a posterior threshold for comparing birth weights. This led to the division of records into the three weight categories, which also constitute the target of the multiclass classification, according to the Table 1.

Fetal weight category	**Threshold**	**N. of recordings**
SGA	weight < Th. SGA	237
NGA	Th. SGA < weight < Th. LGA	407
LGA	weight > Th. LGA	165

Table 1: Summary table of the target. SGA= small, NGA=normal, LGA= large for gestational age. Th=threshold

Subsequently, the following information concerning the maternal medical history was extracted from the database and used as part of the features for the machine learning algorithms:

- Pregnancy_ID: a unique numerical code assigned to each pregnancy;
- Visit_ID: a unique sequential numerical code associated to each recording;
- Gest_Week: an integer number representing the gestational week of the pregnancy;
- Num_Pregnancy: an integer representing how many pregnancies the woman has had previously;
- Age: age of the mother;
- Diabetes' type

A preprocessing phase of the FHR signal was then carried out. In this phase, an outcome of the cCTG consisting of a set of integer values representing the quality of each sample of the signal was used to determine good quality segments of the FHR signal and samples that instead needed to be linearly interpolated. This is due to missing values or physiologically implausible ones. Baseline, accelerations, and decelerations of the FHR signal were also calculated.

Following this, the actual processing phase of the FHR signal was conducted, during which the parameters listed in Table 2 were calculated, representing the additional features employed in this thesis (in addition to those related to maternal characteristics already mentioned). Details about their definitions and implementation can be found in [5]. Regarding the parameters calculated in 1-minute and 3-minute windows, these were considered only in segments of active sleep (a phase of fetal behavioural state marked by movement, accelerations of fetal beats, and high heart rate variability) of the FHR, ignoring those of quiet sleep (considered as associated to fetal sleep). This is due to the fact that from literature is known that an instance of active sleep serves as a sign of fetal health and stands as one of the primary criteria utilized in the system proposed to assess fetal normality by [6]. Moreover studies have already indicate fewer active sleep periods in problematic fetuses with markers showing higher discrimination in active compared to quiet sleep episodes.

For the identification of the active sleep segments, a deep neural network with a 1D encoder-decoder architecture developed in [7] was employed. Its usage also led to the addition of another variable to the set representing the percentage of activity segments in the entire signal. In absence of activity segments, an arbitrary value of -1 was assigned to the parameters that were supposed to be calculated in windows, and an additional binary variable, d_par, was then added to identify such samples.

Parameter's name	Domain	Windows' length
Short Term Variability (STV)	T	1 min
Interval Index (II)	T	1 min
Delta (Δ)	T	1 min
n_acc	T	Global
n_dec	T	Global
Very Low Frequency (VLF)	F	3 min
Low Frequency (LF)	F	3 min
Movement Frequency (MF)	F	3 min
High Frequency (HF)	F	3 min
Approximate Entropy (ApEn)	N-L	Global
Sample Entropy (SampEn)	N-L	Global
Multiscale entropy (MSE)	N-L	Global
Sample Asymmetry (SampAsi)	N-L	Global
Binary Lempel-Ziv (LZC2)	N-L	Global/3 min
Ternary Lempel-Ziv (LZC3)	N-L	Global/3 min
Acceleration capacity (AC)	N-L	Global
Deceleration capacity (DC)	N-L	Global
Deceleration reserve (DR)	N-L	Global
Acceleration Phase Rectified Slope (APRS)	N-L	Global
Deceleration Phase Rectified Slope (DPRS)	N-L	Global
LFprsa	N-L/F	Global
MFprsa	N-L/F	Global
HFprsa	N-L/F	Global

Table 2: Parameters. T= time; F= Frequency; N-L= Non Linear

3. Machine learning

The listed variables were used to train two multiclass classification algorithms for predicting the SGA, NGA, and LGA weight categories of diabetic pregnancies. These models are a multiclass Logistic Regression and a Multilayer Perceptron (MLP). The choice of these models stems from the dual purpose of the work: prediction of the weight classes and interpretability. To ensure the second one, Logistic Regression was selected, which provides insights through the coefficients of the created function. As for predictive capability, reliance was placed on the known ability of neural networks to learn highly challenging tasks thanks to their great versatility.
Before training, the following data preparation steps were performed:

1. Signals that were excessively corrupted (more than 90% of invalid windows i.e., with more than 5% of interpolated samples) were removed, using parameters calculated within 1 and 3-minute windows as a reference. Subsequently, for all parameters, an analysis was conducted to eliminate any outliers, removing values outside the range defined by the 25th and 75th percentiles. This was done only for parameters with at least 10 valid windows;
2. To prevent bias in the classifier and avoid it overfit on a limited number of patients, the number of recordings has been limited to a maximum of 10 per patient, selecting the most recent ones;
3. The categorical variable expressing the type of diabetes with a number is converted to "dummy" variables that expresses the type of diabetes (type 1, 2 and gestational)
4. For model training, the database was divided into a training set and a test set with a ratio of 80% and 20%, respectively, ensuring that recordings belonging to the same patient were placed in the same set. This decision was made to prevent introducing data leakage from training to test set in the model testing phase;
5. Due to missing annotations during recording, there are missing values in the selected variables. To avoid losing information, it was chosen to impute such information based on a method that estimates the missing values in each sample of the

dataset by finding a defined number of nearest neighbors and averaging them. In the training set, this process identifies the nearest neighbors, calculates the estimates, and directly replaces the missing values. In the test set, the information learned during training is used to directly estimate the missing values;

6. All numerical variables, except for those cases where parameter values were set to -1, were then standardized to ensure a common scale of representation. Similar to the handling of missing values, the operation differs for the train set and test set in this case as well;
7. The target variable is not evenly represented in the samples comprising the database. So, it was used a combination of an undersampling technique on the majority class and the oversampling technique SMOTE on the others to reach the equality. This last algorithm increases the number of samples of a minority class by generating synthetic data [8]. The combination of the two methods depends on the characteristics of the two models trained.

The training and testing process for both models follows a standardized procedure. A stratified k-fold cross-validation is used to divide the training set into a defined number of folds, while preserving the class distribution in each fold. This method helps to mitigate the risks of overfitting and provides a more reliable performance estimate. Notably, SMOTE is applied exclusively to the training data to generate synthetic samples. Afterward, the grid search technique is utilized to explore the optimal hyperparameters for both logistic regression and MLP models. Performance evaluation involves calculating the cross-validation score, which represents the best average performance achieved during cross-validation, with balanced accuracy as the metric of choice. Finally, the model with the most effective hyperparameter combination is selected, and its performance on the test set is determined accordingly.

4. Model performances

One primary objective of the work was to train multiclass machine learning classification models capable of predicting the fetal weight class. The Table 3 shows the results on the test set, in terms of various metrics, of the Logistic Regression and MLP.

Metrics	Logistic Regression	MLP
Balanced accuracy	54.7%	52.6%
F1 score	51.5%	50%
Accuracy	50.2%	49.7%
Balanced accuracy majority voting	55.1%	59.6%

Table 3: Results on test sets for the Logistic Regression and MLP for the classification of the 3 classes

The results indicate substantial progress towards the primary objective. Utilizing pre-childbirth clinical data has proven effective in predicting newborn weight categories, with developed models surpassing baseline thresholds in three classes classification accuracy. Since multiple records were associated with the same patient in the database, majority voting was employed to ensure consistency in outcomes and assess the impact of increased recordings on class assignment accuracy. By grouping records for each pregnancy and assigning the most frequent outcome, both Logistic and MLP models show improved balanced accuracy as demonstrated in table 3. This indicates that an increased number of records enhances the reliability of predictions.

5. Understanding the model: XAI application

Another focus of the work is on result interpretation through Explainable Artificial Intelligence (XAI) tools, emphasizing the importance of providing comprehensible tools and additional knowledge to healthcare professionals. The aim was to improve the accuracy of predictions and interventions in fetal weight anomalies to promote positive outcomes for maternal and fetal health. XAI enables systems to explain decisions understandably fostering model comprehension and transparency in AI-based solutions. In this work, **SHapley Additive exPlanations (SHAP)** has been employed [9]. This is a method designed to add explainability ca-

pacities to machine learning models that employs coalition game theory and Shapley values to interpret the predictions, offering insights into both feature importance and interaction effects for enhanced interpretability [9]. For the overall interpretation of parameter contributions to predictions, the SHAP summary plot was used.
In the case of the Logistic Regression model, this was used in addition to (but also compared with) the model's coefficient plots. An example of such global plots is shown in the Figure 1. These plots for the various classes confirm the already known importance of maternal characteristics in the fetal state evaluation. The models are notably impacted, in fact, by the **types of diabetes**, where type 1 and gestational diabetes tend to result in smaller fetal weight, contrasting with type 2 diabetes which leads to larger fetal weight predictions. Additionally, type 2 diabetes is linked to both normal and large fetuses in terms of predicted weight. **Maternal age** tends to predict in particular larger fetuses, as well as a high **number of pregnancies**.
At the same time, however, the importance of parameters such as the low **number of accelerations** in the FHR signal, common for the SGA and LGA classes, is highlighted, which distinguishes them from the normal weight category. Particularly relevant also seems to be the **LZC** index and the **MSE**. A lower entropy, in the time scales, is associated with the NGA, while a higher entropy is associated with the SGA class. Importance has also been given to frequency domain parameters **VLF, LF, MF, and HF** even if these features did not show discriminant capabilities.
Another type of SHAP plot, namely the waterfall plot, was used to highlight the impact of variables on classification of individual samples. The graph was initially used to analyze samples that were incorrectly classified. In this instance, it was noted that parameters deemed globally important also contribute significantly to errors. This occurs when their behavior deviates from the global contribution indicated in the summary plots. Moreover, the same type of plot was also used to investigate those samples for which parameters calculated in windows, in the absence of activity segments in the FHR signal, were assigned a value of -1. For such samples, this plot shows how these parameters are indeed considered most important for prediction purposes as shown in the Figure 2 for a sample of a SGA classes predicted by MLP model.

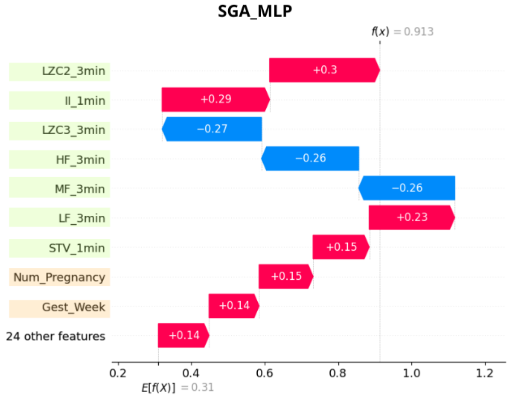

Figure 2: Example of parameters' contribution without activity segment in the FHR

Finally, another SHAP plot, namely the interaction plot, was used to observe a possible change in the most important variables depending on the gestational week and how this relationship influenced the models across different classes.

6. Conclusions

The study aimed to develop machine learning classifiers to predict fetal weight classes (namely SGA, NGA and LGA) in pregnancies complicated by maternal diabetes. Results showed Lo-

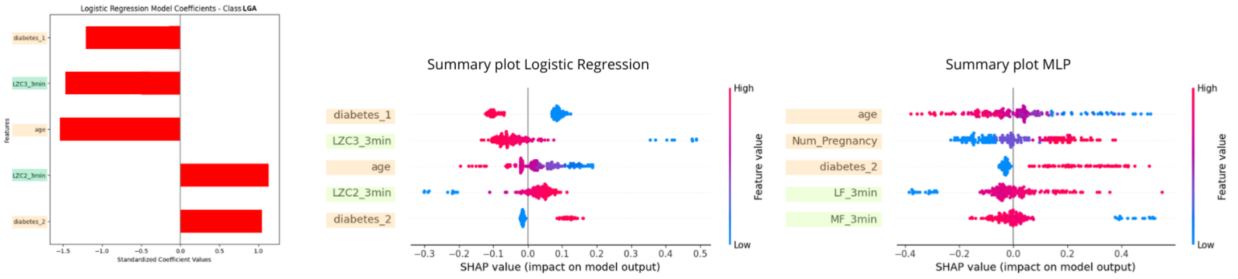

Figure 1: Model coefficients and SHAP summary plots for the LGA class

gistic Regression achieving 54.7% accuracy and MLP model reaching 52.6%, further improved with majority voting. Analyses using the SHAP method underscored the significance of the combination of variables, notably, maternal clinical history and FHR parameters. It has been observed that maternal factors such as type 2 diabetes can influence fetal weight prediction, leading to classifying the fetus as NGA or LGA, likely due to the effect of hyperinsulinemia on fetal growth. Additionally, an increase in accelerations in the FHR signal and a higher presence of activity, indicating advanced development of the fetal autonomic system, may be associated with a lower risk of weight problems for the fetus. Some variables instead have shown a different impact than anticipated by previous studies (such as the entropy related ones as MSE, which shows higher signal complexity values for the small weight class compared to the normal weight one), or less significant than expected (for example the PRSA derived parameters). This reinforces the idea that parameters effective in distinguishing between healthy and diseased groups may not be equally capable of differentiation within a solely pathological group. A challenge in this work concerned the textual format of clinical annotations in the database. This has made it necessary to use less efficient and precise methods of information selection, given the variety of terms, abbreviations, and presence of errors, and it might have led to the loss of relevant information. Nonetheless, integrating functional insights alongside maternal information has been shown in this work to represent a first attempt to propose a methodology that may lead to advancements in medical practice, albeit not ready for immediate clinical use yet. It is acknowledged that the obtained results may appear weak in terms of discrimination capability. However, diabetes complications during pregnancy present a challenging aspect in fetal monitoring and pregnancy management. The approach proposed entails a pipeline for a multi-feature machine learning model with explainability characteristics, which could represent a step forward in a personalized pregnancy medicine approach.

References

[1] World Health Organization. Definition, diagnosis and classification of diabetes mellitus and its complications: report of a who consultation. part 1, diagnosis and classification of diabetes mellitus. Technical report, World health organization, 1999.

[2] M. G Signorini, N. Pini, Malovini, and et al. Integrating machine learning techniques and physiology based heart rate features for antepartum fetal monitoring. *Computer Methods and Programs in Biomedicine*, 185:105015, 2020.

[3] E. Spairani, B. Daniele, Magenes, and et al. A novel large structured cardiotocographic database. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1375–1378. IEEE, 2022.

[4] Kiserud T., Piaggio G., Carroli G., and et al. The world health organization fetal growth charts: a multinational longitudinal study of ultrasound biometric measurements and estimated fetal weight. *PLoS medicine*, 14(1): e1002220, 2017.

[5] M. G. Signorini, N. Pini, A. Malovini, R. Bellazzi, and G. Magenes. Dataset on linear and non-linear indices for discriminating healthy and iugr fetuses. *Data in Brief*, 29:105164, 2020.

[6] L. Stroux, C. W Redman, A Georgieva, and et al. Doppler-based fetal heart rate analysis markers for the detection of early intrauterine growth restriction. *Acta obstetricia et gynecologica Scandinavica*, 96(11): 1322–1329, 2017.

[7] E. Spairani, G. Steyde, L. Subitoni, and t al. A semi-supervised deep learning approach to automate the identification of fetal behavioral states in fetal heart rate tracings. 2024.

[8] N. V Chawla, K. W Bowyer, L. O Hall, and et al. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[9] Christoph M. *Interpretable Machine Learning A Guide for Making Black Box Models Explainable, II edition.* 2023.