POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

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

Tesi di Laurea Magistrale in
Biomedical Engineering - Ingegneria biomedica

Author: **Elena Novelli**

Student ID: 103301
Advisor: Prof. Maria Gabriella Signorini
Co-advisors: Giulio Steyde
Academic Year: 2023-24

Abstract

During pregnancy, careful monitoring is crucial to prevent complications, with particular attention to diabetes mellitus, which increases the danger of issues for both the mother and the fetus. The most commonly employed examination in clinical practice during pregnancy is Cardiotocography (CTG) and this study focuses on the impact of maternal history and parameters derived from fetal heart rate (FHR) extracted with computerized CTG (cCTG) on the classification of fetal weight categories, particularly Small for Gestational Age (SGA), Normal for Gestational Age (NGA), and Large for Gestational Age (LGA). Weight pathologies during fetal development are indeed among the complications most closely linked to diabetes. Machine learning classifiers were trained using maternal clinical features and FHR signal parameters obtained from cCTG systems. Logistic Regression models and Multilayer Perceptron (MLP) models were employed to predict weight categories. The first one achieves a balanced accuracy of 54.7%, while the second one reaches 52.6% in three class classification. Both results are increased if a majority voting is performed, suggesting that a larger number of records can enhance the accuracy of the prediction. For interpreting the results, SHapley Additive exPlanations was used, a method for explaining machine learning models that employs coalition game theory and Shapley values to interpret the predictions of such models, offering insights into both feature importance and interaction effects. The interpretation of results indicates that maternal history variables play a significant role, as it was known in clinical practice, in predicting weight categories. However, this study highlights how the contribution of parameters from cCTG, in particular the complexity index of Lempel-Ziv, the number of acceleration of the signal, the multiscale entropy and the percentage of activity segments, is also fundamental compared to maternal variables. Some variables have shown a different or less significant impact than expected by previous studies, underscoring that parameters distinguishing between healthy and diseased groups might not equally differentiate within a solely pathological group. The study highlighted so the importance of combination of maternal history and fetal signal parameters in predicting fetal weight categories, aiming to enhance fetal well-being monitoring and clinical decision-making in pregnancies compicated by diabetes mellitus.

Keywords: computerized cardiotocography, diabetes, pregnancy, fetal weight, machine learning, explainability

Abstract in lingua italiana

Durante la gravidanza, un monitoraggio attento è cruciale per prevenire complicazioni, con particolare attenzione al diabete mellito, che aumenta il rischio di problemi sia per la madre che per il feto. L'esame più comunemente impiegato nella pratica clinica durante la gravidanza è la cardiotocografia (CTG) e questo studio si concentra sull'impatto della storia materna e dei parametri derivati dalla frequenza cardiaca fetale (FHR) estratta con la CTG computerizzata (cCTG) sulla classificazione delle categorie di peso fetale, in particolare Piccolo per l'età gestazionale (PEG), Normale per l'età gestazionale (NEG) e Grande per l'età gestazionale (GEG). Le patologie legate al peso durante lo sviluppo fetale sono infatti tra le complicazioni più strettamente legate al diabete. Classificatori di apprendimento automatico sono stati addestrati utilizzando caratteristiche cliniche materne e parametri del segnale FHR ottenuti dai sistemi cCTG. Sono stati impiegati modelli di regressione logistica e modelli di perceptron multistrato (MLP) per prevedere le categorie di peso. Il primo raggiunge un'accuratezza bilanciata del 54,7%, mentre il secondo raggiunge il 52,6% nella classificazione a tre classi. Entrambi i risultati aumentano se viene eseguito un voto maggioritario, suggerendo che un maggior numero di registrazioni può migliorare l'accuratezza della previsione. Per interpretare i risultati, è stato utilizzato SHapley Additive exPlanations, un metodo per spiegare i modelli di apprendimento automatico che utilizza la teoria dei giochi di coalizione e i valori di Shapley per interpretare le previsioni di tali modelli, offrendo approfondimenti sia sull'importanza delle caratteristiche che sugli effetti di interazione. L'interpretazione dei risultati indica che le variabili della storia materna giocano un ruolo significativo, come era noto nella pratica clinica, nella previsione delle categorie di peso. Tuttavia, questo studio evidenzia come il contributo dei parametri dalla cCTG, in particolare l'indice di complessità di Lempel-Ziv, il numero di accelerazioni del segnale, l'entropia multiscala (MSE) e la percentuale di segmenti di attività, sia fondamentale rispetto alle variabili materne. Alcune variabili hanno mostrato un impatto diverso o meno significativo rispetto a quanto previsto da studi precedenti, sottolineando che i parametri che distinguono tra gruppi sani e malati potrebbero non differenziare in modo equo all'interno di un gruppo esclusivamente patologico. Lo studio ha quindi evidenziato l'importanza della combinazione di storia materna e parametri del segnale fetale nella previsione delle categorie di peso fetale, mirando a migliorare il monitoraggio del benessere fetale e la decisione clinica nelle gravidanze complicate dal diabete mellito.

Parole chiave: cardiotocografia computerizzata, diabete, gravidanza, peso fetale, machine learning, explainability

Contents

1 | Introduction

For the maternal-fetal system, pregnancy itself does not entail a life-threatening risk, but the antepartum period is typically a delicate phase that requires careful monitoring to prevent negative consequences. Despite advancements in perinatal and medical care, in fact, complications can still arise, significantly impacting maternal and neonatal health.

Among the most significant risks, there is diabetes, a condition that can develop during pregnancy or exist prior to conception. Diabetes increases the risk of complications for both the mother and the fetus, including suboptimal glycemic control, an increased risk of gestational hypertension, and a higher likelihood of preterm delivery and hypoxia. Additionally, diabetes can elevate the risk of developing other complications such as fetal weight related issue (for example, fetal macrosomia) which can lead to difficulties during delivery and increase the risk of birth injuries. Early identification and careful monitoring of complications during pregnancy are crucial to ensuring the best possible care for both mother and child. To this end, various techniques for fetal monitoring are utilized, including cardiotocography, a non-invasive examination that assesses fetal cardiac activity and uterine contractions. This monitoring method has become a primary tool for assessing fetal well-being during pregnancy, enabling medical professionals to promptly detect signs of fetal distress and take necessary measures to ensure a positive outcome for both [1, 2].

The following work will focus on the importance of fetal monitoring through cardiotocography and early identification of complications during pregnancy complicated by diabetes. Through a better understanding of these processes and proactive management of at-risk situations, it aims to provide tools that can help clinicians improving outcomes for mothers and their children.

In this chapter will be presented a small overview on the diabetes and some of its related issues in pregnancy, on the cardiotocography and finally their relationship for the prediction of the health for a newborn.

1.1. Diabetes

As an introductory section of the undertaken research, an overview of diabetes will be presented, encompassing the various types that will be considered in this study, along with its relation with pregnancy. Specifically, among the associated perinatal and neonatal morbidities, after a brief mention of fetal hypoxia and acidosis, the thesis will delve into those related to neonatal weight, upon which it will place particular emphasis.

1.1.1. Diabetes in pregnancy

The term **diabetes mellitus** describes a "metabolic disorder of multiple aetiology characterized by chronic hyperglycemia with disturbances of carbohydrate, fat and protein metabolism resulting from defects in insulin secretion, insulin action, or both. The effects of diabetes mellitus include long-term damage, dysfunction and failure of various organs" [3].

Numerous studies have investigated the relationship between different types of diabetes, pregnancy, associated risks, and birth outcomes. Specifically, attention is focused on type 1 and type 2 diabetes, as well as gestational diabetes mellitus (GDM).
"The **Type 1** category includes most cases which are primarily due to pancreatic islet beta cell destruction (which leads to onset earlier in life, the need for insulin therapy, and the potential development of vascular, renal, and neuropathic complications) and are prone to ketoacidosis. Type 1 includes those cases attributable to an autoimmune process, as well as those with beta-cell destruction and who are prone to ketoacidosis for which neither an aetiology nor a pathogenesis is known" [3]. Type 1 diabetes, also known as juvenile diabetes due to its early onset, is an autoimmune pathology characterized by a deficiency of insulin production. This condition is often diagnosed in the initial years of life. A substantial genetic predisposition is implicated in the manifestation of this ailment, with hereditary factors playing a pivotal role in the individual's propensity to develop the disease [4].
The category named **Type 2** also called type 2 diabetes mellitus (T2DM) "includes the common major form of diabetes which results from defect(s) in insulin secretion, almost always with a major contribution from insulin resistance" [2, 3]. This insulin resistance often becomes apparent in the later stages of life. The development of type 2 diabetes is linked to a spectrum of risk factors. Modifiable factors include obesity and lifestyle choices, while non-modifiable factors encompass genetic predispositions [4].
Gestational diabetes mellitus is a particular type of diabetes specific of the pregnancy period and is in fact defined as “glucose intolerance with onset or first recognition during

pregnancy (during routine testing in pregnancy, generally between 24 and 28 weeks, that does not meet the criteria of pre-existing diabetes)" [3, 5].

The scientific literature shows that hyperglycemia associated with diabetes is one of the most common medical conditions women encounter during pregnancy. The International Diabetes Federation (IDF) estimates that one in six live births (16.8%) are to women with some form of hyperglycemia in pregnancy. 16% of these cases may be due to diabetes in pregnancy (DIP) but the majority (84%) is due to gestational diabetes mellitus [6].

DIP manifests either as pre-existing diabetes predating pregnancy (that is diabetes diagnosed before pregnancy so type 1 or type 2) or as diabetes first identified in the considered pregnancy (when hyperglycemia detected during pregnancy meets criteria for diagnosing diabetes outside pregnancy).
The severity of hyperglycemia existing at conception and during the phase of generation of the organs heightens women's vulnerability to complications and in fact perinatal mortality rates in women with pre-existing diabetes remain increased 1- to 10-fold compared to women without diabetes [7]. Undiagnosed pre-existing diabetes before the pregnancy may entail complications like eye and kidney pathologies, markedly increasing pregnancy risks. Hyperglycemia during the genesis of the organs elevates the risk of spontaneous abortions and congenital anomalies. Diabetes in pregnancy, with its greater risk of hyperglycemia, can lead to fetal growth aberrations and macrosomia, contributing to short-term complications such as obstructed labor, obstetric emergencies, neonatal hypoglycemia, and potential neurological damage as well as hypertension, preterm delivery and caesarean delivery [5, 7]. There is also a risk of the onset or exacerbation of microvascular complications. Thus, meticulous glucose control before conception and throughout pregnancy is recommended.
Worldwide, the age at which T2DM develops is declining, leading to a growing number of women who were previously unaware of their T2DM status becoming pregnant, with their diabetes being diagnosed for the first time during routine pregnancy screenings. Conversely, women with a high risk of diabetes may experience the onset of the condition during pregnancy for the first time. Diagnostic criteria include fasting plasma glucose (FPG) $a \geq b$ 7.0 $mmol/L$ or 126 mg/dL, and/or 2-hour 75-g oral glucose tolerance test (OGTT) value $a \geq b$ 11.1 $mmol/L$ or 200 mg/dL, or random plasma glucose (RPG) $a \geq b$ 11.1 mmol/L or 200 mg/dL associated with signs and symptoms of diabetes [6]. In cases of DIP, the risk of complications is higher due to increased blood sugar levels, and it's unclear whether these levels were high before or during early pregnancy. While diabetes diagnosed during pregnancy could be type 1 or type 2, it's more likely to be type

2. Unlike gestational diabetes, DIP is often detected early, sometimes as soon as the first trimester, through appropriate testing. This emphasizes the importance of identifying and managing diabetes early in pregnancy to reduce potential problems. [6].

Because many women do not receive screening for diabetes mellitus before pregnancy, it can be challenging to distinguish GDM from preexisting diabetes [8]. The GDM group comprises a diverse range of women with different metabolic characteristics when influenced by pregnancy hormones. Different scenarios may include:

- Hyperglycemia that likely preceded the pregnancy.
- Reduced and/or falling insulin secretory capacity (e.g. developing type 1 diabetes).
- Significant insulin resistance from early pregnancy (e.g. polycystic ovary syndrome, women with overweight or obesity) and other combination of factors as the predisposition of type 2 diabetes.

Many obstetricians or obstetric care providers use the two-step screening process to identify the gestational diabetes. This approach is based on first screening with the administration of a 50 g oral glucose solution followed by a 1-hour venous glucose determination. Women whose glucose levels meet or exceed an institution's screening threshold (that varies from 130 mg/dL to 140 mg/dL) then undergo a 100 g, 3-hour diagnostic OGTT. Gestational diabetes mellitus is most often diagnosed in women who have two or more abnormal values on the 3-hour OGTT [8].
GDM is associated with a higher incidence of maternal morbidity including cesarean deliveries, shoulder dystocia, birth trauma, hypertensive disorders of pregnancy (including preeclampsia), and subsequent development of T2DM. Perinatal and neonatal morbidities also increase (especially if the pathology is untreated); the latter include macrosomia, birth injury, hypoglycemia, polycythemia and hyperbilirubinemia but also birt trauma and stillbirth [7, 8].

1.1.2. Fetal hypoxia and acidosis

Fetal acidosis and hypoxia in offspring of diabetic mothers are significant in scientific inquiry. Maternal, placental, and fetal factors can lead to oxygen deprivation in the fetus, causing acidosis. Gestational diabetes may elevate catecholamine levels in fetuses, exacerbating hypoxic stress during delivery [9]. The placenta's metabolism of glucose can increase lactic acid production, affecting fetal blood pH in cases of maternal hyperglycemia [9]. Continuous hypoxia may escalate during delivery, posing risks like brain damage and acidosis, elevating intrauterine fetal death risk [9–11].

Umbilical artery blood gases and lactate serve as reliable fetal hypoxia markers [12]. Infants born to gestational diabetic mothers exhibit biochemical differences at delivery, including lower umbilical vein oxygen and higher lactate and pCO2 levels [11, 13]. Lower fetal oxygen and higher lactate levels may indicate increased fetal metabolism due to maternal hyperglycemia and hyperinsulinemia [13].

The antepartum cardiotocography monitoring discussed in Section 1.2 is utilized to detect fetuses at risk of intrauterine hypoxia and acidemia. However, studies linking this technique to maternal diabetes and fetal hypoxia remain limited, which is why this study chose to focus on this topic at first place. It should be noted, however, that the analysis of this issue did not yield satisfactory results in predicting blood gases and lactate values based on the available information in this project. Therefore, the focus primarily shifted to diabetes-related weight issues, which will be introduced in the next subsection 1.1.3.

1.1.3. Small and large for gestational age

Despite improvements in pregnancy outcomes for women with diabetes, as previously stated, newborns still represent a high-risk population. In addition to those complications already mentioned, attention is now drawn to the adverse effects of maternal diabetes on embryonic growth retardation and excessive fetal development in relation to gestational age.
These categories of fetuses, which will now be presented, are responsible for a significant portion of morbidity and are associated with high healthcare costs once born due to a higher need for and utilization of healthcare services, compared to a normal fetal population [14].

Small for gestational age

Small for gestational age (SGA) denotes a clinical condition characterized by fetal weight falling below the 10th percentile of the general population considering gestational age [15].

Alterations in maternal homeostasis associated with diabetes can negatively impact the fetal developmental environment [15]. The SGA newborns typically exhibit elevated red blood cell counts, reduced body temperature, and glucose levels, heightening their risk of mortality and morbidity during the neonatal period and beyond. Additionally, they face an increased likelihood of developing diseases such as type 1 and type 2 diabetes mellitus, hypertension, obesity, and kidney disease in adulthood [16]. Diabetes can induce lipid peroxidation and diminishes collagen deposition within chorionic villi (CVs), crucial structures for feto-maternal exchange. These alterations in CV conformation have been associated with an imbalanced metabolic environment within the placenta[15]. These metabolic alterations can lead to a reduction in nutrient exchange between mother and fetus, potentially contributing to the development of an SGA fetus [15].

Large for gestational age

While macrosomia is commonly defined as a birthweight of 4000 g, the term "large for gestational age" (LGA) more broadly defines fetal overgrowth relative to gestation.
A baby weighing above the 90th birthweight centile is typically considered large for gestational age.
However, it's worth noting that these terms are often used interchangeably in clinical practice [17].

A widely acknowledged factor contributing to excessive intrauterine fetal growth is maternal insulin resistance, thus correlating with LGA and maternal diabetes (all of the 3 types shown earlier). Elevated maternal blood glucose levels result in increased glucose transportation to the fetus through the placenta, consequently inducing fetal hyperglycemia and hyperinsulinemia, thereby stimulating fetal growth. Studies have demonstrated a significant association between increasing maternal glucose levels and an elevated likelihood of birth weight surpassing the 90th percentile, with this association strengthening across ascending glycemia categories in particular starting from the third trimester [5]. LGA is marked by increased need for caesarean delivery, prolonged and complicated labor due to physical size that can led to birth injuries, postpartum haemorrhage, shoulder dystocia, low Apgar score, admission to neonatal intensive care unit, and severe neonatal morbidity and perinatal mortality [14, 17]. The identification and management of LGA is important, as it significantly increases the risk of other birth complications.

1.2. Cardiotocography

As previously stated, the most commonly employed examination in clinical practice during pregnancy is Cardiotocography (CTG), a monitoring practice utilized in Obstetrics and Gynecology clinics to assess fetal well-being, through the simultaneous analysis of the Fetal Heart Rate (FHR) and the Uterine contraction signals by a pressure sensor (TOCO) [18]. These signals are recorded simultaneously using two distinct transducers placed on the mother's abdomen. The first transducer is positioned near the fetal back to record the FHR, while the second transducer, placed in proximity to the uterine fundus, detects uterine contractions [18].

The computerized version of CTG (cCGT) is now widely employed, particularly to address challenges related to intra- and inter-subject variability, and the limitation of manual interpretation in extracting quantitative information from the FHR signal. This modern version allows for the derivation of various numerical parameters linked to fetal conditions, encompassing morphological, frequency, and nonlinear/complexity indices. These parameters play a critical role in encapsulating the diverse pathophysiological aspects of FHR [19].

The cCTG consists of two main elements:

1. A cardiotocograph that records the fetal heart rate signal, uterine contractions, and fetal motor activity using Doppler or heartbeat detection.
2. Software installed on a standard laptop that processes and analyzes the data collected by the cardiotocograph

As said, the movement of the fetal heart which is externally detected through an abdominal probe via the ultrasound (US)-Doppler method. The Doppler effect relies on the principle that an ultrasound beam is reflected from a surface with a frequency different from that of emission, proportional to the characteristics of the movement of the reflecting surface. The ultrasound beam emitted by a transducer applied to the maternal abdomen and oriented towards the fetal cardiac region detects movements of the cardiac valves and allows the use of the opening and closing echoes of the cardiac valves as triggers to define the duration of the cardiac cycle [20].
A pulse, typically consisting of several cycles of 1–2 MHz sinusoids, is transmitted toward the fetal heart. The reflected pulse, slightly shifted in frequency (Doppler shift) by contractions of the fetal heart, is compared with the transmitted pulse (demodulation).

1.2.1. Fetal heart rate

The Fetal Heart Rate signal is a vital parameter whose non-invasive monitoring plays a pivotal role in assessing fetal well-being, especially in pregnancies with risks and in predicting adverse conditions [21]. In particular, the variability of the fetal heart rate has been identified as a crucial indicator to assess fetal well-being both before and during delivery. A decrease or absence of fluctuations is often associated with fetal distress conditions. Such distress is generally interpreted as a situation that negatively impacts the fetal health status and is closely linked to irregularities in the fetal heart rate signal [21, 22]. The FHR variability serves as a reliable indicator of the synergistic activity of the autonomic nervous system, which regulates heartbeat dynamics. Numerous experimental studies conducted on large populations, both in physiological and pathological conditions, have highlighted the effectiveness of HR analysis in noninvasive but quantitative monitoring of cardiovascular control systems. Additionally, FHR parameters play a crucial role in distinguishing between pathological states, offering valuable insights into the onset of disease conditions. These properties of FHR analysis hold particular promise in monitoring fetal well-being during the antepartum period, as a significant portion of adverse fetal outcomes is attributed to events occurring before the onset of labor, manifesting as alterations in FHR itself.

Furthermore, changes in FHR often precede and serve as predictors of fetal distress and adverse conditions before the emergence of recognizable symptoms. The analysis of FHR signals also provides valuable insights into the neural development of the fetus, encompassing both linear and non-linear aspects. The autonomic nervous system (ANS), which regulates nearly every organ system, contributes in fact to the beat-to-beat variation of FHR, reflecting the influence of the fetus' autonomic nervous system components, including sympathetic and parasympathetic pathways [1, 23].

Fetal behavioral states

Research on infants, both premature and full-term, has revealed essential parameters for assessing neurological health, particularly related to sleep. These studies have highlighted the significance of sleep duration and characteristics. Notably, there is a distinction between quiet sleep (non-REM) and active sleep (REM) based on observations of muscle tone, eye movement, brain activity, and heart rate. These states are important markers throughout the day, reflecting the fetus's well-being.

When interpreting FHR tracings, it's vital to consider fetal rest-activity patterns. Fetal heart rate changes correspond to different sleep states, indicating fetal health. A healthy fetus experiences cycles of active and quiet sleep starting from around 28 weeks of gestation. Active sleep phases, marked by movement, accelerations, and high heart rate variability (HRV), is a signal of fetal well-being. On the other hand, quiet sleep periods represent physiological period of rest intervals, characterized by minimal movement and low HRV [24].

Already in the 1970s, the first experimental observations of fetal quietude and activity cycles were presented by Dawes and Ruckebush. Sterman and Hoppenbrowers hypothesized the presence of fetal sleep and wakefulness phases. The identification of two distinct patterns of fetal cardiac activity, cyclically alternating and fundamentally physiological, was proposed by Timor-Trisch and confirmed by Romanini in 1984. These findings led to significant changes in obstetric routines [25].

An example of a FHR signal with visible pattern of active and quite sleep can be seen in Figure 1.1.

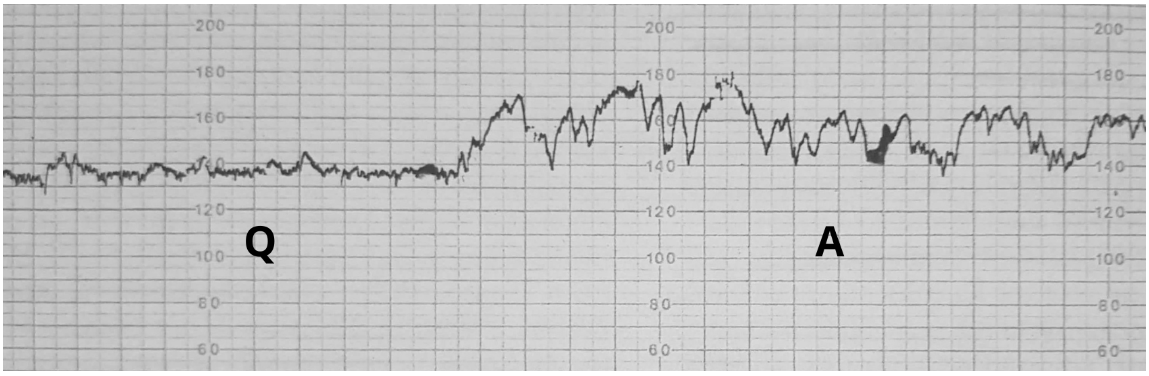

Figure 1.1: Example of an FHR signal with quiet pattern (Q) followed by an activity period (A). Adapted from [25]

Delving deeper into the cycle of "active and quiet sleep", different behavioral states can be explored.

The definition of a behavioral state is based on simultaneous monitoring of fetal heart rate, fetal body movements, and rapid eye movements. To correctly identify fetal states, it's essential to record heart rate using a cardiotocograph, while body movements and eye movements must be recorded using two different ultrasound machines. After completing the recording, fetal variables are analyzed using the moving window technique, usually for 3 minutes, to reduce the impact of acute and short-duration events and increase the stability of the observed phases. The states can then be described according to the classification by Nijhius and Prechtl:

- 1F (Quiet sleep): Absence of spontaneous motor activity or detection of startless somatic movements, absence of ocular motor activity; fetal heart rate (FHR) shows quiet or silent pattern, with no accelerations and reduced variability.

- 2F (Active sleep): Presence of spontaneous fetal motor activity involving the body and limbs; presence of spontaneous ocular activity, both with rapid and slow movements of the fetal lens; FHR pattern shows numerous accelerations coinciding with fetal motor activity.

- 3F (Quiet awake): Absence of somatic movements; continuous ocular movements; FHR pattern characterized by larger fluctuations compared to 1F, but with no accelerations.

- 4F (Active awake): Continuous motor activity, involving fetal trunk structures with rotational movements; presence of rapid ocular movements; unstable FHR pattern due to prolonged and wide accelerations, which merge with brief periods of tachycardia.

From a clinical standpoint, understanding the progressive development of behavioral states during normal pregnancy provides a valuable reference framework for assessing fetal health. This allows for comparing observed patterns in pathological cases with standards of normality established for the corresponding gestational age. Knowledge of these differences can guide clinical practice in managing high-risk pregnancies and in promptly identifying potential complications [24, 25].

1.2.2. Applications of cCTG: monitoring pregnancies complicated by maternal diabetes or growth disturbances

As stated, diabetes mellitus, along with its consequences, can be a risk factor and therefore a subject of particular scrutiny in such pregnancies. CTG can be employed for this purpose and works in literature testify it.

For example, in Signorini et al.'s study titled "Linear and Nonlinear Parameters for Analyzing Fetal Heart Rate Signals from Cardiotocographic Recordings," the focus is on exploring the connection between CTG and gestational diabetes to enhance antepartum fetal monitoring. The work highlights the limited capability of current CTG analysis in providing precise indications of fetal status, especially in the presence of gestational diabetes. A novel methodological approach is proposed, involving the analysis of spectral parameters of fetal heart rate and the application of nonlinear algorithms. These new parameters aim to provide more detailed information on fetal status, including the ability to early recognize anomalies associated with gestational diabetes. The goal is to improve early diagnosis of common fetal pathologies, enabling timely preventive interventions to enhance pregnancy outcomes and reduce risks to both the child and the mother.

Another article from S. M. Lobmaier et al. called “Influence of gestational diabetes on fetal autonomic nervous system: a study using phase-rectified signal-averaging analysis “ aimed to evaluate the influence of gestational diabetes on fetal cardiovascular function and the autonomic nervous system using cardiotocography and the signal processing technique called phase-rectified signal averaging (PRSA). The study highlighted increased activity of the autonomic nervous system in fetuses of diabetic mothers during late gestation. Analysis of fetal cardiovascular function and ANS using PRSA may offer better surveillance compared to conventional techniques, potentially linking gestational diabetes with future cardiovascular dysfunction in newborns.

The CTG has also been widely used for investigations concerning fetal weight, particularly regarding intrauterine growth restriction.
Intrauterine growth restriction (IUGR) is a pathological decrease in fetal growth rate

whereby the fetus cannot reach its full genetic growth potential and is associated with significantly increased mortality and morbidity.

In the article “Doppler-based fetal heart rate analysis markers for the detection of early intrauterine growth restriction”, Lisa Stroux et al. aimed to assess whether markers reflecting the impact of intrauterine growth restriction on the cardiovascular system, computed from a Doppler-derived heart rate signal, would be suitable for its antenatal detection. Using a cardiotocography archive of IUGR cases and healthy controls matched for gestation and gender, the study evaluates the discriminative power of short-term and long-term variability of the fetal heart rate and metrics characterizing sleep state distribution within a trace. Results indicate that heart rate variability markers, along with information on sleep states, can contribute to the early detection of IUGR, particularly before 34 weeks of gestation.

Also Singorini et al. again have dealt with this topic in “Integrating machine learning techniques and physiology based heart rate features for antepartum fetal monitoring”. The paper endeavors to develop an accurate model for the early identification of IUGR during pregnancy using cardiotocography . Fifteen machine learning methods were tested to discriminate between healthy and IUGR fetuses using features of the heart rate extracted from CTG recordings. Results showed that the Random Forest method achieved the best performance. Nonlinear heart rate indices demonstrated greater discriminative capability between fetal conditions, but a combination of CTG features contributed to improved classification accuracy. This study presents a valid approach for diagnosing IUGR in the antepartum period, providing an interpretable link between machine learning outcomes and quantitative assessments of fetal well-being.

1.3. Aim of the work

As previously stated, cardiotocography has been employed in studies related to fetal weight disorders at birth as well as in studies focused on diabetes in pregnancy. However, the relationship between these 2 main topic has not been extensively explored within this context. The aim of this study was to harness measurable parameters from fetal heart rate signals, acquired via cCTG, and maternal clinical information, focusing on pregnant women diagnosed with diabetes. The goal was to develop a machine learning classifier capable of predicting fetal weight. This classifier categorizes fetuses into three groups: small, normal, and large for gestational age where these categories have been determined by the distribution of the estimated fetal weight (EFW) defined by the World Health Organization (WHO). This was done while investigating the relationship between parameters derived from FHR, clinical maternal information and fetal weight itself.Specifically, the significance of the CTG's parameters was examined in comparison to and in correlation with general physiological parameters already accessible to healthcare professionals without the need for cardiotocographic examination, such as the generic characteristics of the pregnant woman (that can be age, gestational week or number of pregnancy and above all the presence and type of diabetes). The importance of variables derived from maternal medical history is already known, and progress has been made in selecting groups of parameters derived from CTG analysis that are indicative of fetal well-being. Therefore, further investigation into their relationship is desired to address the expressed issue and provide an initial step towards an enhanced approach to guiding and assisting clinicians in identifying potential high-risk pregnancies, with the primary goal of safeguarding fetal well-being.

2 | Materials and Methods

2.1. Database description and 2CTG software

The data employed in this thesis work were obtained from a very large database of CTG recordings. These recordings were gathered at the Department of Neuroscience, Reproductive Sciences, and Dentistry within the Gynecology and Obstetrics Unit, School of Medicine, Federico II University of Naples, Italy. This data collection spanned from 2013 to 2021 and was obtained as a results of routine antepartum fetal monitoring examinations. The cohort comprised 9,476 pregnant individuals, yielding a total of 24,492 tracings. Each tracing was categorized by medical professionals to denote either a healthy pregnancy or one with various complications. Additionally, personal details regarding the maternal-fetal system but also maternal clinical history were included, along with a set of quantitative parameters assessing different statistical characteristics of the cCTG signals.The dataset is further detailed in [26]. Monitoring was conducted in a controlled clinical setting with patients positioned on an armchair. cCTG records, lasting between 20 to 60 minutes, were obtained utilizing Philips Avalon family monitors equipped with ultrasound transducers and transabdominal tocodynamometers. These fetal monitors employ autocorrelation techniques to analyze Doppler signals of heartbeats, converting them into heart periods (equivalent to RR intervals) and subsequently into heart frequencies in beats per minute. Philips monitors generate FHR values every 250 milliseconds (4 values/second) [27].

The cardiotocograph was linked to the **2CTG system**, enabling dual readings of FHR and toco signals twice per second, resulting in 120 data points per minute [28]. A sampling frequency of 2 Hz was chosen as a balanced compromise to ensure sufficient bandwidth and acceptable accuracy for advanced analyses, including nonlinear parameters. The TOCO signal is sampled at the same frequency as the FHR signal to ensure accurate alignment of the data, even though the frequency band of the TOCO signal is lower than that of the cardiac variability.

The 2CTG software performs real-time calculations for every minute of recording, processing both the baseline and all classical analysis parameters (including accelerations, decelerations, short- and long-term variability, contractions, maximum and minimum FHR) within seconds. This allows for continuous reanalysis of the entire preceding tracing with each new minute of recording, ensuring that the user has the most up-to-date information for all the time points [29]. The software interface provides user-friendly interaction. 2CTG can capture and analyze the tracing concurrently with the operator inputting patient demographics and medical history data. This means that monitoring can commence immediately upon connecting the cardiotocograph transducers, and data entry can be performed via the computer keyboard while the examination is ongoing. The functionality of 2CTG involves transmitting FHR and tocometry values, which are recorded by the cardiotocograph every 250 milliseconds, to the PC in serial fashion [29].

The database provided for this work was anonymized (ensuring that no personal sensitive information pertaining to the patient was included) and presented, as said, features regarding the pregnant woman and her clinical history and outcome from the cCTG.
As regards the first group there are:

- PREGNANCY_ID: a unique numerical code assigned to each pregnancy, which remains the same even if the woman is associated with multiple records in the database. If a woman has multiple pregnancies, she is associated with different ID;
- VISIT_ID: a unique sequential numerical code associated to each recording. The higher the number, the more recent the registration;
- GEST_WEEK: an integer number representing the gestational week of the pregnancy. With this data, the gynecological standard understands the integer number of weeks elapsed since the last menstruation. Each rounding is therefore done downwards;
- NUM_PREGNANCY: an integer representing how many pregnancies the woman has had previously;
- AGE: an integer number indicating the age of the pregnant woman;
- LAST_MENSTRUATION: the date of the last menstruation;
- OUTPATIENT_CODE: a numerical code ranging from 1 to 12 assigned by the physician to categorize the health condition of the maternal-fetal unit. It considers the findings from other visits and exams outside the cCTG and existing medical history;
- NOTE: a textual variable containing general information annotated by the clinician during the visit. Such information may include the date of delivery if it has oc-

curred and the mode (cesarean or spontaneous). It may also indicate the hospital department, the newborn's weight and/or the weight of the pregnant woman. Additionally, other fetal information such as gender, *Apgarscores* at 1 and 5 minutes, and blood gas values (PO_2, PCO_2, pH, *lactates*, and $BECF$) may be provided;

A first type of cCTG outcome are instead:

- FHR: the fetal heart rate signal detected by the software 2CTG;
- QUALITY: a set of variable length of integer values that represent the recording quality for each sample. This quality metric is determined by the 2CTG2 system using the results of the autocorrelation process. A score of 32 indicates excellent quality, while a score of 64 indicates acceptable quality, and a score of 128 indicates inadequate quality;
- TOCO: the trace representing the uterine contractions;
- FMP: during the examination, the mother reported instances of fetal movement by pressing a button on a handheld device connected to the fetal monitor, creating a binary array known as the Fetal Movement Profile. True values indicate moments when the mother perceived fetal movement;

2.1.1. Subset selection and target

The following steps detail the procedures executed on the recently introduced database, focusing on variables related to pregnant women, leading to the selection of traces and variables for subsequent preprocessing and signal processing phases. Special emphasis will be placed on the choice and selection of the final target of the work while mentioning the work done in the initial phase of the project regarding the study of fetal hypoxia and acidosis. All subsequent procedures were conducted using Matlab R2021b.

1. Given the importance previously highlighted regarding the issue of diabetes and the stated purpose of the study, the first step involved selecting only recordings from diabetic pregnant women from the main database. This selection was accomplished by creating a text analysis function that searched into the "NOTE" section for a range of expressions or words related to diabetes (e.g., "diab", "dm", "dgm", "gdm", "mellito", "d.g.") using regular expressions. The function also accounted for potential misspellings by the clinician. This resulted in the selection of 1869 recordings, corresponding to 7.63% of the initial database. It is worth mentioning that during this initial phase, the method applied excluded all recordings related to twin pregnancy from the selection. The rationale of this choice will be detailed in a

few steps;

2. For the selected recordings, an additional variable encoding diabetes type was added (1: type 1 diabetes, 2: type 2 diabetes, 3: gestational diabetes). This information was again obtained through regular expression searches within the note section;

3. The information contained in the "LAST_MENSTRUATION" variable was used and compared with the delivery date (which was present in most cases) from the note to determine the delivery week. Checks were performed to verify the plausibility of the results obtained, ensuring that the dates fell within possible limits;

4. Another step involved extracting the fetal weight value again from the notes. The strategy was to searches for the word "weight" (along with possible variations) in the "NOTE" string. Once the keyword was identified, it was looked for a four-digit numerical value within a specified range since fetal weight is expressed in grams in the notes (while the mothers' weight is expressed in kilos). If a match was found, it was considered as the weight and returned. If the value was less than 500, it is considered a typo. Excluding twin births ensures that there were no issues with assigning the weight to the proper owner;

5. Through the same search method, all terms related to fetal blood gas values measured at birth such as pH, indications of fetal arterial oxygen partial pressure (pO_2) or carbon dioxide partial pressure (pCO_2), as well as excess base of fetal circulation ($becf$) or the presence of *lactate*, were identified. Then, numerical values related to these indices (also considering the possibility of typos) were retrieved.

6. Textual research in the notes has allowed to obtain information also related to the mode of delivery, thus being able to distinguish between pregnancies brought to term with spontaneous delivery, cesarean delivery, urgent or not, and also indication of a previous cesarean delivery.

7. The number of records was increased by the presence of two additional secondary databases collected by the same research group, containing other examinations conducted on diabetic pregnant women. Proper records were attributed to the correct patients using their unique codes ($PREGNANCY\_ID$) as keys. For all those pieces of information not directly obtainable from the associated records in the main database, the previously explained methods were reapplied;

Brief discussion on the prediction of fetal blood gas values

As stated in Chapter 1.1.2, the condition of fetal acidosis and hypoxia in offspring of diabetic mothers is a significant scientific research topic. Maternal, placental, and fetal factors can cause oxygen deprivation in the fetus, leading to acidosis. Gestational diabetes, in particular, can increase catecholamine levels in fetuses, exacerbating hypoxic stress during childbirth.

With the available information, was decided to train a machine learning regression model capable of predicting the values of blood gas variables at birth, treating each variable separately as a target and omitting the others. The aim was to explore the tendency of some fetuses towards severe hypoxia or acidosis, using variables derived from both the FHR signal and the mother's clinical history as predictive features. The main targets were pCO_2, as hyper- and hypocapnia are correlated with respiratory and neurological complications, pO_2, and pH [30].

In all cases, the results were extremely inconclusive considering metrics such as Mean Squared Error, Mean Absolute Error, Root Mean Squared Error, and the Coefficient of Determination R^2, which reached a maximum value of 0.2.

To interpret this result, two gynecologists in charge of the project at the Federico II Hospital in Naples were consulted. They associated the outcome with the fact that childbirth can drastically alter blood gas and lactate values, which can vary significantly from the values predicted considering other variables such as maternal medical history and parameters extracted from the FHR signal acquired through cCTG, referring to periods prior to childbirth by weeks as well.

While it was known that childbirth could be a traumatic event for the fetus, the motivation behind predicting blood gas values was to investigate whether different fetal development, observable through temporal, frequency-based, or even nonlinear parameters extracted from the FHR signal, along with information about the mother's medical history, could anticipate such an event. Additionally, the aim was to understand if fetuses with a more developed system could better withstand this acute event.

Subsequently, attention was directed towards another aspect related to diabetes, namely fetal weight.

Final target

As stated earlier, the main and definitive focus of the study has been on weight issues resulting from diabetes presence. Given the purpose of the study, only those records containing weight values were ultimately selected from the final database. Additionally, among the extracted variables listed in the subset selection process, only those related to pre-partum information were retained (except for variables concerning the week of delivery and birth weight for reasons explained here). This exclusion also encompassed those variables related to found blood gas values similar. At the end of this selection process, the considered database consisted of 809 registrations corresponding to 149 patients.

From now on, focus will be only on the topic of weight, and all future steps explained in the work refer to this subject.

Weight categories, such as "Small for gestational age" , "Normal for gestational age" and "Large for gestational age", were defined based on a study conducted by the World Health Organization (WHO). This multinational prospective observational longitudinal study on fetal growth in pregnancies led to the creation of fetal growth charts for estimated fetal weight (EFW) as can be seen in Figure 2.1

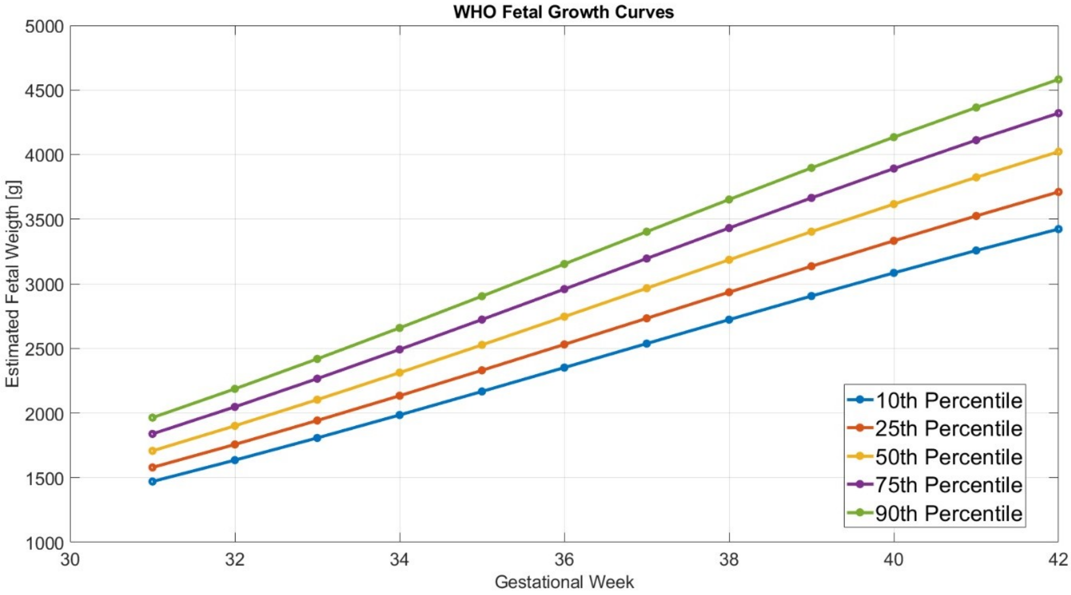

Figure 2.1: WHO chart of EFW.

To level the estimated quantiles and obtain a clearer and more uniform view of fetal growth, WHO used polynomial functions of gestational age [31].
To establish reference intervals, quantile regression was used. This technique makes an inference on the regression coefficients for the conditional quantiles of a variable without making assumptions about its distribution. Quantile regression is particularly useful in the study of distribution changes and has shown that fetal growth in the population is not symmetrical with gestation [31].

For the purpose of this study, which aims to categorize fetal weight, a function was created to calculate an estimated weight. This function uses a 3rd-degree polynomial model, where the coefficients represent the quantiles of the distribution of estimated fetal weights. In the original study, the independent variable is the gestational week but here is used the week of delivery, which was previously calculated. This operation was performed to find an a posteriori threshold to which compare the birth weight. Fist was found a threshold for the definition of SGA fetuses using the coefficients referring to the 10th percentile of the growth charts and then was found a threshold to identify the LGA fetuses using the 90th percentile. The choice of percentiles reflects the most widespread definitions of SGA and LGA (as reported in the appropriate section 1.1.3).
Each fetal weight was then compared with the two thresholds obtained and to the corresponding recording was assigned a target value according to the following Table 2.1.
The fetuses with weights between the two indicated thresholds were considered normal for gestational age (NGA) corresponding to "non-problematic" pregnancy.

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

Table 2.1: Summary table of the target.

As shown in the table, the most represented weight class in the database is that of fetuses considered normal followed by the small ones and the large ones. Nevertheless, it is worth noting that the distribution percentages of the target are 50.3 % for the normal class and 49.7 % for classes associated with weight-related conditions within the dataset. This distribution holds true even when considering the actual number of patients present in the database. Out of the 149 actual pregnant women, only 57% carry a non-growth affected fetus. This distribution deviates from what would be expected from a normal population, considering the definition given to the weight categories. In fact, if one considers that the "normal" population covers between the 10th and 90th percentiles, its representation in the dataset would be expected to be around 80% of the registrations. The fact that it represents only 50-60%, with the remainder related to a problematic population, can be viewed as an initial confirmation of the complications, risks, and challenges that diabetes poses during pregnancy, thereby justifying the analysis being undertaken.

It is important to specify right away that at the end of this target selection phase, information regarding the numerical value of birth weight and the week of delivery have been eliminated. Therefore, all subsequent analysis was conducted considering only variables (whether derived from maternal clinical history or parameters extracted from the FHR signal) available during pre-delivery pregnancy monitoring.

2.2. Preprocessing

Before proceeding with the analysis, it was necessary to preprocess the raw fetal signals provided by the cCTG. To process them, information from the "QUALITY" variable provided by the cCTG itself was used. Specifically, all samples for which the quality score was less than or equal to 64 were considered acceptable and left unchanged. For all samples with a higher score, or with values equal to 0, or deemed physiologically implausible through other checks, linear interpolations were performed.
The effect of this phase on a FHR signal can be seen in the figure 2.2.

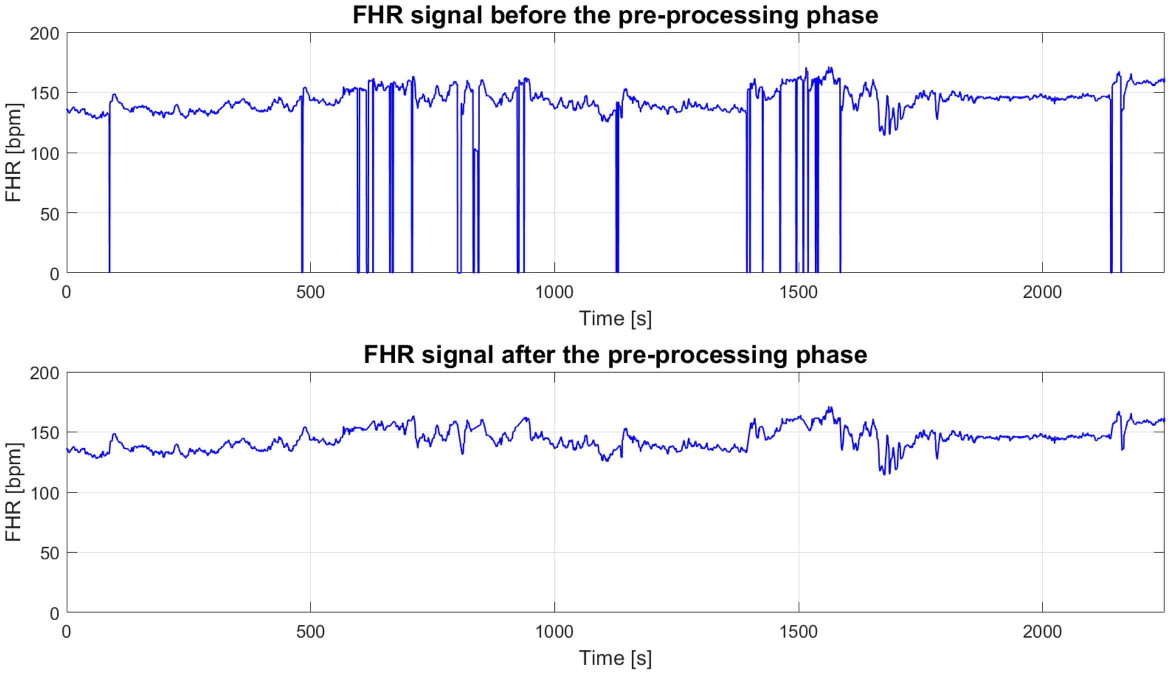

Figure 2.2: Preprocessing effect on a FHR signal. Interpolation in the second image can be seen where the signal in the first one was particularly corrupted.

The parameters described in section 2.3 are then calculated only for those signal segments where there is less than a 5% percentage of interpolated consecutive samples. These segments are considered to be of too low quality for their respective parameters to be considered significant, but their direct elimination could alter the signal morphology and affect the calculation of the parameters themselves.

In the preprocessing phase, baseline, accelerations, and decelerations of the FHR signal are also calculated. The baseline in fetal heart rate monitoring refers to establishing the average trend of the fetal heart rate signal, excluding physiological events like accelerations and decelerations. This results in the baseline FHR, expressed in beats per minute (bpm), which serves as a reference for identifying accelerations, decelerations, and estimating FHR variability.

In the utilized system, the baseline is continuously computed during signal recording using a multi-step digital filter based on adaptations of the Mantel et al. algorithm [32, 33]. This method dynamically updates the baseline value as new signal data is received, minimizing the impact of sudden signal fluctuations on parameter estimation. Acting as a filter, the baseline smooths out irregularities in the FHR signal, preventing inaccurate parameter estimations.

During baseline calculation, the acquired FHR signal is subsampled every 0.5 seconds, averaging every group of 5 distinct points to produce a signal consisting of 24 points per minute. This approach results in a more refined and precise representation of the fetal heart rate trend [32].

Accelerations are defined as events on the CTG trace during which the FHR, expressed in bpm, remains persistently above the estimated baseline value for a fixed period of time (in the system used, the threshold is 5 bpm). Further categorization of accelerations involves setting different threshold criteria to quantify the increase in FHR in bpm compared to the estimated baseline value, resulting in the distinction between small and large accelerations. Similarly, the identification of decelerations, whether they are large or small, follows an analogous process [32, 34].

2.3. Signal processing and parameters

Starting from the expressed preprocessing and the obtained signal, various parameters have been computed, which will be added to the maternal features mentioned earlier to become variables used in fetal weight classification.

A primary distinction among these parameters can be made between parameters in the time domain, parameters in the frequency domain, and nonlinear parameters.

Secondly, depending on the parameters and their significance, they have been calculated either in overlapping at 50% sliding windows of fixed length equal to 1 minute, or in 3-minute windows, or globally along a 20-minute segment of the signal (which is noted to be the minimum length for all) of better quality.

In the following table, the parameters calculated from the signal are introduced.

Parameter's name	Domain	Windows' length
Short Term Variability (STV)	Time	1 min
Interval Index (II)	Time	1 min
Delta (Δ)	Time	1 min
n_acc	Time	Global
n_dec	Time	Global
Very Low Frequency (VLF)	Frequency	3 min
Low Frequency (LF)	Frequency	3 min
Movement Frequency (MF)	Frequency	3 min
High Frequency (HF)	Frequency	3 min
Approximate Entropy (ApEn)	Non linear	Global
Sample Entropy (SampEn)	Non linear	Global
Multiscale entropy (MSE)	Non linear	Global
Sample Asimmetry (SampAsi)	Non linear	Global
Binary Lempel-Ziv (LZC2)	Non linear	Global/3 min
Ternary Lempel-Ziv (LZC3)	Non linear	Global/3 min
Acceleration capacity (AC)	Non linear	Global
Deceleration capacity (DC)	Non linear	Global
Deceleration reserve (DR)	Non linear	Global
Acceleration Phase Rectified Slope (APRS)	Non linear	Global
Deceleration Phase Rectified Slope (DPRS)	Non linear	Global
LFprsa	Non Linear/Frequency	Global
MFprsa	Non Linear/Frequency	Global
HFprsa	Non Linear/Frequency	Global

Table 2.2: Parameters derived from the FHR signal

Regarding the parameters computed in windows, three distinct methodologies have been followed in increasing order of complexity and relative appropriateness for the analysis:

- **APPROACH 1**: Initially, 1-minute and 3-minute windows were considered and defined throughout the entire signal. For each signal in the database, a number of values were obtained for all parameters indicated as 1 min and 3 min in the table, corresponding to the number of windows defined within the signal itself. Then, to obtain a single value for each parameter to be assigned to each signal, the window values were averaged;

- **APPROACH 2**: The second approach follows the same window identification methodology and parameter calculation but adds the additional variable of the percentage of activity segments present throughout the entire signal. This variable is called *Act_per*;

- **APPROACH 3**: In the final version of the work, 1-minute and 3-minute windows were considered and identified only in those segments of the FHR signal associated with periods of fetal activity of the fetal signal, excluding segments belonging to quiet zones. Again, the average was taken across the various segments to arrive at a single parameter value for each one considered, and the column with the percentage of activity was added back in. Any future analysis of the work done will refer to this approach;

For approaches 2 and 3, the identification of fetal states of activity and quietness was performed employing the model presented in [35], which was previously developed by our research group. In this framework, a deep neural network (DNN) with a 1D encoder-decoder architecture was implemented, aiming to recognize and automatically segment FHR recordings into active and quiet periods [36]. The process combines the use of Hidden Markov Models with segment annotations by clinicians, achieving an accuracy of 88%, deemed more than sufficient for the analysis [36].
The application of the network generated the following outputs for each signal:

- The percentage of activity throughout the entire signal and the percentage of quietness;

- All the indices indicating the start and end of activity segments and the same for quiet segments;

This network was then configured to ensure that segments of both activity and quietness shorter than 3 minutes were not captured, thus avoiding the presence of parameters belonging to multiple phases.

The choice to focus solely on the presence of activity and parameters in such sections is due to the fact that an instance of active sleep serves as a sign of fetal health and stands as one of the primary criteria utilized in the Dawes/Redman system to assess normality [24]. Additionally, several studies have demonstrated a lower occurrence of active sleep periods in problematic fetuses (such as those with IUGR) and the markers linked to active sleep episodes exhibit greater discriminatory capacity compared to those associated with quiet sleep episodes [24]. Given the complementary of the percentages and indices of the start and end of activity and quiet sleep, it would have been redundant to also report the indications from the quiet phase so all the indication regarding this phase have been removed.

However, there are signals within the database that have been identified as devoid of activity moments and thus consist purely of quiet sleep. These signals were not removed because they were considered to potentially contain relevant information given the characteristics already expressed regarding activity and quiet cycles. Nonetheless, since it was not possible to calculate the parameters at 1 minute and 3 minutes for these signals, in these cases, they were all set to a fixed value of -1.
To further distinguish these missing parameters to which placeholder values have been assigned, an additional boolean variable called d_par has been added to the database. This variable is set to 1 when there is no activity and 0 otherwise. This serves as an additional tool that informs the classifier that the parameter values in those specific rows are placeholders but should be taken into account.

An in-depth analysis of the parameters that will be used for the classification algorithm and presented in the table 2.2 will now be provided.

2.3.1. Time domain parameters

Before computig time-domain parameters pre-processing steps were applied, as deescribed in 2.2. The signal, originally expressed in bpm, was transformed into milliseconds. Then, another step is computed to derive a downsampled version of the signal, called T [ms], by averaging the signal in non-overlapping windows of 5 samples, resulting in a "sampling frequency" of 0.4 Hz. This procedure makes it impossible to quantify beat-to-beat variability but renders the analysis more robust to noise and consistent with the definitions employed in the Dawes and Redman criteria.

One aspect examined in the fetal signal is the variability of the FHR signal. Defined as the variation in the duration of successive intervals, it is a key indicator of the sympathetic and parasympathetic nervous system's regulatory activity on the fetal heart sinus node. Its adaptability and rate of change reflect fetal heart and autonomic nervous system health. This phenomenon can be assessed on different time scales, including short, medium, and long-term variability, depending on the availability of signal samples. Understanding these variations can provide crucial clinical insights into fetal adaptation and the presence of potential pathologies [29].

The variability in the short term was calculated for each minute of recording using two different parameters:

Short Term Variability (STV) defined as the mean of the absolute value of the difference between consecutive values of T[ms] expressed by the formula 2.1

$$STV = \frac{\sum_{i=1}^{23}[T(i+1) - T(i)]}{23}, \quad i = 1, ..., 23 \tag{2.1}$$

and the **Interval Index (II)**, calculated as the coefficient of variation (expressed by the standard deviation) of the differences between all FHR values within a one-minute recording averaged over 2.5-second periods, following the formula 2.2

$$II = \frac{\mathrm{std}[I(i+1) - T(i)] \cdot |i|}{STV}, \quad i = 1, ..., 24 \tag{2.2}$$

Medium-term variability is instead calculated using the parameter **DELTA** as the difference between the highest and lowest FHR values for each minute of recording according to the formula 2.3

$$\Delta = \max(T(i)) - \min(T(i)), \quad i = 1, ..., 24 \tag{2.3}$$

Additionally, the **number of accelerations and decelerations (n_acc and n_dec)** present in each signal were calculated over the entire signal itself. Accelerations are defined as events on the CTG trace during which the FHR, remains persistently above the estimated baseline value for a fixed period of time. Similarly are defined decelerations [34].

2.3.2. Frequency domain parameters

The presence of specific spectral components is strongly correlated with the activity of neural cardiovascular control systems, as established. Frequency domain parameters, extracted from power spectral density (PSD) analysis, are utilized to measure the activity of the sympathetic and parasympathetic branches of the autonomic nervous systems, responsible for regulating heart rate variability [37].
Spectral analysis was conducted through autoregressive (AR) modelling, with model parameters computed recursively employing the Levinson–Durbin algorithm [38].

Unlike the PSD of an adult subject, the fetal HR spectrum exhibits four distinct contributions, all of which have been considered in the analysis [29, 38]:

The **Very Low Frequency (VLF)** band, defined between 0 Hz and 0.03 Hz, related to long period and nonlinear contributions;

The **Low-Frequency (LF)** band, defined between 0.03 Hz and 0.15 Hz, mostly related with the activity of the sympathetic nervous system;

The **Movement-Frequency (MF)** band, defined between 0.15 Hz and 0.5 Hz, related to maternal breathing and fetal movements;

The **High-Frequency (HF)** band, defined between 0.5 Hz and 1 Hz, related to fetal breathing and parasympathetic activity;

The actual parameters VLF, LF, MF and HF derived from the aforementioned bands have been defined as the power in the bands calculated computing the integral of PSD and have been expressed in natural units (ms^2). These parameters have been calculated in windows of 3 minutes length.

2.3.3. Non-Linear parameters

Entropy paramaters

Approximate entropy (ApEn), a mathematical tool encompassing various statistical measures, is used to estimate a system's complexity. As reported and demonstrated by Pincus et al. in [39], its application in this research is based on the idea that its measurement can be a marker of fetal well-being, where a high value (indicating high complexity) suggests a healthy fetus and low complexity indicates possible pathologies. Unlike Kolmogorov's original aim, ApEn was made to analyze data without making assumptions about the system [40].
ApEn is so a way to understand how connected and predictable a system is. Low ApEn values usually mean the data is regular and predictable, while high values suggest randomness. In relationship with pregnancy and fetal analysis was originally used to predict acidosis in cases with very regular patterns [21].
Operatively, "ApEn(m,r,N) is approximately equal to the negative average natural logarithm of the conditional probability that two sequences that are similar for m points remain similar, that is, within a tolerance r, at the next point considering N the number of sampled of the input series" [39].
In the present analysis it was calculated according to the work of Pincus et al. [41].

ApEn shows dependency on data record length, often yielding lower values for shorter records, and lacks relative consistency across datasets. The use of **Sample Entropy (SampEn)**, which is a modified version of approximate entropy addressing the bias effect caused by self-matches, was considered for this reason. According to the definition "SampEn calculates the negative natural logarithm of the conditional probability that two sequences similar for m points remain similar at the next point, excluding self-matches

from the calculation" [42].
Similarly to ApEn, a lower value of SampEn suggests greater self-similarity in the time series. Another advantage of this metric is that the SampEn algorithm is simpler compared to the ApEn algorithm, requiring approximately half the time to calculate [42]. However, both metrics have been taken into account in the analysis with the aim of obtaining as much information as possible from the entire signal.

According to the definition proposed by Costa et al. **Multiscale entropy (MSE)** was also used [43]. This parameter quantifies signal complexity by analyzing entropy across different temporal scales, integrating information from Takens' theorem and stochastic approaches. It distinguishes between meaningful structural richness and randomness, addressing limitations of traditional entropy measures like approximate entropy and sample entropy. MSE considers variations in signal structure at multiple levels of temporal detail, offering insights into underlying dynamics. Unlike traditional measures, MSE captures complexity accurately even in systems with deterministic and stochastic components [43]. This parameter was already used for the analysis of fetal weel-being in [40] where was proven that MSE values increase for normal fetus.

Sample Asimmetry

Sample Asimmetry captures alterations in the distribution shape of HR intervals induced by decreased accelerations and/or transient declines in heart rate [44]. It highlights the asymmetry present in frequency histograms relative to their central tendencies, such as the mean or median. Unlike other metrics for assessing heart rate variability, sample asimmetry offers the ability to independently quantify the impact of accelerations and decelerations [44].
It is expressed by the equation 2.4

$$SampleAsymmetry = \frac{\sum_{i=1}^{n}(\widetilde{FHR}|\widetilde{FHR} > 0)^2}{\sum_{i=1}^{n}(\widetilde{FHR}|\widetilde{FHR} < 0)^2} \tag{2.4}$$

where $\widetilde{FHR}$ stands for the detrended version of the signal.

Lempel-Ziv parameters

The Lempel-Ziv Complexity (LZC), developed by Lempel and Ziv, is a measure of the algorithmic complexity of a binary sequence. This index derives from the LZ data compression algorithm, which replaces repetitions of subsequences with references to previously observed sequences. LZC thus represents the minimum amount of information needed

to describe the sequence, based on information theory [45–47]. There are two encoding options for the mentioned references to subsequences: a binary one, where references are encoded using a binary system, particularly efficient in terms of storage space, and one using a ternary system, more flexible, to represent the position of subsequences in the previous sequence [47]. The calculation of complexity is the same in both cases and measures the effectiveness of this compression process. The higher the complexity index, the more complex the sequence and the more difficult it is to compress [45].
For random sequences, the Lempel-Ziv complexity is equal to the length of the sequence itself since compressing it would result in information loss. Complexity depends on the number and frequency of different subsequences appearing in the sequence, reflecting the emergence of new patterns over time.

Depending on how the sequences have been encoded, two parameters of Lempel-Ziv complexity can be distinguished:
If the encoding is done in binary form as expressed by equation 2.5, then it is referred to as **Binary Lempel-Ziv complexity (LZC2)**

$$\text{Binary encoding} = \begin{cases} 1, & \text{if } RR[ms]_{n+1} > RR[ms]_n + p \cdot RR[ms]_n \\ 0, & \text{if } RR[ms]_{n+1} \leq RR[ms]_n + p \cdot RR[ms]_n \end{cases} \tag{2.5}$$

If the encoding is done in ternary form as expressed in 2.6, then it is referred to as **Ternary Lempel-Ziv complexity (LZC3)**.

$$\text{Ternary encoding} = \begin{cases} 1, & \text{if } RR[ms]_{n+1} > RR[ms]_n + p \cdot RR[ms]_n \\ 0, & \text{if } RR[ms]_{n+1} < RR[ms]_n - p \cdot RR[ms]_n \\ 2, & \text{if } RR[ms]_n - p \cdot RR[ms]_n \leq RR[ms]_{n+1} \leq RR[ms]_n + p \cdot RR[ms]_n \end{cases} \tag{2.6}$$

p represents a minimum quantization level for a symbol change in the coded string. The introduction of this parameter limits the effect of additive noise [47]. For the binary formula it is set to 0.02 while for the ternary case it is equal to 0.01.

Due to their definition, these Lempel-Ziv's complexity indexes, benefit from a longer signal duration. With a longer signal, in fact, the sequence dictionary can become richer and more varied, enabling better data compression. However, these algorithms are also highly sensitive to noise, which is certainly present in the signals under examination. To address these dynamics and maximize the benefits of these parameters, both indices were calculated globally and over 3-minute lentgh windows.

PRSA derived parameters

Phase-rectified signal averaging (PRSA) offers an effective method for analyzing quasi-periodic oscillations within noisy and non-stationary signals [48]. Despite challenges such as phase resetting and noise interference, PRSA enables the examination of system dynamics. The technique relies on identifying anchor points (APs), which are selected based on the average signal value before and after a specific instant within a chosen time window [49]. These APs serve as reference points for aligning oscillatory fluctuations, a process known as phase rectification. Following alignment, the signal's surroundings are averaged to smooth out variations [48]. The process just described can be seen in the Figure 2.3 where the various steps are illustrated.

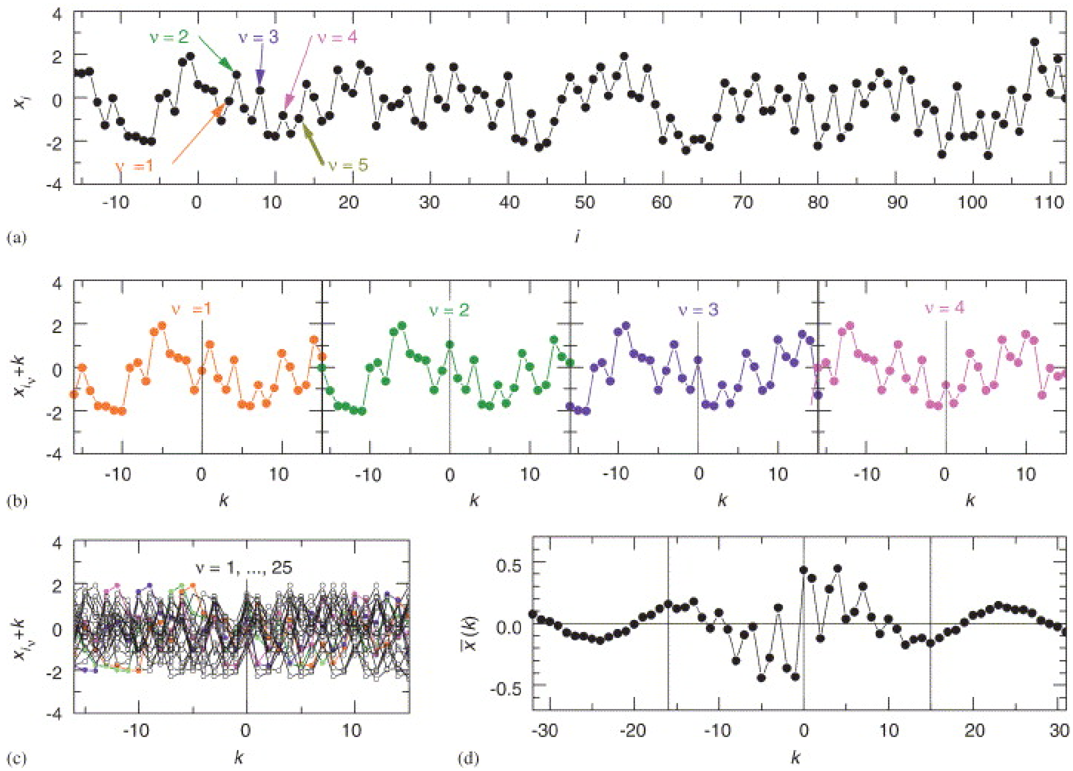

Figure 2.3: PRSA tecnique adapted from Bauer et al. [48]. (a) Anchor points are selected from the original time series. (b) Windows of defined length are defined around each anchor point. (c) The surroundings of many anchor points (all located in the centre) are shown on top of each other. (d) The PRSA curve resulting from averaging over all surroundings is shown versus the offset k from the anchor points

PRSA aims to evaluate autonomic regulation of heart rate even in the presence of phase de-synchronizations, missed beats, and signal losses [35].

The different choice of APs leads to the definition of deceleration-related PRSA (PRSAdec) and acceleration-related PRSA (PRSAacc) curves.
The deceleration-related PRSA curve (PRSAdec) is computed by defining the samples of the RR[ms] signal that meet a specific condition called deceleration anchor points (APdec) as follow:

$$APdec = \{t : \frac{1}{T}\sum_{i=0}^{T-1} RR[t+i] > \frac{1}{T}\sum_{i=1}^{T} RR[t-i]\} \tag{2.7}$$

Each deceleration anchor point (APdec) defines a window of length 2L (with L chosen to be set to 40 [50]), capturing the original signal values ranging from APdec-L to APdec+L-1. In this analysis, the parameter T is set to 1.
A comparable process is undertaken to derive PRSAacc by reversing the direction of the inequality.

For the aim of this work, from the PRSA analysis, various parameters have been derived: **Acceleration and Deceleration Capacities (AC and DC)** are the 1st determined using PRSA. These metrics have became relevant and very used due to their ability to effectively handle noisy signals and their sensitivity to various medical conditions, such as fetal distress [35]. Essentially, they measure the heart rate's ability to speed up or slow down, although they are not directly associated with specific activities in the nervous system [35]. Research indicates that AC and DC are valuable for monitoring FHR, aiding in distinguishing between healthy fetus and those experiencing for example growth issues [51]. They also demonstrate significant correlations with markers of oxygen deprivation [51].

The operative definition of this two parameters comes from the already mentioned PRSA curves.
DC is in fact defined as :

$$DC = \frac{1}{2S}\sum_{i=1}^{S} PRSAdec[L+i] - \frac{1}{2S}\sum_{i=0}^{S-1} PRSAdec[L-i] \tag{2.8}$$

where the parameter s is the scale and is equal to 1.
The definition of AC is similar but based on PRSAacc. (it's worth nothing that AC is a negative quantity) [35].

In relation to these two parameters, the **Deceleration Reserve (DR)** was then introduced. It's aim is to detect asymmetric patterns in the signal, encompassing both rising and falling trends frequently encountered in FHR during labor. It specifically highlights any disparities between deceleration and acceleration capabilities. DR serves as an indicator of whether the prevailing trend in the time series is predominantly upward (positive DR) or downward (negative DR) [35]. Its connection with AC and DC can be see by its definition:

$$DR = DC + AC \tag{2.9}$$

Other parameters considered are the **Acceleration and Deceleration phase-rectified slope (APRS and DPRS)**.

They have been computed as the slope of the PRSA curve computed in the APs according to the formula 2.10

$$APRS = \left.\frac{\partial X(i)}{\partial i}\right|_{i_{AP}} \tag{2.10}$$

This parameter is a quantitative indicator of both the average increase and decrease in FHR amplitude (absolute change of heart frequency so its acceleration and deceleration) and also of the duration of these events. These two measures are used to quantify fetal well-being [49].

The last parameters calculated from the signal have been proposed to quantify the oscillations in PRSAdec. First of all the scalogram of this parameter is computed as follow:

$$x_W^{PRSA}(s,p) = \sum_{k=-L}^{L-1} x_k^{PRSA} \cdot \frac{w(|k-p|)}{s} \tag{2.11}$$

The spectrogram is derived from the square of the wavelet coefficients generated by the analytic Morse wavelet, where the parameter y is set to 3, and the time-bandwidth product is 60 [52]. This spectrogram, denoted as PRSA_Spt, is computed at k=0, resulting in a single spectrum. The parameters **LFprsa, MFprsa, and HFprsa** are then calculated by integrating PRSA_Spt over the predetermined frequency bands and normalizing the result by the total power of PRSA_Spt.

2.4. Machine Learning

The subsequent sections will detail the steps taken and methodological decisions made that culminated in the development of two machine learning multiclass classification models. These models are adept at categorizing a record, and consequently a fetus, into one of the three previously outlined weight categories.
The mentioned models are a multi-class Logistic Regression and a Multilayer Perceptron (MLP), a linear and a non-linear model respectively.

The choice of the two models is driven by the dual objective of the work: to train a classifier capable of predicting the fetal weight category of a diabetic pregnant woman and to observe which features (among maternal and those derived from cCTG) have the greatest impact on this classification and their relationship. With interpretability in mind, Logistic Regression was chosen because, by definition, it provides insight into its functioning through the coefficients' values of the constructed function, which links features and target. Focusing on the predictive ability of the model, reliance was placed on the universal approximation theorem for neural networks formulated by Cybenko, which states that even with a single hidden layer, a multilayer perceptron can approximate any function that links input and output [53]. These models will be described in more detail in the section 2.4.2.

2.4.1. Data preparation

Removal of corrupted and outliers

Starting from the obtained database, an attempt was made to eliminate signals and their corresponding recordings that were too corrupted by noise, using the parameters calculated in the 1-minute and 3-minute length windows as reference. Rows were then identified and removed where more than 90% of the windows calculated values for the index under consideration were missing. After this operation, the database comprises 769 rows.

Subsequently, for all parameters, an analysis was conducted to remove any outliers (data points that significantly deviate from the majority of other values in the dataset). A Matlab function based on the percentile method was used, which removes data points falling outside the range defined by the 25th and 75th percentiles, known as the lower and upper quartiles. In this case, it was the individual parameter value calculated in that window that was removed, not the entire signal row. However, this was done only for those parameters that had at least 10 values (thus for which at least 10 valid windows

had been identified) to make the process meaningful.
Subsequently, the mean of the values was calculated for each parameter to arrive at a single value to be assigned to that signal, as explained at the beginning of the 2.3 chapter.

These were the last steps carried out in Matlab; all subsequent procedures were performed in Jupyter Notebook.

Visits Record Limitation

As a standard practice at the hospital from which the data was sourced, if other tests indicate or suggest various pathologies that could lead to complications in pregnancy, this is labeled as "at risk," and the expectant mother is usually closely monitored (including with cCTG exams) weeks before normal monitoring would begin. Consequently, these patients have a higher number of recordings than usual. This situation applies, for instance, to pregnant women with diabetes. In the available database, most pregnant women have multiple exams (the distribution can be observed in the Table 2.3), with extreme cases where, due to related pathologies, one patient's outcomes from up to 30 visits are available.

N. of subjects	N. of visits
19	1
19	2
28	3
16	4
15	5
8	6
9	7
6	8
7	9
5	10
1	11
4	12
3	14
2	15
1	16
1	19
1	21
1	30

Table 2.3: Distribution of the number of visits associated to the women in the database. For each number of registration is reported the numbers of $Pregnancy\_ID$ that have such number of registration related

To prevent bias in the classifier and avoid it learning too much from a limited number of patients, the number of recordings has been limited to a maximum of 10 per patient, selecting the most recent ones according to the sequential $Visit\_ID$ variable.

This was the final step where row elimination was performed. Therefore, the database appears in the subsequent machine learning phase with a size of 692 rows and 34 columns.

In the Figure 2.4 it is possible to observe the correlation matrix of the variables that will be used for training the models.

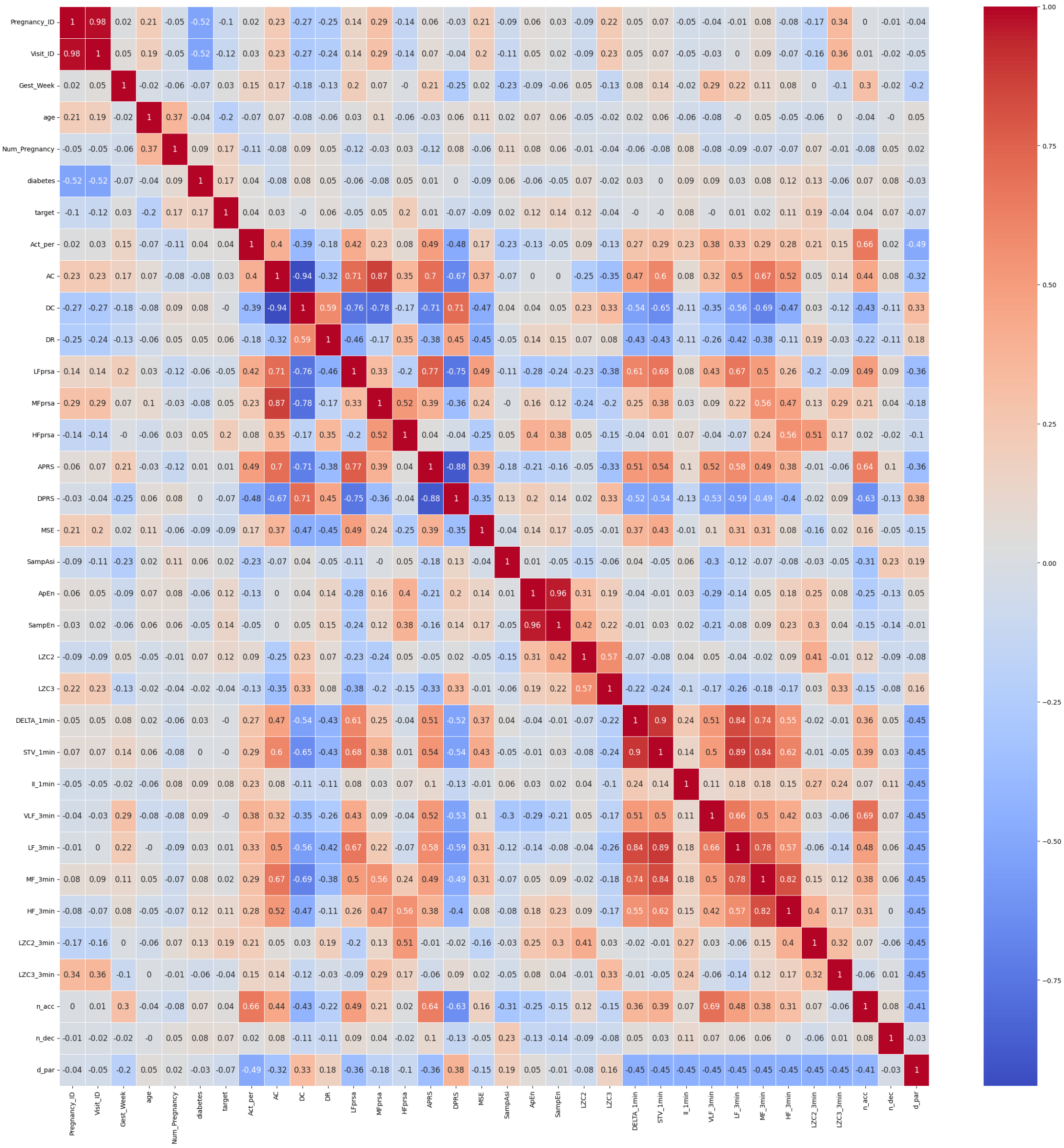

Figure 2.4: Correlation matrix of the variables

Handling categorical variable

For the subsequent stages of model construction and training, the categorical variable expressing the type of diabetes with a number from 1 to 3 is converted to "dummy"

variables. This process involves creating 3 (in this case) new binary variables named *diabetes_1*, *diabetes_2*, *diabetes_G*. A value of 1 is then assigned if the variable is present, and 0 otherwise.
This is done to represent the categories in a numerical format suitable for machine learning models.

Train and test split

For model training, the database was split into a training set and a test set with a ratio of 80% and 20%, respectively. The split was performed ensuring that different recordings from the same patient were not placed in the same set using a specific method that generates indices to create the training and test sets, allowing the data to be divided based on unique pregnancy identifiers. These instructions are used to define specific stratification in the sample domain maintaining the division proportions set.
This choice aims to avoid introducing bias in the model testing phase, where the model could otherwise encounter information from patients or pregnancies already seen and learned during training, compromising its robustness.

Missing values

Due to missing annotations during registration, there are 16 missing values for the pregnant woman's age and 3 missing values regarding the number of pregnancies already had by the woman under examination. To avoid losing information, it was chosen not to delete the corresponding rows but to impute the missing values with estimated values. The approach is based on the K-Nearest Neighbors algorithm, which for each sample with a missing value, finds its defined number of nearest neighbors in the dataset based on the available values and uses them to estimate the value to replace by averaging them. The application of this method differs for the training and test sets:

In the first case, the process combines the identification of the nearest neighbors, the calculation of estimates for missing values, and the replacement of missing values directly in the training set. This method allows training the model and performing the transformation simultaneously, improving computational efficiency and ensuring that the model is trained with complete data.

In the second case, the information learned during training is used to calculate estimates for missing values in the test set without another calculation. This ensures that the model is not influenced or modified by the test data during the process of imputing missing values.

Standardization

All numerical variables were then standardized to ensure a common scale of representation. The chosen operator is the $MinMaxScaler$, which rescales features to a range between 0 and 1 by subtracting the minimum value and then dividing by the difference between the maximum and minimum values of each variable. This pushes the variable values into the predefined range. Similar to the handling of missing values, the operation differs for the train set and test set in this case as well. For the standardization of the test set, the minimum and maximum values previously calculated from the training set are used to ensure that the data transformation is consistent and to avoid any distortion that could negatively impact the model's performance.

Standardization is applied to all numerical variables except for those cases where parameter values were set to -1, that, by the way, are still indicated by the binary variable d_par. This was done to ensure again that the information carried by these parameters wouldn't be lost or joined to the information of other parameters.

Balancing of the data

As previously stated, the target variable is not evenly represented in the samples comprising the database, with the NGA category accounting for 50% of the rows, the SGA class for 29%, and the LGA class for 21%. This data imbalance is also reflected in both the train and test sets, even after the explained registration elimination phases 2.4.1.

By leaving the class distribution in the train set this way, there's a risk of creating bias in the model towards the most represented class, which would be predicted more likely at the expense of the less represented ones. This would undermine the analysis objective, especially considering the clinical significance of the work, where it's more important to predict the "problematic" classes than the healthy ones. The cost of misclassifying an abnormal and interesting data point as a normal example is much higher in this case than the cost of the reverse error.
Additionally, balancing the train set allows the model to generalize properly to new data, reducing the risk of overfitting. It's important to note that this balancing is not applied to the test set, which remains composed of unseen data that should represent the distribution in reality.

The classes in the train set were balanced using a combination of two techniques: undersampling and oversampling with the SMOTE algorithm [54].

Undersampling involved randomly sampling a subgroup of the majority class (NGA). For training the Logistic regressor classifier, a number of samples equal to that of the second most represented class, SGA, was selected. Meanwhile, a subgroup equivalent to 73% of the original composition of NGA (but still higher than the number of samples of the other 2 classes) was reached when training the multi-layer perceptron.

In addition, an oversampling technique called **Synthetic Minority Over-sampling Technique or SMOTE** was employed [54]. This algorithm increases the number of samples of a minority class by generating synthetic data, rather than simply oversampling with replacement. Specifically, SMOTE selects examples that are close in the feature space, drawing a line between these examples and creating a new sample along that line by calculating the difference between the selected examples, multiplying this difference by a random number between 0 and 1, and adding it to the reference sample. This process generates points along a line between two existing examples, thereby expanding the decision region of the minority class [54].
By default, this algorithm generates samples until reaching the number of samples in the most represented class. As mentioned earlier, this number will depend on the model trained with such data. In the case of the logistic regressor, SMOTE is practically applied only to the LGA class until reaching a number of samples equal to SGA (to which NGA had already been brought). Meanwhile, in the case of the MLP, both the LGA and SGA classes undergo oversampling, resulting in 200 samples per class. As will be explained later in section 2.4.2, SMOTE will be applied only to the training set in k-fold cross-validation to avoid introducing data leakage.

The different amount of undersampling and oversamplig between the 2 models primarily depends on their nature. The Logistic model prefers a smaller amount of data but of higher quality for its operation. This is a linear and simple model, less prone to overfitting and that can be effectively trained even with a limited amount of data, provided that they accurately represent the distribution of real data. Although the SMOTE algorithm is efficient in fact, the synthetic nature of its data is still recognizable to the classification algorithm. For training the MLP, a nonlinear model, on the other hand, as many data as possible are required to limit the risk of overfitting since it's characterized by a big number of parameters.

The Table 2.4 displays the actual numbers of training and test sets after the undersampling operation for the two models:

Model	Train set	Test set
Logistic Regression	418	167
Multilayer Perceptron	442	177

Table 2.4: Train test and test set dimensions before the training phase

2.4.2. Models training

The steps listed so far enable the actual training phase of the two models. The objective of these models is to classify a record, and therefore a fetus, into one of the 3 weight categories (SGA, NGA, and LGA) using information derived from the maternal clinical history and parameters derived from the FHR signal.
For the actual training of the model, all variables that were definitely not significant ($Pregnancy\_ID$ and $Visit\_ID$ and those related to postpartum information ($delivery\_week$) as said were excluded. It is worth noting in fact that the focus is on the information about fetal health obtainable by the clinician during prenatal examinations.

As mentioned, there are two machine learning models that have been trained for fetal weight classification into one of the 3 categories: a multiclass Logistic Regression and a Multi-Layer Perceptron.

Logistic Regression is a classification algorithm designed primarily for binary classification problems, i.e., problems with two classes or values for the target variable, but it can be extended to handle multiclass classification problems. In the latter case, it is referred to as **Multinomial Logistic Regression**. This extension mainly involves adapting the softmax function to predict a multinomial probability distribution, assigning probabilities to each output class. The softmax function plays a crucial role in transforming class scores into probabilities. During the training process, the loss function used is cross-entropy (based on information theory and entropy), which measures the discrepancy between the probability distribution predicted by the model and the actual probability distribution of the training data. Cross-entropy is therefore crucial in computing the loss during model training, as it provides a measure of the model's prediction error.
Its main hyper-parameters are:

- The inverse of the regularization strength, C;
- The maximum number of iterations;
- The solver (algorithm to use in the optimization problem);
- The penalty

The **Multilayer Perceptron (MLP)** is a type of artificial neural network widely used in the field of artificial intelligence. Its name comes from its layered composition, where there is an initial part that receives the input, followed by a variable number of hidden layers containing neural units that perform nonlinear transformations on the input data, and finally the layer that produces the network's output. The connections between these units have weights that are updated during training to reduce error. Of particular importance is the application of a non linear activation function by the neurons in the hidden layers, which introduces nonlinearity into the model. This allows the MLP to learn complex relationships in the data.
Its main parameters are:

- The number of hidden layers;
- The activation function used;
- The solver for weight optimization;
- The alpha parameter
- The number of iterations to perform

For both models, the training and test process occurs in the same way:

1. A pipeline was built to concatenate the application of the SMOTE algorithm with model classifier training; This is done to ensure that the oversampling is done only in the actual train sets;
2. The technique of stratified k-fold cross-validation is used to divide the train set into k subsets (folds) while maintaining the class proportions unchanged in each fold. This helps to robustly evaluate the model's performance on different parts of the the set, reducing the risk of overfitting and providing a more accurate estimate of performance. Introducing SMOTE into the pipeline ensures that the generation of synthetic data occurs only on the training data and not on the test part in each fold of the cross-validation, thus avoiding data leakage;
3. The grid search technique is used to search for optimal hyperparameters for the logistic regression model and the MLP one. Grid search trains the model with all possible combinations of these hyperparameters, evaluating the model's performance on each combination using the defined cross-validation. At the end of the grid search, it returns the model with the best combination of hyperparameters found;
4. For performance evaluation, the cross-validation score is calculated as the best average performance obtained during cross-validation using balanced accuracy as the

metric. The choice of balanced accuracy is due to its ability to calculate the average accuracy across classes, penalizing models less for correctly predicting minority classes compared to simple accuracy, ensuring that each class contributes fairly to the overall score. This is particularly useful in the case of class imbalance and allows for a more accurate assessment of model performance. It's formulation is the follow (note that the library used calculates Balanced Accuracy in the multiclass case as the average of the recalls obtained for each class):

$$\text{Balanced Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i} \tag{2.12}$$

where N is the number of classes, TPi is the number of true positives (i.e., instances of class i correctly classified as belonging to class i) and FNi is the number of false negatives (i.e., instances of class i erroneously classified as not belonging to class i). In addition to balanced accuracy, other two metrics were used to show at the end the models' performance: accuracy and F1 score. The accuracy value indicates the percentage of correct predictions made by a model out of the total number of predictions made. The latter is a metric that combines precision and recall, considering both false positives and false negatives. It is particularly useful when classes are imbalanced. It evaluates how well a model balances precision and recall among the different classes in a multiclass classification.

5. Finally, the score on the test set is calculated using the model with the best combination of hyperparameters found during the grid search;

2.5. Explainable Artificial Intelligence

In the overall aim of the presented work, the intention to provide a tool and knowledge that can be used by medical professionals should not be overlooked. In this regard, the tool of Explainable Artificial Intelligence (XAI) comes into play, which refers to the ability of artificial intelligence systems to explain the internal functioning of a system, algorithm, or model in a manner compatible with human thought, to understand the reasoning behind the choices made by the AI that led to a particular decision [55]. It is clear from its definition how this field is growing exponentially when applied to the healthcare domain, particularly in Clinical Decision Support Systems (CDSSs) [55].
Challenges in the healthcare domain that XAI aims to address include associating the knowledge of a trained model with medical characteristics or understanding how the presence or absence of certain medical feature information affects the performance of a model and its interpretation of features [56].

XAI is therefore a crucial tool to enable healthcare providers to effectively understand and assess artificial intelligence (AI)-based solutions. However, it also serves as a powerful tool for engineers themselves, who have numerous reasons to seek to understand what lies behind the black box of machine learning models. Although machine learning involves managing the variables at play, this does not automatically mean that the resulting model is entirely transparent. Machine learning models can be complex, and the relationships between variables can be nonlinear. Moreover, explain ability can aid in model validation and debugging in case of errors or unexpected results. Understanding how the model arrives at a particular prediction can help developers identify and correct issues in the model or input data, as well as understand if it is influenced by biases in the training data or algorithmic choices.

These issues are evident in nonlinear models like MLP but are actually present in linear models as well, such as Logistic Regression. Although the coefficients of logistic regression provide information about the relationships between input variables and output, they do not necessarily offer a complete explanation of the model's decision-making process, and not all interactions between variables or effects of correlated variables influencing the model's predictions may be revealed.

Multiple XAI tools/algorithms have been provided and can be found in the literature.
In this work, **SHapley Additive exPlanations (SHAP)** has been employed. This method was used in this work to investigate the possible impact of features on trained models, seeking to understand the difference in importance between variables derived from maternal clinical history and parameters extracted from the FHR signal, as well as their

relationship.

2.5.1. SHAP

SHAP (SHapley Additive exPlanations), introduced by Lundberg and Lee in 2017, is an explainability method that represents a significant advancement in interpreting the predictions of machine learning models. This method, based on coalition game theory and Shapley values, provides a robust conceptual framework for explaining model decisions clearly and accurately [57, 58].

The main goal of SHAP is to explain the reasoning behind model predictions on an individual basis, identifying the specific contribution of each feature in the decision-making process. In other words, it aims to provide a clear interpretation of how the model reasoned to arrive at a specific prediction for a data instance. This involves analyzing the specific contribution of each feature or input variable in the model's decision-making process. At the same time, it also allows for understanding how these features interact with each other to determine the final prediction [57].
To do this, SHAP uses Shapley values, which come from coalition game theory. These values assign a fair "payout" or weight to each feature, representing and based on how much each variable contributes to the final outcome predicted by the model. This provides a detailed understanding of how each input influences the final result. SHAP specifies the explanation as:

$$g(z_j) = \phi_0 + \sum_{j=1}^{M} \phi_j z_j^k \tag{2.13}$$

where g is the explanation model, $z' \in \{0,1\}^M$ is the coalition vector, M is the maximum coalition size and $\phi_j \in \mathbb{R}$ is the feature attribution for a feature j that is the values of Shapley [57]. A coalition vector in SHAP is a binary vector representing the presence or absence of each feature in the context of model interpretation. It indicates which combinations of features are considered together to assess the contribution of each to the model's predicted outcome [57].

Three are the key properties of this model: Local Accuracy, ensuring precise explanations on an individual basis; Missingness, effectively handling missing features for proper interpretation; and Consistency, ensuring that explanations align with changes in the model. These properties ensure an accurate understanding of how each feature influences predictions, even in complex scenarios [57].

A distinctive aspect of SHAP is its representation of explanations as a linear model of additive feature attribution. This model allows for a clear visualization of the weight of each feature in the model's prediction, providing an intuitive and interpretable explanation. This is indeed the key point of its usage, which also represents the reason for its choice in this field: its ability to represent explanations in graphical form. This means that Shapley attributions, indicating the contribution of each feature to the model's prediction, can be visualized through intuitive and interpretable graphs. These graphical representations allow for a more immediate understanding of the influences of individual features on model predictions, facilitating the analysis and interpretation of results [57].

Another strength is its ability to provide both global and local explanations.
The global explanations provided by SHAP enable understanding the overall behavior of the model across the entire dataset. This can be particularly useful for identifying general trends, relationships between features, and their importance in the model's decision-making process [57].
For this purpose, the **summary plots** graph have been used in this work.
This graph, an example of which can be seen in the Figure 2.5, combines the feature importance with their effects. Each point represents a Shapley value for a specific instance and feature at a time. The vertical axis orders the features from top to bottom by their mean absolute SHAP values for the entire dataset, so based on their importance, while the horizontal axis expresses the SHAP value for each feature. Blue color indicates a low value of the variable under consideration, while red color indicates a high value.

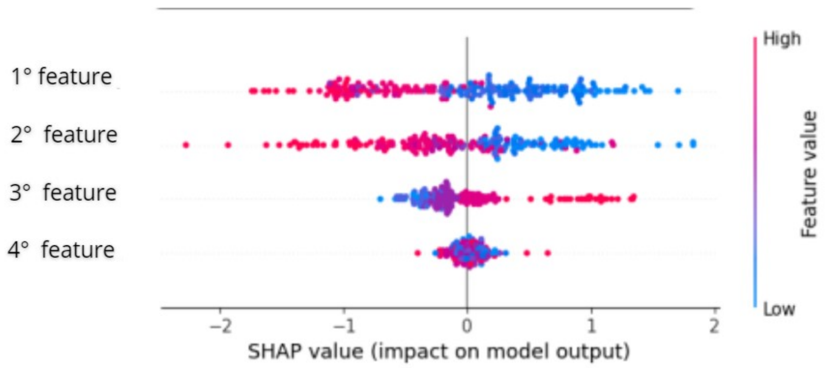

Figure 2.5: Example of SHAP summary plot. Adapted from [57].

On the other hand, SHAP's local explanations also focus on individual predictions, providing detailed insight into how the model reasoned about a specific data instance. This is valuable for understanding why a particular prediction was made by examining the contribution of each feature in that specific context.
To analyze how features influence individual predictions, **waterfall plots** have been used.

As can be seen in Figure 2.6, they visually represents Shapley values ϕ_i, j as arrows indicating whether they increase or decrease the prediction $f(xi)$ compared to the expected prediction $E[f(x)]$.

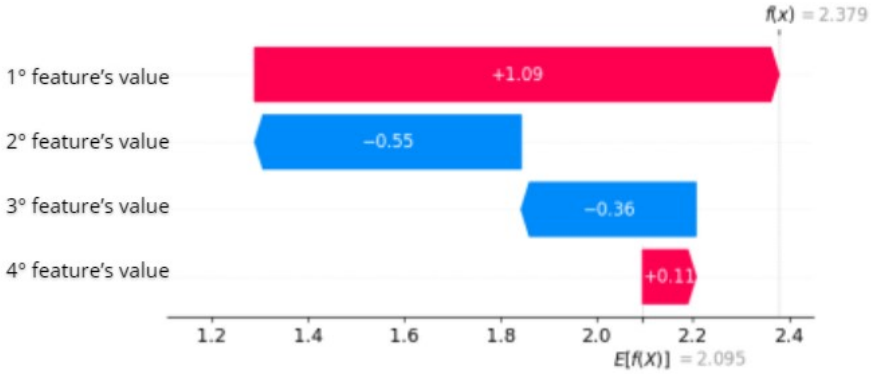

Figure 2.6: Example of SHAP waterfall plot. Adapted from [57].

Finally, interactions among the most interesting features were also investigated, and for this purpose, an **interaction plots** have been used. This kind of plot shows how the SHAP contributions of different model attributes vary based on the values assumed by other variables. Through this type of visualization, it's possible to better understand how the relationships between variables influence the model's predictions, allowing for a better interpretation and understanding of its behavior. An example can be seen in Figure 2.7

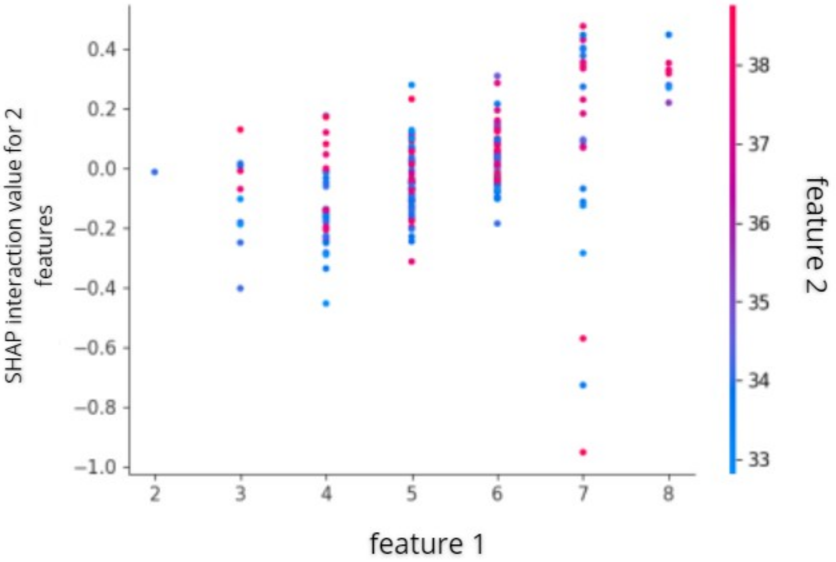

Figure 2.7: Example of SHAP interaction plot. Adapted from [57].

At the implementation level, SHAP offers several variants, including KernelSHAP and TreeSHAP, which allow for efficient and accurate explanation of model predictions. KernelSHAP uses a kernel-based approach to estimate feature contributions, while TreeSHAP is specifically designed for tree-based models such as decision trees and random forests. In this work, the former type was utilized.

Specifically an Explainer object, that is an object that uses machine learning models to explain the contribution of each feature in the model's decision-making process, is created, configured to use the previously trained model to make probability predictions on the training set. Finally, Shapley values are computed for the test set, and the contributions to predictions are visualized.

In the Result chapter (3), the results of applying SHAP to the present study will be presented.

3 | Results

The purpose of this research is to train classifiers capable of categorizing pregnancies into weight categories such as SGA, NGA, and LGA. Another fundamental objective is to explore, among the available maternal characteristics and parameters calculated over the FHR detected by the cCTG system, which ones are most relevant to achieve this goal with the additional aim to understand the existing connections between these elements.

Following the training of the two models, the results achieved in balanced accuracy, F1 score and accuracy for the classification task of the three weight classes, are presented in Table 3.1.

	Logistic Regression		**MLP**	
Metrics	Train	Test	Train	Test
Balanced accuracy	55%	54.7%	59.8%	52.6%
F1 score	53.6%	51.5%	58.4%	50%
Accuracy	54.6%	50.2%	60%	49.7%

Table 3.1: Train and test results considering various metrics of the 3 classes classification

From the Table 3.1, it is evident that the model with the best performance is Logistic Regression, which achieves higher results in testing across all metrics used compared to MLP. In particular should be noticed its value in test for the balanced accuracy, the highest between the test results.

For visualizing the performance of the models on the test sets, the confusion matrix was also used, in its normalized version.
In the Figure 3.1 can be seen the confusion matrix of the Logistic Regression model

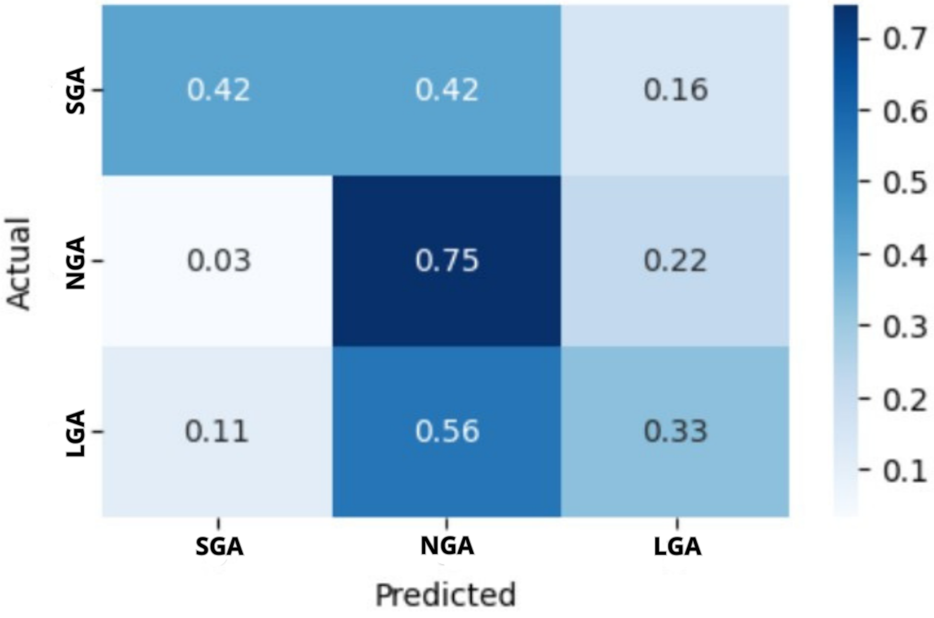

Figure 3.1: Confusion matrix of Logistic Regression

In the Figure 3.2 can be seen the normalized confusion matrix of the MLP model instead.

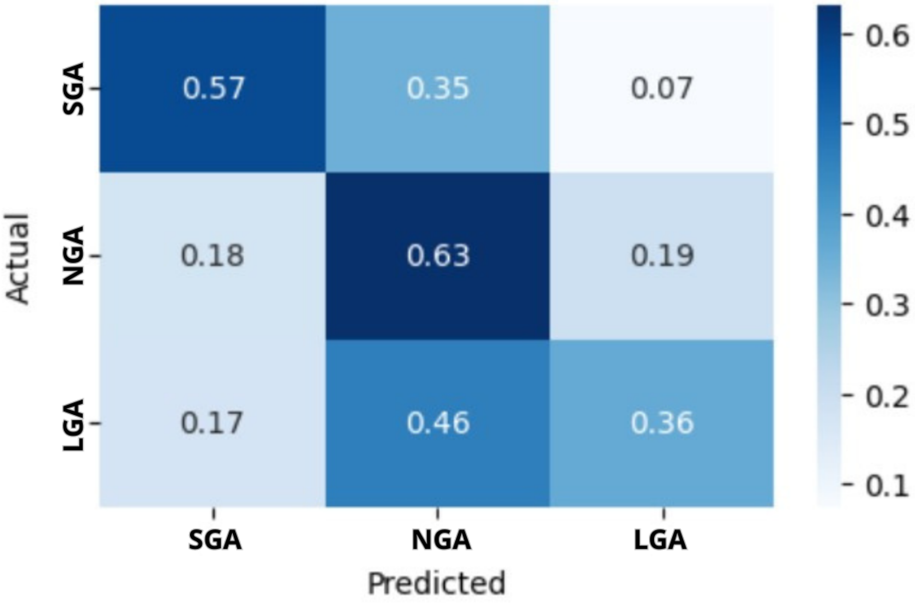

Figure 3.2: Confusion matrix of Multilayer Perceptron

For both models, it can be observed from these matrix that the most prevalent class in the database, namely the NGA class, also achieves the highest level of balanced accuracy.

Can be also seen that both models' primary source of error lies in the misclassification of large fetuses as normal. This issue of false negatives should be further examined and addressed because in medicine, the presence of this type of classification errors can lead to overlooking relevant clinical cases.

3.1. Majority voting

As previously stated, multiple records can be associated with the same patient, up to a maximum of 10. In the original database, at the beginning of the work, all records for the same patient had the same target (the calculation assigned the same category to all records due to the fixed parameters used that are the week of delivery and birth weight). Therefore, it was desired to verify whether the trained models would lead to the same outcome.
Moreover, it was also intended to understand if a large number of recordings could favor the assignment of the correct class.

The targets assigned to the various samples in the test were so compared, relying on the unique numerical code identifying pregnancies, with the original ones that had been assigned. This comparison revealed that recordings from the same pregnant woman were not always placed in the same category. It was decided so to investigate how the balanced accuracy would vary if the two classification algorithms were forced to assign the same label/target to all records for the same patient. Essentially, by grouping all records for the same patient relying on the unique code, the target that had been most frequently assigned, after the training of the model and its application on the test set, was reassigned to all records of that group.

The balanced accuracy of the two models was then recalculated. The results show an increase up to 55.1% of balanced accuray in the case of Logistic Regression, while a substantial increase up to 59.6% was observed in the case of the MLP. Performing multiple recordings has thus proven to be very effective for better target assignment and consequently for improved diagnosis, especially in the case of the MLP model.

3.2. Understanding the model: XAI application

As already mentioned, another focus of the work is on the interpretability of the obtained results. The interpretability of results in artificial intelligence is crucial, especially in sensitive sectors like healthcare. Transparency and the ability to interpret a model allow clinicians to understand the algorithm's decision-making process. This aspect plays a key role in building trust: medical professionals tend to have greater trust in a model when they understand its motivations and this also helps clinicians make informed clinical decisions. They can assess the adequacy of the model's recommendations in relation to other clinical information available to them and, if deemed necessary, may choose not to adhere to the predictions provided by the model.

Another aim is to understand which variables are most capable of defining, within a high-risk category such as diabetic pregnancies, which pregnancies are at higher risk for complications related to the condition (it's important to remember, in fact, that the classification performed is not in relation to a healthy control group).

In the next section, relevant results will be primarily presented through the SHAP plots showcased in Section 2.5.1.
In all the graphs, the parameters' names will be presented with the abbreviation mentioned in the Table 2.2. The acronym $3min$ refers to parameters calculated in windows of three minutes length, while $1min$ refers to one minute windows.

3.2.1. Summary plots and Model Coefficients

SGA plots

In the following plot 3.3, can observed the influence of various parameters on the prediction of the SGA class by the Logistic Regression model. This is done both through visualization of the coefficients of the function that links dependent variables and target, and through the use of the SHAP summary plot. SHAP is in fact a powerful interpretation tool and in this case this is particularly useful, considering that in a multiclass classification, a simple line linking input and output is not created as in the case of a binary classification.

In general, a coefficient plot shows how a standard deviation change in the variable affects the model's prediction. Variables with larger standard deviations correspond to longer bars in the plot, indicating greater importance of the variable. The contribution of the variable can be positive (bar in the right half-plane) or negative (bar in the left half-plane) to the prediction.
On the other hand, a SHAP summary plot displays in descending order the 20 variables that have the greatest impact on the model ranking them by their mean absolute SHAP values for the entire dataset, referring to instances of a specific class for which the distributions are shown for each variable. The more points on the right half-plane, the greater the impact on the prediction of that class by the values of that variable (low values if blue points, high values if red points).

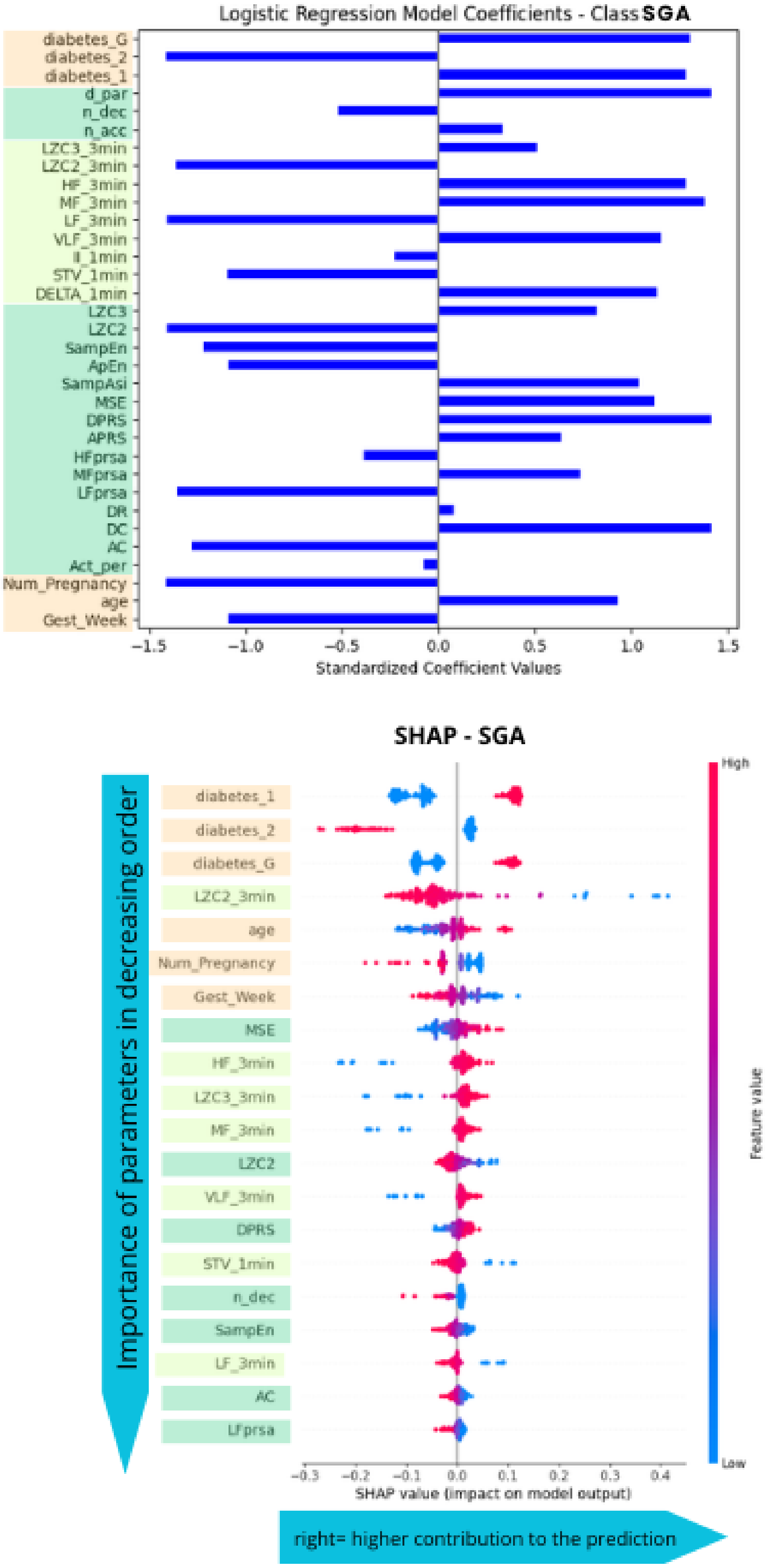

Figure 3.3: Illustration of the model coefficients and the SHAP summary plot for the SGA class of the Logistic Regression model.

As can be observed, both graphs agree on the positive or negative impacts of the variables, although they do not always agree on the order of importance of the variables. In particular, it is noted that:

- Both types of graphs show how the Logistic Regression model places great importance on **diabetes types** for predicting the SGA class, as indicated by the length of the bars in the coefficient plot and by the position on the y-axis in the summary plot. Both the presence of **type 1 diabetes** and **gestational diabetes** increase the probability of predicting an SGA fetus, as does the absence of **type 2 diabetes**.
- The **coefficient of Lempel-Ziv binary complexity** measured over a 3-minute window is also shown to be highly impactful. A high value in this parameter, so an higher complexity of the signal, indicates a lower likelihood of an SGA fetus compared to other weight classes. Conversely, the contribution of the **ternary complexity** parameter, slightly less impactful than other variables, also calculated over 3 minutes, positively influences the prediction of the target class. Another relevant factor, highlighted in the coefficient plot but less prominent in the SHAP summary plot, is the binary complexity coefficient calculated over the entire signal, behaving similarly to its windowed counterpart.
- **High power values in the high, movement**, and **very low frequency bands** increase the likelihood of predicting an SGA fetus. So the parasympathetic, maternal breathing and fetal movements and also non linear contribution that have been associated to the aforementioned parameters, led to this prediction. Conversely, the prediction likelihood of this class decreases with power in the **low frequency band**, although it should be noted that the contribution of this variable is high in the coefficient plot but marginal in the SHAP summary plot.
- Regarding maternal features, significant relevance is attributed the **number of pregnancies**. This variable tends to decrease the likelihood of an SGA classification as its values increase. Conversely, **maternal age** contributes positively to the prediction, with higher values associated with a greater probability of SGA compared to other weight classes.
- The negative contribution of **Sample Entropy** is notable in the coefficient plot but less emphasized in SHAP. However, both agree on the importance of **Multiscale Entropy**, which increases the likelihood of predicting an SGA fetus.
- The **variability in the short term** of the signal seems to not be associated with the probability of being SGA.

- High values of the variable **DPRS** contribute to the logistic regression model's prediction of the SGA class compared to other weight classes.
- **Acceleration capacity**, on the other hand, decreases the probability of predicting the target class. It's worth noting that this is not in accordance with what already found in [22] for IUGR fetus.
- A difference between the 2 types of plots is that the coefficient plot emphasizes the positive contribution of the dummy variable *d_par*, which indicates parameters in the absence of activity segments, the negative contribution of **approximate entropy**, and the contributions of the **Medium-term variability (DELTA)** variable. This is not reflected in the SHAP plot, which does not rank these variables among the top 20 most impactful.

From this plot can be seen, in general, that measures related to the parasympathetic system (such as HF and DPRS) exhibit opposite behavior compared to measures of the sympathetic system (identified by parameters LF, AC, LFprsa, and STV) and push for a positive prediction of the class in question. Generally, this may suggest a greater contribution of the parasympathetic system in SGA fetuses.

Now the SHAP summary plot still related to the SGA class of the MLP is shown in Figure 3.4.

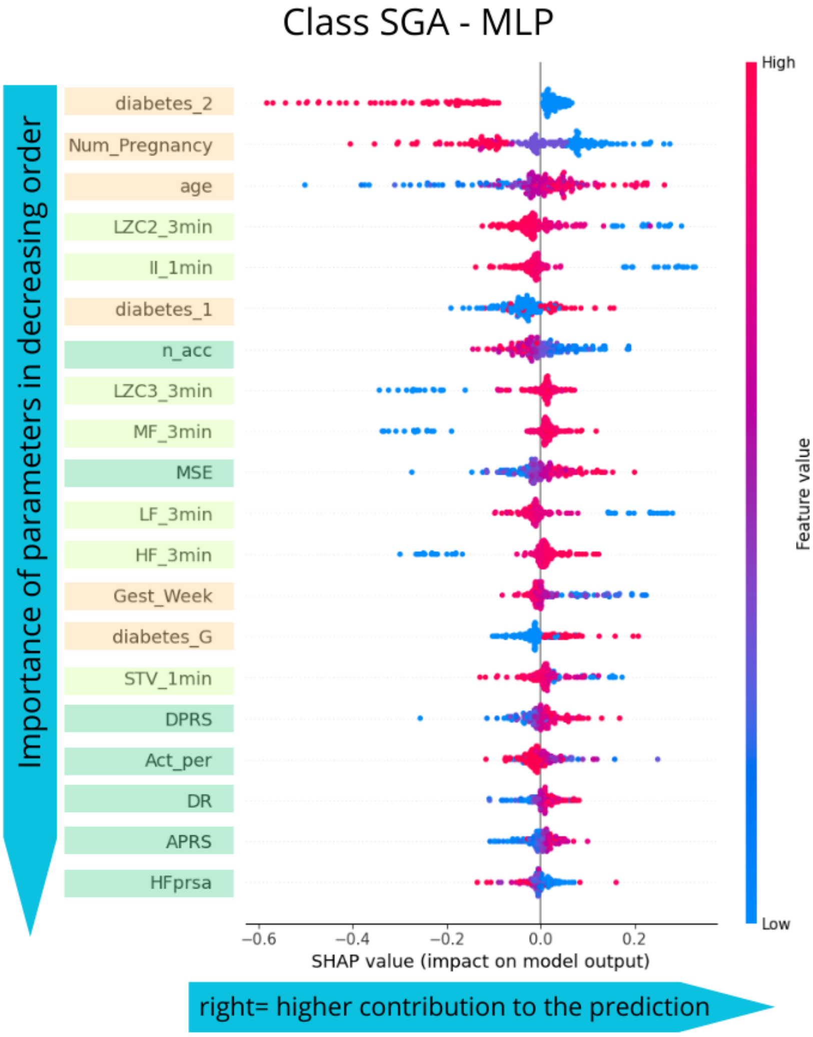

Figure 3.4: Illustration of the SHAP summary plot for the SGA class of the MLP model.

The first thing that needs to be observed is the consistency regarding the direction of impact on the model prediction by the variables that are considered important by both the MLP and the Logistic Regression (particularly looking at the top 20 choices by SHAP for the latter model).

Divergences are observed regarding the parameter **II_1min**, which expresses short-term variability calculated in 1-minute windows, as it was not considered relevant by the Logistic Regression. It is not among the top 20 variables in the summary plot of the Logistic Regression and provides a small negative contribution to the prediction according to the coefficient plot. However, the MLP model considers its negative contribution to the prediction of a fetus being SGA as highly relevant, saying that the more a signal is variable, the less the class SGA will be probable, and place it among the top 5 variables by impact.

A similar consideration applies to the variable **APRS**, whose positive contribution is not relevant for the summary plot of the logistic regression and is secondary for the coefficient plot. However, for the MLP model, high values of this variable clearly contribute to a fetus being predicted as SGA compared to other weight classes.

Furthermore, for the MLP model the **number of accelerations** present in an FHR signal are very important to decrease the probability of predicting the SGA class, whereas in the Logistic Regression, their contribution is secondary compared to the one of the number of decelerations.

NGA plots

The same scheme will now be used to present the results for the NGA class, starting from the illustration in the following coefficient plot of logistic regression and the corresponding summary plot (3.5).

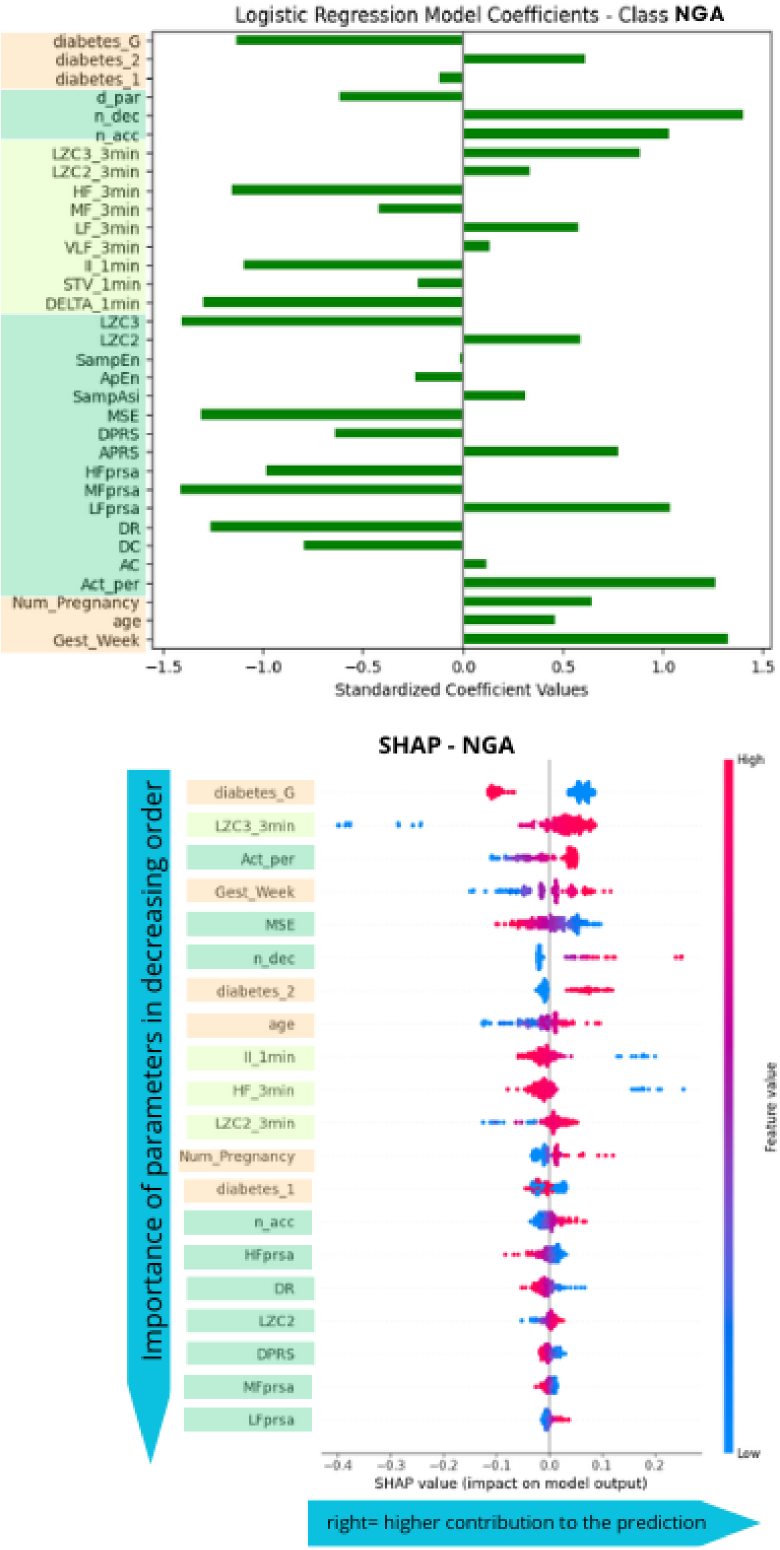

Figure 3.5: Illustration of the model coefficients and the SHAP summary plot for the NGA class of the Logistic Regression model.

Again, also in this case both graphs agree on the positive or negative impacts of the variables, although they do not always agree on the order of importance of the variables.

- Among the **types of diabetes** in this case, the absence of **gestational diabetes** is what drives the prediction of the fetus as NGA. Next in importance is the contribution of **type 2 diabetes**. The presence of this type of diabetes consistently favors the prediction of the class under consideration compared to other weight categories. The coefficient plot shows how marginal the contribution of **type 1 diabetes** is. In the summary plot, there is no clear influence observed, as there are both red and blue points on the same plane, indicating that the absence and presence of this diabetes type decrease the probability of predicting the NGA class.

- Both representations demonstrate the significance of the **number of decelerations** in the FHR signal for predicting the NGA class compared to other weight categories. Similarly, the **number of accelerations** is considered important with the same direction of contribution in the SHAP plot and in the coefficient plot.

- The **ternary Lempel-Ziv complexity index** calculated in 3-minute windows is also highly relevant, contributing to the prediction of the class under consideration for high values so for complex signals. This is particularly evident from the coefficient plot, while it is slightly less intuitive from the SHAP plot, where it is clear that low values are against the prediction of the class, but there is a little more ambiguity regarding high values (red points) that are present also in the negative half-plane. The coefficient plot shows the opposite contribution of the same parameter calculated globally across the entire signal, whereas the summary plot does not consider this variable to be sufficiently relevant.

- Both representations show that a higher presence of **activity segments** in the FHR signal increases the probability of predicting the fetus as NGA rather than SGA or LGA.

- A higher **number of pregnancies** also contributes to predicting the NGA class.

- The **short-term variability parameters** II_1min **and** $DELTA\_1min$ reduce the probability of assigning the NGA class, although the latter is not actually considered as relevant as indicated by the coefficient plot. Lower variability of the signal seems to be associated with the class NGA.

- **Multiscale entropy** is a parameter for which low values so more regular signal increase the probability of predicting the class under consideration.

- Great importance is then given to the **power in the high frequencies of** a signal processed with **PRSA** technique, indicating a higher probability that the fetus is normal compared to other weight classes if this value is low.

- Among the variables of power in the bands, the values of **power in the high frequencies**, associated with the parasympathetic system, is relevant in reducing the probability that the fetus is predicted as NGA.
- Lastly, the contribution of **deceleration reserve**, although negative, is relevant.

Now the SHAP summary plot still related to the NGA class of the MLP is shown in Figure 3.6.

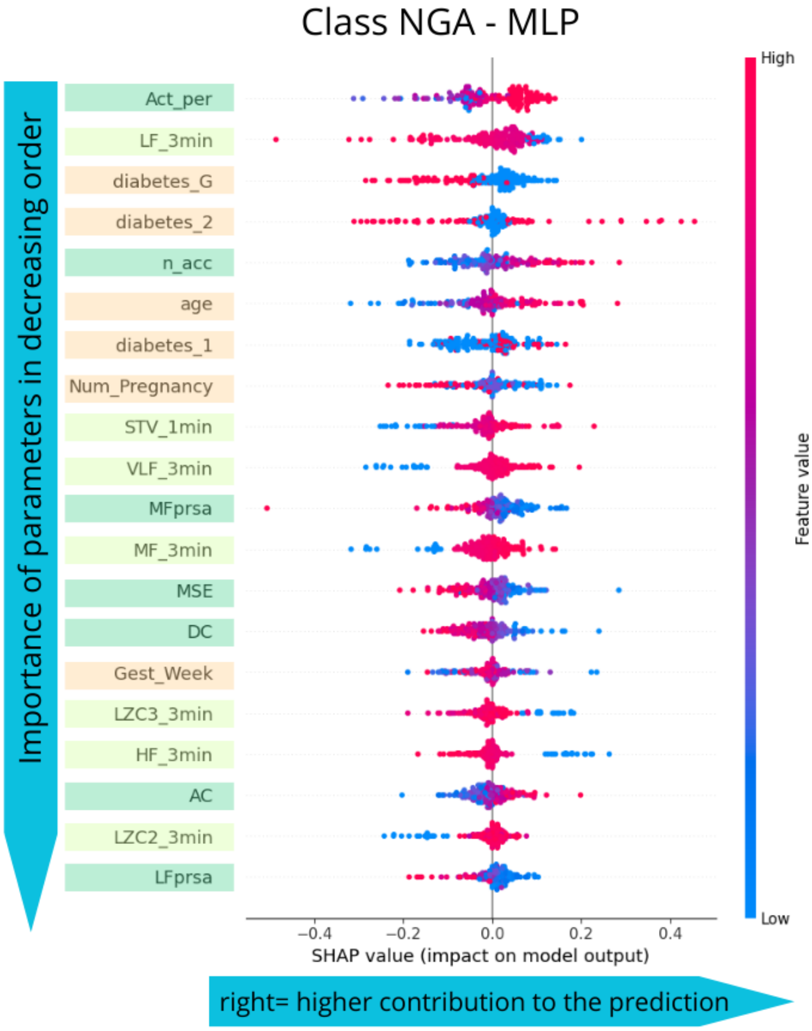

Figure 3.6: Illustration of the SHAP summary plot for the NGA class of the MLP model.

In this case, it should be noted that there is a considerable difference between the variables considered important for the MLP model and for the Logistic Regression.

Another characteristic of this plot is the less clear separation into distinct positive and negative semi-planes of points representing high and low parameter values.

Consistent with the Logistic Regression model's graphs, the **percentage of activity segments** on the FHR signal and **gestational diabetes** behave similarly, both of which are of great importance for this model as well.

Totally irrelevant for this model is the presence of the **type 2 diabetes**, which gave a positive contribution for the Logistic Regression. In this case, the presence of this type of diabetes sometimes reduces and sometimes increases the probability of assignment to the NGA class. Similarly, the behavior of **type 1 diabetes** does not seem to provide indications about the weight category in which the sample will be placed.

Significantly, the **number of pregnancies**, despite not showing a very clear distinction between the semi-planes, seems to lean towards a negative contribution to the prediction of the NGA class. This behavior contradicts that predicted by the Logistic Regression model.

Opposite contribution compared to the first model shown is also given by the **ternary Lempel-Ziv complexity index** calculated in 3-minute windows and the power in the **low frequencies of** a signal processed with the **PRSA** technique. Both provide a negative contribution here to the prediction of the NGA class compared to the other weight classes.

Finally, the MLP model attributes great importance to the power calculated in the **low frequencies** i.e. to the contribution of sympathetic nervous system. Although not clearly, high values of this parameter decrease the probability that a fetus will be predicted as NGA. In the Logistic Regression, this parameter was not considered particularly relevant, and in the summary plot it was not placed by SHAP among the top 20 variables in importance.

LGA plots

Finally, the result of the class LGA will be presented starting again from the illustration of the explanation of Logistic Regression model in Figure 3.7

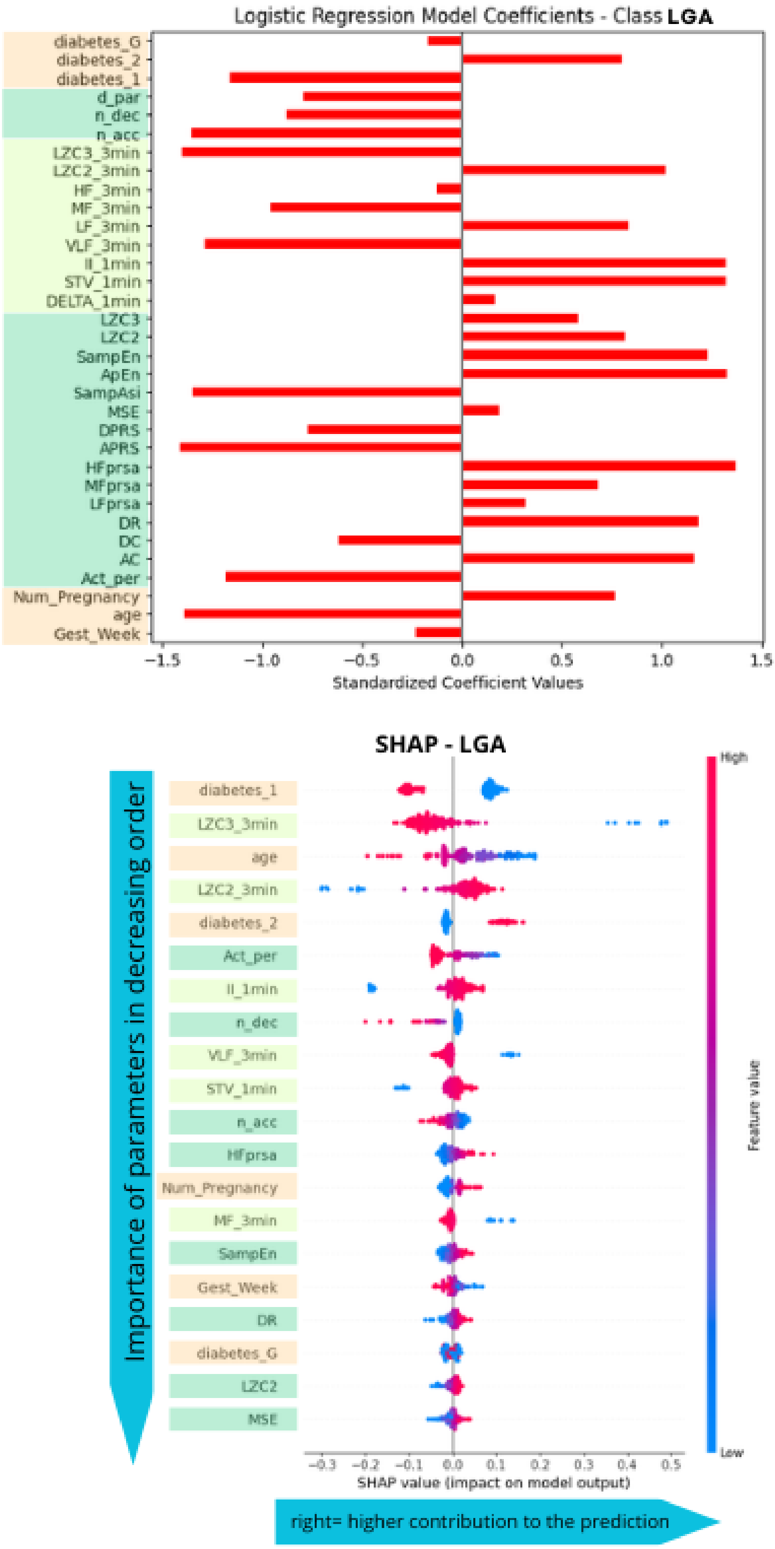

Figure 3.7: Illustration of the model coefficients and the SHAP summary plot for the LGA class of the Logistic Regression model.

In this case as well, there is consistency regarding the impact direction of the variables

on the prediction of the LGA class by both representations but often not in the order of importance.
Can be observed that:

- **Type 1 diabetes** is considered an extremely important variable for both representations. Its absence tends to increase the probability that a fetus will be predicted as LGA compared to other weight categories. The contribution of **Type 2 diabetes** instead promotes the prediction of this class. The last type of diabetes, the **gestational** one, is marginal. In this case does not have a clear contribution according to SHAP and has a very small negative contribution according to the coefficient plot.
- The **binary and ternary Lempel-Ziv complexity** coefficients calculated in 3-minute windows are very relevant. The ternary version of this parameter seems to decrease the probability of the LGA class for high values, while the opposite behavior is observed for its binary counterpart.
- Particularly relevant and impactful on the prediction in this case are the **short-term variability parameters** (**II_1min and STV_1min**) of the signal. An higher variability of the signal seems to indicate a greater probability for a fetus to be LGA with respect to the other weight classes.
- A higher **percentage of active segments** in the FHR signal significantly decreases the probability that a fetus is predicted as LGA compared to small or normal.
- Among the power bands, the most relevant is the one calculated in the **very low frequency**, which has been associated with nonlinear contribution, that impacts negatively to the model.
- On the other hand, opposite behavior is observed for the power calculated in the **high frequency band** from the signal processed with the **PRSA** technique, whose contribution is significant for this model. Also derived from PRSA, the substantial negative contribution of the parameter $APRS$ is noteworthy in the coefficient plot.
- **Acceleration capacity** and **deceleration reserve** indicate a higher probability of predicting the LGA class for high values, but only the second parameter is considered relevant by SHAP.
- The coefficient plot highlights with the length of the bars the significance of entropy measures such as **sample entropy** and **approximate entropy**, which positively contribute to the prediction. However, only sample entropy is ranked among the top 20 by SHAP. The same applies to **sample asimmetry**, where the coefficient plot shows a significant impact on reducing the probability of predicting the LGA

class, while SHAP does not consider it relevant.

– Regarding **accelerations** and **decelerations** in the FHR signal, the contribution of n_dec is more significant than that of n_acc according to SHAP, whereas it is the opposite according to the coefficient plot. However, both representations agree that these parameters decrease the probability of assigning the LGA class to samples.

Finally, also the summary plot of the MLP model for the class LGA will be shown in Figure 3.8

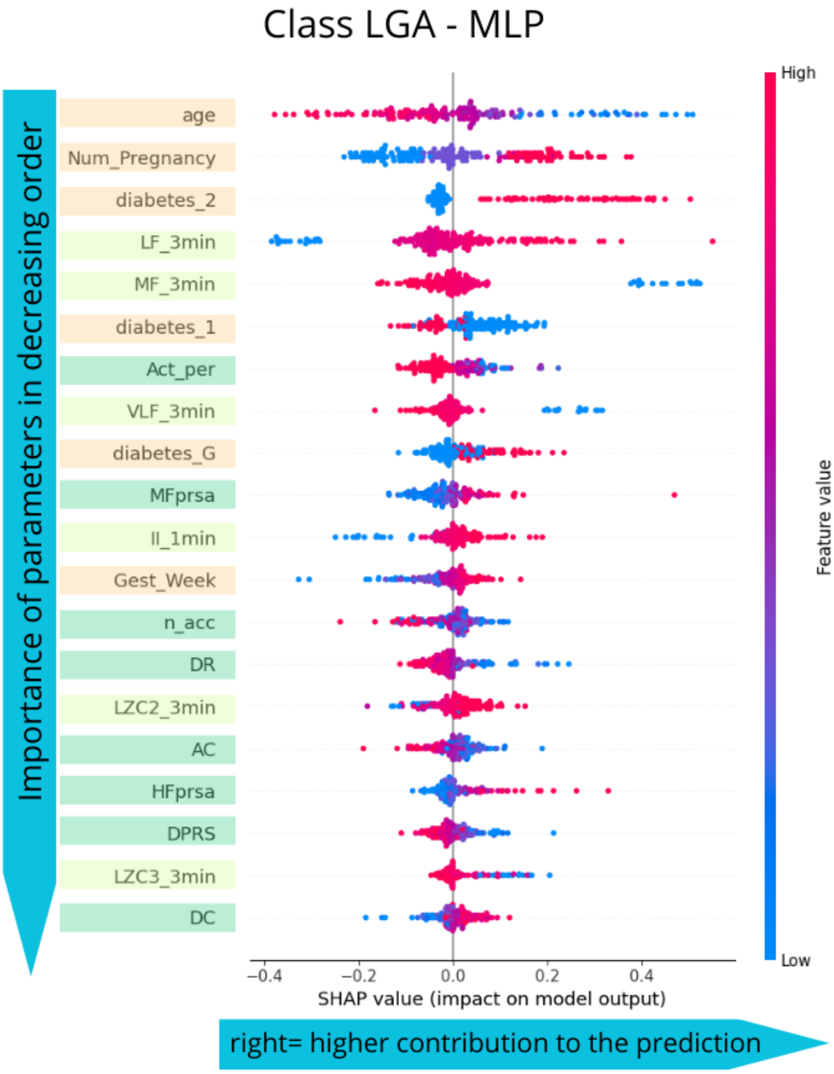

Figure 3.8: Illustration of the SHAP summary plot for the LGA class of the MLP model.

As can be seen, also for this class as for NGA, there are a lot of differences in the order of importance of the variables and in the direction of their contributions between the summary plot of this model and the graphs of the Logistic Regression.

A first difference that can be noticed with the MLP model compared to Logistic Regression is the different order of importance of the features. For example, in the case of MLP, less significance is given to the **Lempel-Ziv complexity indices** calculated over 3-minute windows, although they are still ranked among the top 20 variables.

Moreover, the power in the **movement and low frequency bands** appears to have greater relevance here: a lower presence of MF, associable to maternal breathing and fetal movement, increases the likelihood of an LGA fetus compared to other weight classes. A higher contribution from LF, that has been connected sympathetic system, on the other hand, increases the probability of the fetus being LGA.

The presence of **gestational diabetes** seems to contribute to predicting the LGA class. This type of diabetes was considered inconsequential in the SHAP plot of the Logistic Regression, whereas it was considered a negative contribution to prediction, albeit marginally, in the coefficient plot of this model.

Among the power calculated in signals processed with the **PRSA** technique, for the MLP, the one calculated in the movement frequency seems to be more relevant than that in the high frequency, unlike what was observed for Logistic Regression.

Unlike Logistic Regression, MLP does not consider the **number of decelerations** present in the FHR signal to be significant.

Finally, the **gestational week** and the **deceleration reserve** have opposite impacts compared to those highlighted for Logistic Regression. According to the MLP, the higher the gestational week, the higher the probability that the fetus is LGA compared to other weight classes, while the impact is opposite for the deceleration reserve.

The other variables considered important, on the other hand, have consistent contributions between the two models.

3.2.2. Waterfall plots on results

As stated in Chapter 2.5.1, another useful tool provided by this method is the waterfall plot, which allows to observe how variables have influenced the prediction of the target for a single sample. It is recalled that the waterfall plot is like a bar graph where, similar to the coefficient plot of Logistic Regression, the longer the bar, the greater the impact of that variable on the prediction. The predicted value by the model for the specific instance under examination is represented by $f(x)$, the baseline in the plot. The value of $f(x)$ is the one assigned by the softmax function in the case of multiclass classification. The bars of the variables can point towards $f(x)$ or in the opposite direction, thus expressing the direction of their contribution.

Misclassified

One use of this type of graph can be to investigate the contribution of variables to incorrect predictions to see which variables have led to errors.
For example in the following plot 3.9 will be presented the waterfall plot of a sample that belongs to the SGA class but it was predicted by the MLP as an NGA instead.

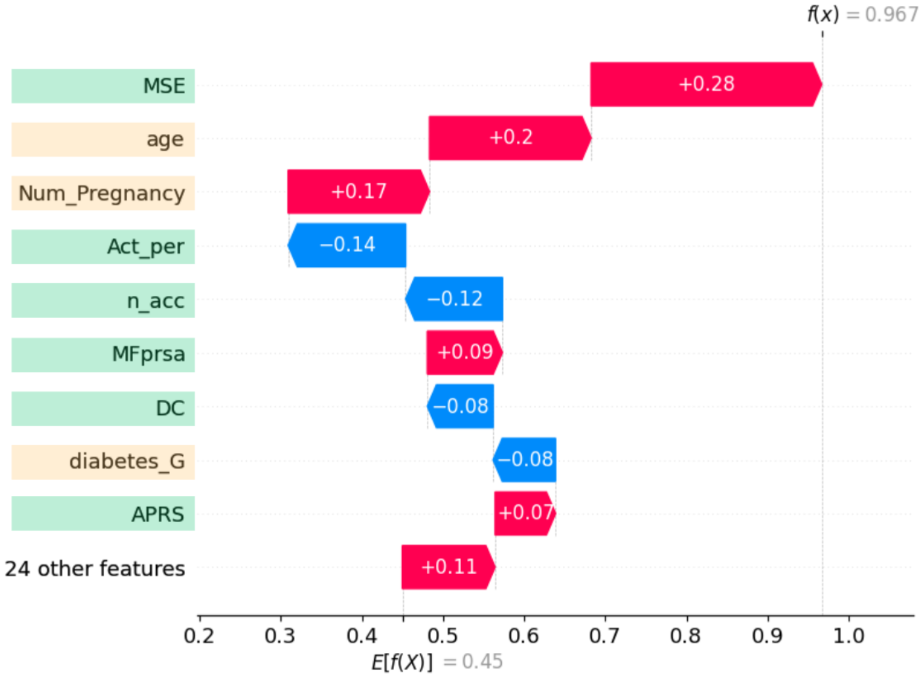

Figure 3.9: SHAP waterfall plot of an SGA sample that was predicted as NGA class by th MLP model.

Comparing it with the summary plot of this class, from the waterfall plot, it can be observed how the number of pregnancies, which for the MLP decreases the probability of predicting SGA, in this case pushes for an incorrect prediction. Unlike the global case, importance is then given instead to the power calculated in the movement frequency of the signal processed with PRSA.

An example of a fetus that should have normal weight for his gestational age but was predicted as SGA can be observed in the following graph (3.10).

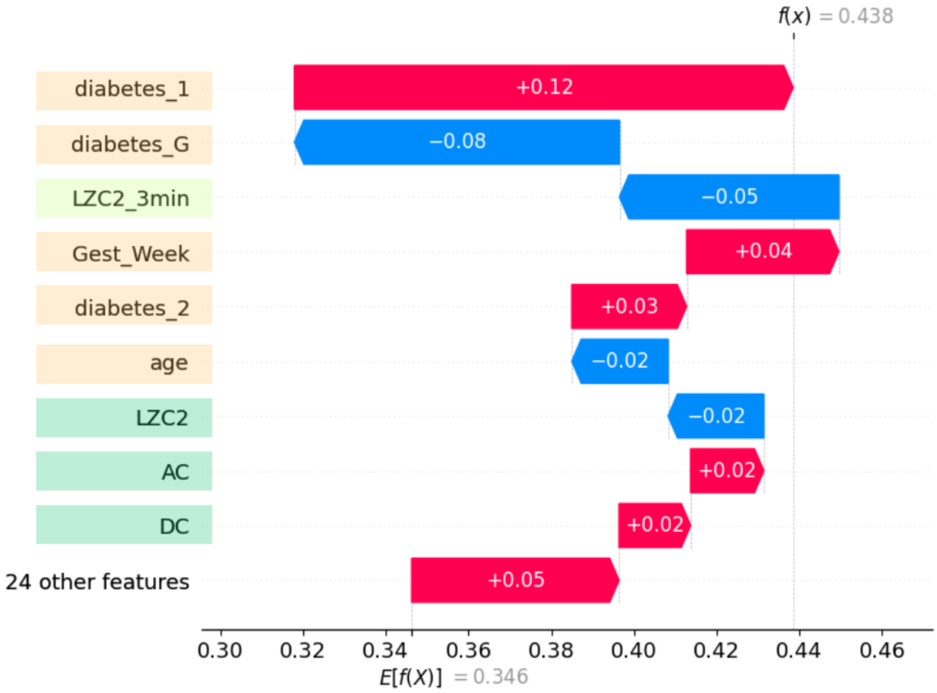

Figure 3.10: SHAP waterfall plot of an NGA sample that was predicted as SGA class by the Logistic Regression model.

As can be observed from this graph, the main factor responsible for predicting this sample as small is type 1 diabetes, which, according to the Logistic summary plot of this class, was supposed to have an inconsistent or even negative contribution.

In the end, another example is the following waterfall plot 3.11 showing a fetus that was supposed to be classified as LGA but was instead predicted as normal by the Logistic Regression.

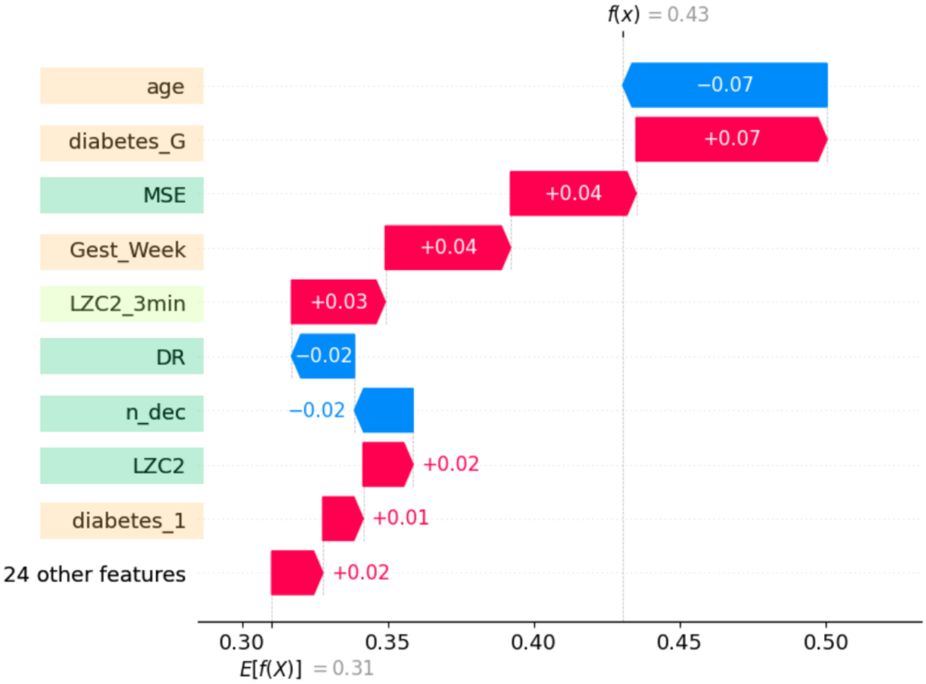

Figure 3.11: SHAP waterfall plot of an LGA sample that was predicted as NGA class by the Logistic Regression model.

Comparing it with the summary plot of the correct class, LGA, in this specific case, we can observe that the gestational diabetes has a significant impact on the prediction of the wrong class here. This type of diabetes was not very discriminant for the SHAP summary plot of this model for this class and even showed a negative contribution in the coefficient plot. Also the type 1 of diabetes behaves in the opposite way as expected looking at the global case and has very less importance than expected. The same opposite contribution is seen for the deceleration reserve.

These plots aforementioned further emphasize the importance for each model of the contributions from some of the most significant variables shown in the previous section (3.2.1). It has just been demonstrated in fact that if their contribution (positive or negative) to the prediction differs from the global case, this can lead a sample to be misclassified.

Parameters in the absence of activity segments

In section 2.3, discussing the approaches followed, it was stated how a "dummy" value of -1 had been definitively assigned to all parameters calculated in 1 and 3-minute windows in cases where the signal lacked activity segments and consisted only of quiet periods. Waterfall plots can be used to observe if there are differences in contributions from variables in samples characterized by these parameters.

Below, waterfall plots for both models will be presented for all 3 classes, considering samples whose parameters in windows were set to -1.

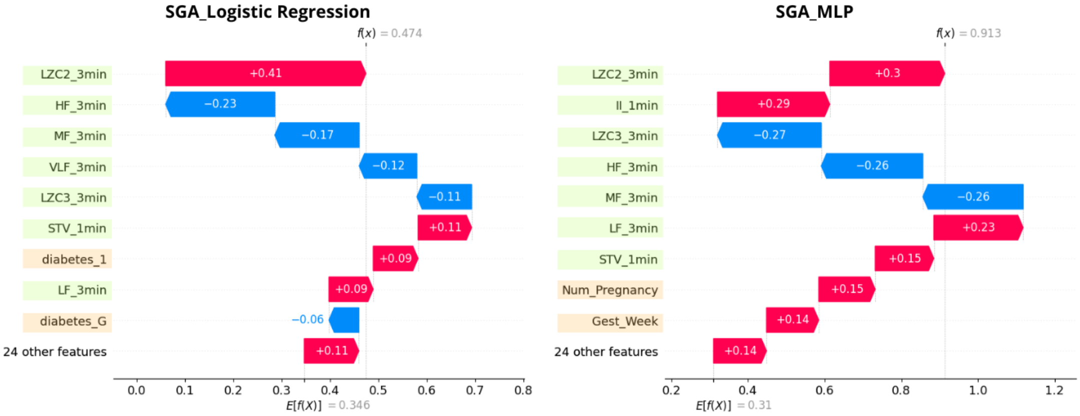

Figure 3.12: SHAP waterfall plot of an SGA samples without activity segments for both model.

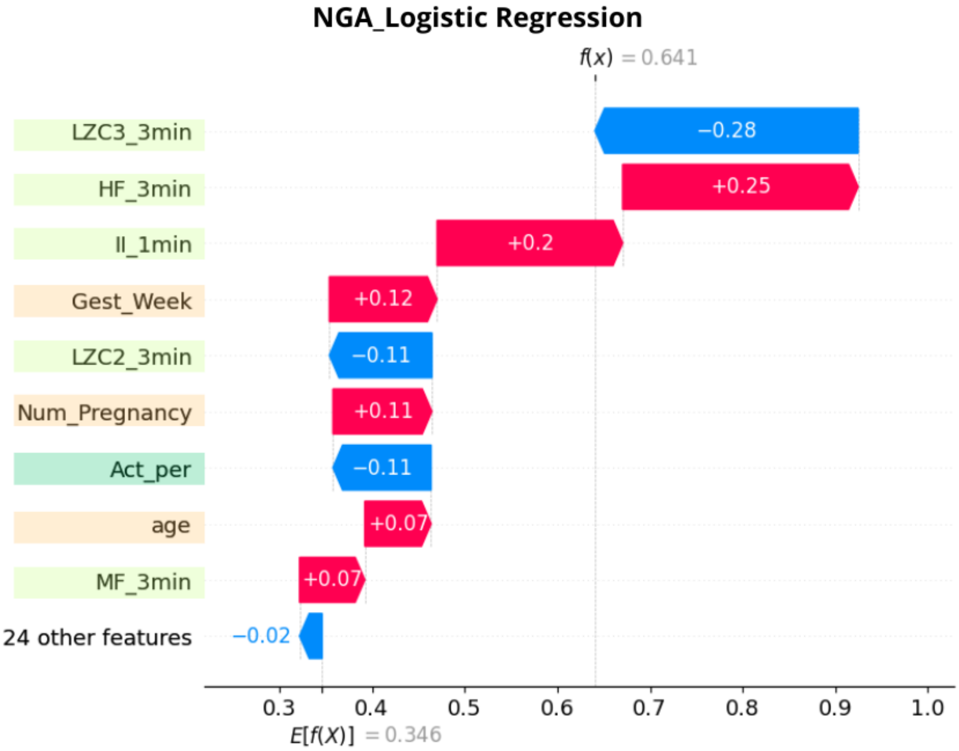

Figure 3.13: SHAP waterfall plot of an NGA samples without activity segments for Logistic Regression model. In the test set of MLP model, there were no samples of this type belonging to the NGA class.

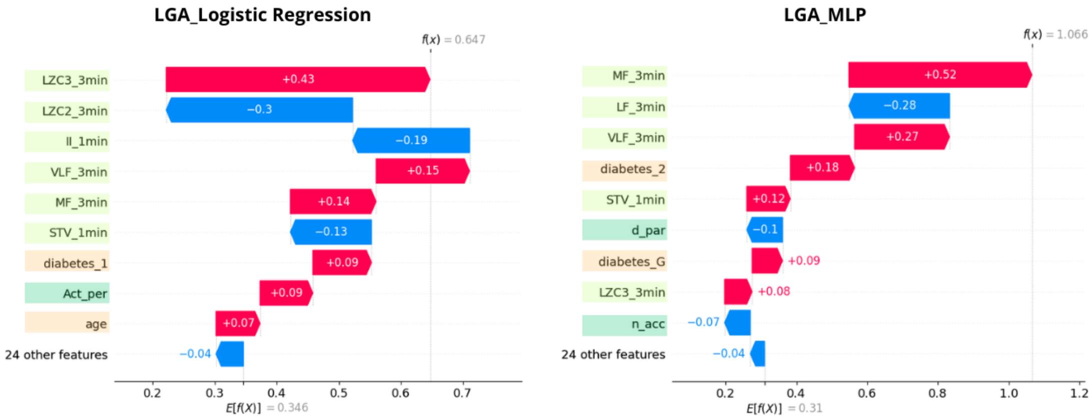

Figure 3.14: SHAP waterfall plot of an LGA samples without activity segments for both model.

Observing these plots, while keeping in mind that the waterfall plot can depend on the individual instance represented, it is immediately noticeable how the parameters calculated in 3-minute and 1-minute windows give by far the largest contribution, both positive and negative and are the most represented. This demonstrates that setting these variables to a value of -1 indeed influences the model, which learns that these samples carry specific information.

3.2.3. Interaction plots for gestational week

The last tool provided by SHAP that is employed in this thesis is the one that allows to see the relationship between two variables of the model and their impact on its prediction. In particular, for the three weight classes, the aim is to investigate the evolution of the main variables in relation to the gestational week. The fetal variables considered can significantly change depending on the gestational week. During pregnancy, in fact, the fetus goes through various stages of development and growth, and this is reflected in the values of the various fetal variables. Although in this specific case, the gestational week range is quite narrow, and therefore technically not subject to significant changes, it was desired to investigate how the impact of maternal parameters and features varies based on the trend of this variable.

For each class and each model, a graph representing the interaction between the gestational week and one of the variables considered most impactful for that model will be presented. The gestational week will be represented by the colors of the dots, with blue associated to low values and red associated with high value. On the x-axis in this case is represented the growing (from left to right) of the value of the variable being studied in relation to the gestational week, while on the y-axis the growing (from bottom to top) on its impact on the model.

SGA interaction plot

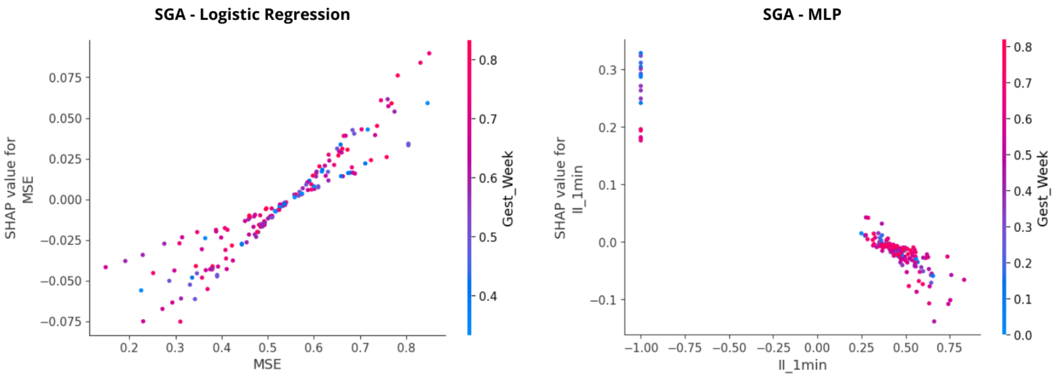

Figure 3.15: SHAP interaction plot of an SGA relevant variable with gestational week for Logistic Regresssion (on the left) and MLP (on the right) model.

According to the Logistic Regression model, it can be observed that there is an almost linear relationship between the increase in MSE values and its positive contribution to the model. In particular, although for most values this happens regardless of the gestational week, in the uppermost part, the most relevant, the prevalent presence of red points indicates high gestational weeks and therefore greater complexity with increasing weeks.

Looking instead at the graph on the right, concerning the MLP model, can observed how the trend of this variable, indicating short-term time domain variability, reflects its overall contribution. In particular, it can be noted again how the values set at -1 are relevant according to SHAP, and in particular how these values are mainly associated with a low gestational week value.

NGA interaction plot

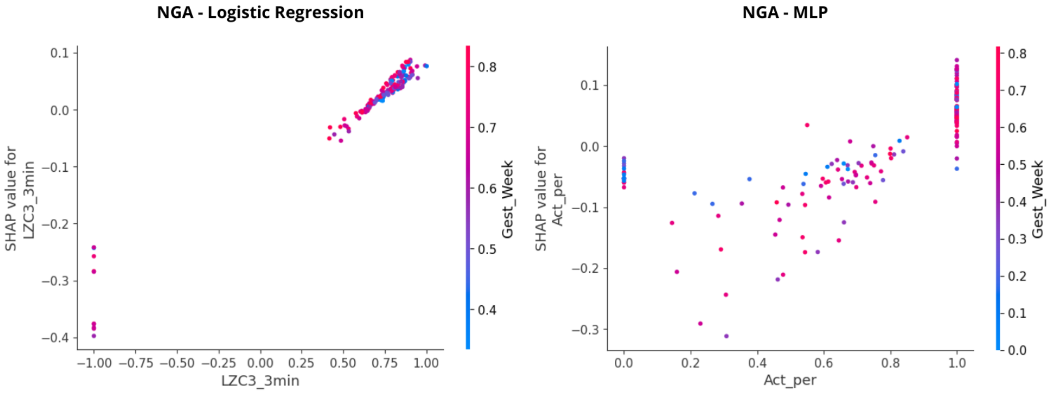

Figure 3.16: SHAP interaction plot of an NGA relevant variable with gestational week for Logistic Regresssion (on the left) and MLP (on the right) model.

According to the graph on the left, more complex signals will tend to make the model more likely to predict a fetus of normal weight compared to other classes, in accordance with the global contributions of this variable. This is almost independent of the gestational week at which the pregnant woman is.

It is interesting to observe the relationship between the percentage of activity segments in the FHR signal and the gestational week for the NGA class according to the MLP model. In the case of absence of activity (left part of the second plot), there is a small contribution to the prediction of the class with associated gestational weeks predominantly low. However, where the greatest contribution for the assignment of the NGA class occurs, practically when the activity percentage is 100%, the gestational week is higher. In the central part of the plot, a more homogeneous distribution can be observed, although low values of gestational week still seem to provide a small positive contribution (except in the last part, as already mentioned).

LGA interaction plot

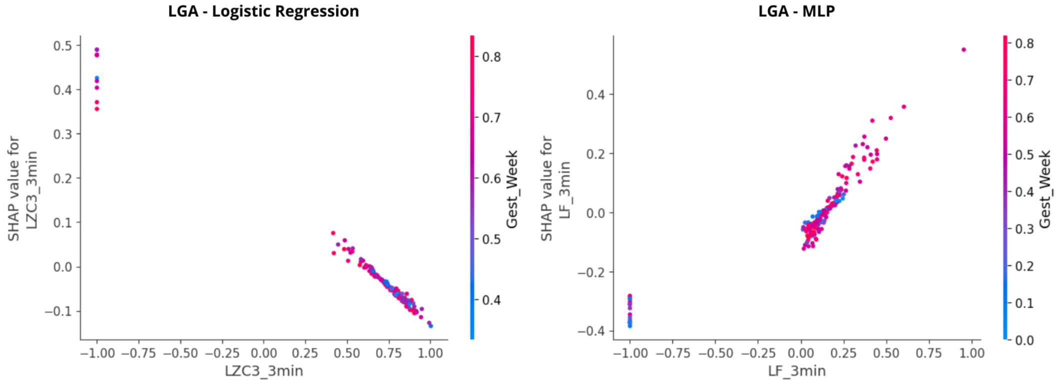

Figure 3.17: SHAP interaction plot of an LGA relevant variable with gestational week for Logistic Regresssion (on the left) and MLP (on the right) model.

In this case, the complexity index of the signal (in its ternary version), according to the Logistic Regression, have an opposite trend compared to the normal case. This is in accordance with the global contribution expressed. In correspondence with the values that were put equal to -1 in absence of activity segments, there is the higher SHAP value and this is in combination with fairly high gestational week values. For all the other values of $LZC3\_3min$ seems to be independence with the gestational week.

In accordance with what can be observed from the summary plot of the MLP model regarding the LGA class, an high power value in the low-frequency bands, therefore an higher contribution of the sympathetic system, contributes to a fetus being predicted as large for gestational age. This is clearly associated with a higher gestational week value as this contribution increases, as can be seen in the top right of the graph.

4 | Discussion

The main objective of this research was to train machine learning classification algorithms capable of categorizing pregnancies based on fetal weight, specifically identifying cases of Small for Gestational Age, Normal for Gestational Age, and Large for Gestational Age. This classification is crucial for preventing and managing potential complications during pregnancy and childbirth.

Another crucial goal of this study was to analyze and determine which among the multiple maternal characteristics, already available to clinicians from the medical history of the subject, and parameters derived from fetal signals collected through computerized cardiotocography (cCTG) are most relevant for fetal weight classification. Additionally, the aim was to explore the interactions and correlations existing between these features and parameters, in order to better understand how they are associated.

Through the analysis of this data, the hope is to give a contribution on the improvement of the accuracy with which doctors can predict and intervene in cases of fetal weight anomalies in already high-risk cases such as diabetic pregnancies. It is crucial to know whether this condition will lead to further risks/complications or not thus contributing to ensuring more positive outcomes for both maternal and fetal health.

4.1. Performance of the models

The results show significant progress towards achieving the primary objective, although it has not been fully attained as this tool is not yet ready for clinical application. Anyway the available variables, which include clinical data collected before childbirth, have proven effective in predicting the weight category in which the newborn will be classified. It is important to note that, in a classification model that distinguishes between three weight categories, a classifier operating randomly would have a success probability of 33%, corresponding to random selection among the three classes (SGA, NGA, LGA). The models developed in this research significantly surpass this baseline threshold. The Logistic Regression model achieved a balanced accuracy of 54.7%, while the Multilayer Perceptron attained a balanced accuracy of 52.6%. The results were further improved by performing majority voting, with Logistic reaching 55.1% and MLP reaching 59.6% showing that incrementing the number of recordings helps for a better diagnosis. These results not only surpass the effectiveness of a random classifier but also indicate a step forward in accurately predicting the fetal weight category using machine learning algorithms.

As already outlined in the Results chapter (3), other metrics have been used in addition to balanced accuracy to evaluate the performance of the models. They confirm the superior predictive capability of Logistic Regression model with respect to MLP. The confusion matrix was also employed to further investigate the actual value of balanced accuracy across classes.

For both models, it can be observed that the class most prevalent in the database, namely the NGA class, it's the best predicted class in terms of balanced accuracy. Following this, the classification of SGA and LGA reflects the actual distribution of available samples. This happens despite the classes being balanced. There could be unique features or different distributions that affect the model's performance. However, it should be noted that class balancing was meaningful for the reasons outlined in section 2.4.1, and furthermore, the lack of balancing would have led to a decrease in accuracy of approximately 10% for both models and an increase in overfitting.

As said, the results obtained surpass the performance of a random classifier, although they are not yet optimal. It is crucial to consider that, despite fetal weight disturbances and hypoxia being among the most common complications associated with diabetes, these topics are poorly explored in scientific literature. Current clinical practice mainly relies on anatomical data, such as abdominal measurements. Therefore, this study represents an important initial step towards a deeper understanding. It is important to note that in this work, there was a strong focus on using functional data reflecting fetal autonomous development, along with maternal clinical data, to estimate an anatomical parameter such as fetal weight. This approach offers a different perspective for assessing fetal well-being. The results indicate that this predictive approach is generally valid and feasible, suggesting that this research direction deserves further investigation.

4.2. Discussion upon variables

As previously stated, another objective of the work is to make it as clear as possible which variables contribute most to predict a particular weight category, distinguishing especially between maternal variables available to clinicians, including those derived from other types of exams or the pregnant woman's medical history, and those parameters calculated from the signal extracted with the cCTG. This is also to further discuss the benefits of its use.

After demonstrating in Chapter 3, through the use of an explainability method, which variables are considered most relevant by the two proposed models for each of the target classes, the physiological meanings of the values of these parameters will now be explored, also distinguishing to which group the most relevant features belong.

Referring to the summary plots for both models it is possible to infer the percentage of importance of parameters derived from the FHR signal and the extent of relevance of variables related to the pregnant woman's clinical history. This kind of plot shows in fact the top 20 variables that had the greatest impact on predicting the target class. In the case of Logistic Regression, this plot is preferred over the coefficient plot for consistency in comparison with MLP.

For the Logistic Regression model this ranking is expressed by the following Table 4.1:

	Top 5		**Top 10**	
	Maternal	Signal derived	Maternal	Signal derived
SGA	80%	20%	60%	40%
NGA	40%	60%	40%	60%
LGA	60%	40%	30%	70%

Table 4.1: Ranking of variables for the Logistic Regression model. The top 5 and the top 10 percentage of variables are shown

The same is done for the MLP model in the Table 4.2

	Top 5		**Top 10**	
	Maternal	Signal derived	Maternal	Signal derived
SGA	60%	40%	40%	60%
NGA	40%	60%	50%	50%
LGA	60%	40%	50%	50%

Table 4.2: Ranking of variables for the MLP model. The top 5 and the top 10 percentage of variables are shown

If the top 5 positions are considered, for both models, the contribution of variables derived from maternal clinical history is more impactful for both critical weight categories compared to the NGA class, which appears to focus more on the impact of parameters calculated from the FHR signal. However, there is greater balance when considering the top 10 positions of variable impact.

As observed from the table, the trained Logistic Regression model, identifying a fetus that might be small for gestational age, is the one that gives the highest overall weight to variables derived from the clinical history of pregnant women, compared to the MLP and other classes. On the other hand, the same model in prediciting the LGA gives in general the most importance to the FHR signal derived parameters.

Most relevant variables

As can be seen from the graphs in general, the analysis conducted on the impact of variables in predicting the classes has certainly confirmed the importance, already known in the literature, of information related to maternal clinical history.

Particularly relevant for Logistic Regression are, in fact, the **types of diabetes**, with the presence of type 1 and gestational diabetes pushing for a small fetal weight, while the opposite contribution comes from type 2 diabetes. The presence of the latter type of diabetes seems to be associated with both normal and large fetuses in terms of predicted weight. This may indicate that hyperinsulinemia associated with type 2 diabetes can promote fetal growth until reaching critical situations. However, it should be noted that the contribution of these variables to the MLP model is slightly different but also less discriminatory.

Regarding again the purely maternal characteristics for both models, it should be noted how a younger **age** of the mother tends only to make the fetus predicted large. Moreover, only in this case a positive contribution to the prediction is given by a high **number of pregnancies**. These maternal variables are therefore discriminating between LGA fetuses and the other weight categories.

There is then the indication of the **gestation week** where its higher numerical value is only associated with normal fetuses while a low value to the problematic weight conditions as regards the Logistic regression model, while the contribution given in the case of MLP is less clear. However, considering that the gestational week at which the recording is made is more of a clinical choice than a physiological factor, caution should be exercised in its interpretation. This variable could therefore be used as an aid in interpreting other parameters, as suggested in the 3.2.3 chapter.

The same summary plot graphs, however (and the same concept is summarized in table

4.1 and 4.2), have shown how the variables derived from the FHR still hold great importance in prediction. In fact, there is no marked difference in impact between the two categories of variables.
Some FHR signal derived parameters have been proven in fact to be able to distinguish between the 3 weight classes. This shows how to improve overall performance, it is useful to integrate diverse and heterogeneous information, which together provide a comprehensive picture.

A distinguishing characteristic, common for the SGA and LGA classes that sets them apart from the normal weight category, is that in the latter, there is a significant impact on the prediction of a high **number of accelerations** in the FHR signal. In problematic classes, instead, a low number of accelerations are prevalent.

In general, for the NGA class, importance is given to the large **percentage of activity segments** present in the FHR signal. This is in agreement with the theory that, as already expressed in section 2.3, a sign of fetal health and one of the main criteria used in the Dawes/Redman system to assess normality is the active sleep. It is also in agreement with those studies that have also highlighted a lower frequency of active sleep periods in problematic fetuses [24]. Consistently, the SGA and LGA classes are influenced in their prediction by low values of activity segments in the FHR signal.
The mode of contribution of these parameters mentioned first and their importance reflects the fact that the maturation and functioning of the autonomic nervous system in the fetus are more established in fetuses with normal weight than in problematic weight classes.

Particularly relevant for the Logistic Regression model seem to be then the **Lempel-Ziv complexity** indices calculated in windows of 3 minutes length. At least one of the binary and ternary indices always occupies a place in the top 5 positions among the most significant parameters for all weight categories. It is recalled that the difference between the two indices lies in the way references to subsets of data are encoded: with the binary or ternary system. Once the sequence is encoded, the algorithm with which the complexity is calculated is the same. The only case, however, in which these are consistent in their contribution is for the case of normal fetus where both agree in stating that a higher value; therefore, more complex signal with less repetitive patterns, tends to be associated with NGA fetus. In the SGA case, in fact, it is only the index deriving from ternary encoding of the sequences to favor the prediction of this class, while in the LGA case it is the index deriving from binary encoding that do the same. The importance of this parameters is also confirmed by the fact that its different contribution is responsible for an incorrect classification for the SGA class.
In all cases these indices are preferred to their version calculated on the entire signal.

Variations of this type of complexity index have been already found in diabetic population in [50].

It is interesting to note how, despite the opposite indication given by this parameter, both the SGA and NGA classes have in common the importance of **multiscale entropy**, which considers the variability of the FHR signal at different time scales, where even in the first case (SGA class) this parameter represents the only global one present among the top 10 positions for both models while in the second it is the only measure of entropy considered relevant again by both models. A lower entropy, therefore a lower complexity in the time scales, is associated with the NGA class. A higher entropy is instead associated with the SGA class. This is contrast with what can be seen in other studies regarding fetal well-being such as [40].
By comparing with the Lempel-Ziv complexity index previously discussed, it can therefore be stated that the evaluation of complexity is related to the way it is calculated.

For what concerns the contributions of the **VLF, LF, MF** and **HF** parameters (that have been associated to non linear components, sympathetic, fetal movements and maternal respiration and parasympathetic contribution respectively), they always appear in the most important positions regarding the impact on the weight classes, especially for the MLP model but there is not one predominant or that allows to clearly distinguish between the weight categories.

Non relevant variables

Finally, a consideration must also be made on those variables that had been provided to the models as information for their training but that were not considered relevant or significant in light of the interpretation provided by the explainability algorithm.

One of the least present parameters in the graphs is the **APRS**, a quantitative indicator of the average increase FHR amplitude, which appears, not with a relevant position, only in the summaryplot graph of SHAP of the MLP model for the SGA class. A better position is occupied by its equivalent that considers the average decrease of FHR amplitude even if, also in this case, never with particularly relevant contributions.

Sample Asimmetry is another parameter that is not generally considered important for SHAP framework across different weight classes. It is slightly relevant only for the coefficient plots of Logistic Regression for SGA and LGA.

Finally, among the less relevant parameters, the medium-term variability time index Δ (**DELTA** in the plots) should be mentioned. Only Logistic Regression, in its coefficient plot, prefers its contribution over that of the other temporal variability indices.

Overall Insights on Variable Analysis

The explainability analysis conducted revealed that some features, both maternal and derived from the fetal signal, are effective in differentiating between the various categories of fetal weight. On the contrary, other features proved to be less capable in this task. The study also highlighted that variables previously identified as discriminative between healthy and pathological populations may behave unexpectedly when applied within a population uniformly affected by a specific condition, such as diabetes. It is important to remember that the "normal" category discussed so far in the classification refers to fetal weight but is still relative to fetuses of diabetic mothers, thus considered pathological. Effects contrary to those expected have been observed, such as in the case of entropy, or even a lack of significant correlation, as for parameters associated with the contributions of the sympathetic or parasympathetic system or those attributable to a signal processed with PRSA technique, which in previous studies were considered discriminative. This underscores how the distinction within a pathological class is an underexplored field in research and makes the results obtained, in terms of explaining the contributions of the features, a significant starting point for further investigation.

5 | Conclusions

Pregnancy itself does not pose a life-threatening risk to the maternal-fetal system, but the antepartum period requires careful monitoring to prevent negative consequences. Indeed, complications can arise that significantly impact maternal and neonatal health. Among the most significant risks is diabetes: both type 1, type 2, and gestational diabetes increase the risk of complications for both mother and fetus. Among the diabetes-related complications, particular focus has been placed on weight disorders, especially the issues of small for gestational age and large for gestational age, using the most common diagnostic tool in clinical practice during pregnancy, namely cCTG.

The aim of this study was to train machine learning classifiers to predict which weight class the fetus of diabetic mothers would fall into, considering a third "normal" class in addition to the two mentioned issues, using information derived from both maternal clinical history and the FHR signal obtained from cCTG. The achieved results yielded an accuracy of 54.7% for the Logistic Regression model and 52.6% for the MLP model, which were further increased with majority voting. These results are significantly higher than those of a random classifier and should be considered encouraging.

Moreover, the analyses conducted using the SHAP method on the significance of the variables utilized, derived from maternal clinical history and referencing parameters extracted from the FHR signal, have highlighted how the impact of the former group is highly significant, as already known, especially considering, for example, the type of diabetes. However, their predictive ability was shown to be enhanced by the inclusion of the listed parameter group. In particular, concerning the parameters, the Lempel-Ziv complexity index, the number of signal accelerations, the MSE, and the percentage of activity segments should be highlighted.

The explainability analysis also highlighted that certain variables, previously deemed discriminative, exhibited unexpected behaviors when applied to uniformly affected diabetic maternal populations. This emphasizes the importance of delving deeper into distinctions within pathological classes, shedding light on potential nuances overlooked in prior research.

The integration of these types of information, which are therefore capable of providing insights from a functional perspective, not just anatomical as in standard clinical practice, along with the proposed classification tool, represents a significant initial advancement in the medical context (although still far from immediate clinical application). A first tool for accurate estimation of fetal weight can indeed influence and guide clinical decisions and intervention strategies, thereby improving health prospects for both mother and fetus. Furthermore, these results pave the way for further research to refine and optimize classification models in this area, progressively considering a more appropriate and informative set of parameters with the goal of achieving even higher levels of accuracy and reliability.

5.1. Limitations

It is necessary to consider some factors that have limited the analysis of the available data.

As mentioned in the introductory section of the database 2.1, the main information guiding the selection of specific samples from the Federico II Hospital database in Naples (particularly diabetes, its type, and fetal weight at birth) were derived from clinical annotations expressed in the $NOTE$ variable. These information were presented in textual form rather than as structured variables, which necessitated the use of less efficient and precise search methods, given the variety of terms, words, abbreviations, and characters used in different records. Another disadvantage of this lack of data structuring is the inevitable presence of transcription errors. Representing variables in this way required detailed analysis of each annotation to record all possible variants of the same information. This led to extended processing times, both computational and manual, a loss of efficiency, and the risk of omitting variables written in previously unidentified ways. Consequently, this reduced the number of available samples and potentially excluded the analysis of relevant information provided by non-selected samples.

Another limitation in terms of information stems from the division between the gynecology and obstetrics departments with other units such as pediatrics especially at communication level. This division implies the lack of communication of certain information in reference to the development over time of complications arising from maternal diabetes that could be used as additional discriminating variables for analysis. This is not only relevant to the specific objective under examination but more generally to fetal well-being monitoring. Having additional information regarding the outcome of childbirth or parameters immediately following this phase could lead to a focus on other variables among those available during pregnancy.

5.2. Future developments

The presented work had, as its main prerequisite in the sample selection, the focus on diabetes. This condition of pregnant women was considered as the sole clinical issue, without considering other conditions. An important aspect that could be considered as a subsequent phase of analysis is to also take into account other pathologies related to diabetes. Among the main ones, attention could be directed towards obesity and hypertension.

Maternal **obesity** is an escalating risk factor that can negatively impact offspring health, increasing the likelihood of obesity and diabetes. During pregnancy, the presence of obesity or diabetes can exacerbate obstetric complications, necessitating careful medical management [59]. There is a continuum in fetal exposure to hyperglycemia, ranging from type 1 diabetes to type 2 diabetes, to gestational diabetes, with an increasing impact on fetal development. Studies have highlighted that elevated maternal weight is closely associated with a significantly higher risk of developing GDM [60]. Excess glucose in the mother's blood, caused by this type of diabetes, can cross the placenta, leading to fetal overgrowth and hence a higher birth weight. Moreover, a high body mass index before pregnancy is linked to an increased risk of insulin resistance in the second trimester, which may result in negative consequences for the fetus, such as an increased risk of developing metabolic disorders in adulthood. Current recommendations emphasize the need to monitor weight gain during pregnancy, considering the long-term effects on the increased risk of heart disease and other chronic conditions. This link between fetal growth and future disease risk underscores the importance of proactive management and preventive interventions [59].

Chronic **hypertension** in pregnancy presents as elevated blood pressure that already exists and is documented before conception. Recent studies have highlighted a significant prevalence of this condition in women with both type 1 and type 2 diabetes [61]. With advancing age and prolonged duration of diabetes, there is an increase in chronic hypertension, which is associated with higher rates of prematurity and neonatal morbidity. Additionally, women who experienced hypertensive complications during pregnancy show long-term consequences such as the onset of chronic hypertension and cardiovascular diseases [62]. From a pathophysiological perspective, hypertension can compromise blood flow to the uterus, causing vasoconstriction, intrauterine growth restriction, hypoxia, and placental detachment. These alterations can negatively influence the neonate's birth weight, resulting in lower-than-normal weight at birth [61].

These associated pathologies could be considered as additional variables to be used both to increase the amount of information available for the pure prediction of weight classes and for a better interpretation of the results to understand the actual contribution of the various conditions on fetal weight issues.

Bibliography

[1] Edoardo Spairani, Beniamino Daniele, Maria Gabriella Signorini, and Giovanni Magenes. A deep learning mixed-data type approach for the classification of fhr signals. *Frontiers in Bioengineering and Biotechnology*, 10:887549, 2022.

[2] Gestational Diabetes Mellitus. Acog practice bulletin. *ACOG: Washington, DC, USA*, 2018.

[3] World Health Organization et al. 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.

[4] Annette Boles, Ramesh Kandimalla, and P Hemachandra Reddy. Dynamics of diabetes and obesity: Epidemiological perspective. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1863(5):1026–1036, 2017.

[5] HAPO Study Cooperative Research Group et al. Hyperglycemia and adverse pregnancy outcomes: the hapo study cooperative research group. *Obstetrical & Gynecological Survey*, 63(10):615–616, 2008.

[6] Moshe Hod, Anil Kapur, David A Sacks, Eran Hadar, Mukesh Agarwal, Gian Carlo Di Renzo, Luis Cabero Roura, Harold David McIntyre, Jessica L Morris, and Hema Divakar. The international federation of gynecology and obstetrics (figo) initiative on gestational diabetes mellitus: A pragmatic guide for diagnosis, management, and care. *International Journal of Gynecology and Obstetrics*, 131:S173–S211, 2015.

[7] Denice S Feig, Howard Berger, Lois Donovan, Ariane Godbout, Tina Kader, Erin Keely, Rema Sanghera, et al. Diabète et grossesse. *Can J Diabetes*, 42(3):S255–82, 2018.

[8] Practice bulletin no. 180: Gestational diabetes mellitus. *Obstetrics and gynecology*, vol. 130,1, 2017.

[9] H.Valensise D.Arduini. *elementi di cardiotocografia clinica, II edizione, cap. 8 pag 183-186*. CIC edizioni internazionali, 2007.

[10] Joachim A Behar, Zeev Weiner, and Philip Warrick. Special session on computational fetal monitoring. In *2019 Computing in Cardiology (CinC)*, pages Page–1. IEEE, 2019.

[11] Bahia Namavar Jahromi, Nahid Ahmadi, Nader Cohan, and Mehdi Roshan Nia Jahromi. Comparison of the umbilical artery blood gas, nucleated red blood cell, c-reactive protein, and white blood cell differential counts between neonates of diabetic and nondiabetic mothers. *Taiwanese Journal of Obstetrics and Gynecology*, 50(3): 301–305, 2011.

[12] Per Olofsson. Umbilical cord ph, blood gases, and lactate at birth: normal values, interpretation, and clinical utility. *American Journal of Obstetrics and Gynecology*, 2023.

[13] E Taricco, T Radaelli, G Rossi, MS Nobile de Santis, GP Bulfamante, L Avagliano, and I Cetin. Effects of gestational diabetes on fetal oxygen and glucose levels in vivo. *BJOG: An International Journal of Obstetrics & Gynaecology*, 116(13):1729–1735, 2009.

[14] Stefan Kuhle, Bryan Maguire, Hongqun Zhang, David Hamilton, Alexander C Allen, KS Joseph, and Victoria M Allen. Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: a retrospective cohort study. *BMC pregnancy and childbirth*, 18:1–9, 2018.

[15] Christian Giommi, Marta Lombó, Nina Montik, Michela Paolucci, Valentina Notarstefano, Giovanni Delli Carpini, Andrea Ciavattini, Antonio Ragusa, Francesca Maradonna, Elisabetta Giorgini, et al. Gestational diabetes mellitus and small-for-gestational-age: An insight into the placental molecular biomarkers. *International Journal of Molecular Sciences*, 24(3):2240, 2023.

[16] Yasushi Tsujimoto, Yuki Kataoka, Masahiro Banno, Shunsuke Taito, Masayo Kokubo, Yuko Masuzawa, and Yoshiko Yamamoto. Gestational diabetes mellitus in women born small or preterm: Systematic review and meta-analysis. *Endocrine*, pages 1–8, 2022.

[17] Kate McMurrugh, Matias Costa Vieira, and Srividhya Sankaran. Fetal macrosomia and large for gestational age. *Obstetrics, Gynaecology & Reproductive Medicine*, 2024.

[18] Grivell, RM, Alfirevic, Z, Gyte, GML, , Devane, and D. Antenatal cardiotocography for fetal assessment. *Cochrane Database of Systematic Reviews*, (9), 2015.

[19] Maria G Signorini, Nicolò Pini, Alberto Malovini, Riccardo Bellazzi, and Giovanni

Magenes. Integrating machine learning techniques and physiology based heart rate features for antepartum fetal monitoring. *Computer Methods and Programs in Biomedicine*, 185:105015, 2020.

[20] H.Valensise D.Arduini. *elementi di cardiotocografia clinica, II edizione, 4-5.* CIC edizioni internazionali, 2007.

[21] Xiaotian Li, Daan Zheng, Shufeng Zhou, Dakan Tang, Caiyan Wang, and Guoqiang Wu. Approximate entropy of fetal heart rate variability as a predictor of fetal distress in women at term pregnancy. *Acta Obstetricia et Gynecologica Scandinavica*, 84(9): 837–843, 2005.

[22] Salvatore Tagliaferri, Andrea Fanelli, Giuseppina Esposito, Francesca Giovanna Esposito, Giovanni Magenes, Maria Gabriella Signorini, Marta Campanile, Pasquale Martinelli, et al. Evaluation of the acceleration and deceleration phase-rectified slope to detect and improve iugr clinical management. *Computational and Mathematical Methods in Medicine*, 2015, 2015.

[23] Maria Ribeiro, João Monteiro-Santos, Luísa Castro, Luís Antunes, Cristina Costa-Santos, Andreia Teixeira, and Teresa S Henriques. Non-linear methods predominant in fetal heart rate analysis: a systematic review. *Frontiers in Medicine*, 8:661226, 2021.

[24] Lisa Stroux, Christopher W Redman, Antoniya Georgieva, Stephen J Payne, and Gari D Clifford. 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.

[25] H.Valensise D.Arduini. *elementi di cardiotocografia clinica, II edizione, cap.4 pag 112-119.* CIC edizioni internazionali, 2007.

[26] Edoardo Spairani, Beniamino Daniele, Giovanni Magenes, and Maria G Signorini. 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.

[27] Paul Hamelmann, Rik Vullings, Alexander F Kolen, Jan WM Bergmans, Judith OEH van Laar, Piero Tortoli, and Massimo Mischi. Doppler ultrasound technology for fetal heart rate monitoring: a review. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 67(2):226–238, 2019.

[28] Giovanni Magenes, MARIA GABRIELLA Signorini, MANUELA Ferrario, and

F Lunghi. 2ctg2: A new system for the antepartum analysis of fetal heart rate. pages 781–784, 2007.

[29] H.Valensise D.Arduini. *elementi di cardiotocografia clinica, II edizione, cap. 11 pag 251-256*. CIC edizioni internazionali, 2007.

[30] Anucha Thatrimontrichai, Kan Charernjiratragul, Waricha Janjindamai, Supaporn Dissaneevate, Gunlawadee Maneenil, Manapat Phatigomet, and Nattachai Anantasit. Correlation and prediction of arterial partial pressure of carbon dioxide from venous umbilical blood gases. *The Turkish Journal of Pediatrics*, 64(1):85–91, 2022.

[31] Torvid Kiserud, Gilda Piaggio, Guillermo Carroli, Mariana Widmer, Josè Carvalho, Lisa Neerup Jensen, Daniel Giordano, Josè Guilherme Cecatti, Hany Abdel Aleem, Sameera A Talegawkar, 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.

[32] H.Valensise D.Arduini. *elementi di cardiotocografia clinica, II edizione, cap. 11 pag 243-249*. CIC edizioni internazionali, 2007.

[33] R. Mantel, H.P. van Geijn, F.J.M. Caron, J.M. Swartjes, E.E. van Woerden, and H.W. Jongsma. Computer analysis of antepartum fetal heart rate: 1. baseline determination. *International Journal of Bio-Medical Computing*, 25(4):261–272, 1990.

[34] R. Mantel, H.P. van Geijn, F.J.M. Caron, J.M. Swartjes, E.E. van Woerden, and H.W. Jongswa. Computer analysis of antepartum fetal heart rate: 2. detection of accelerations and decelerations. *International Journal of Bio-Medical Computing*, 25 (4):273–286, 1990.

[35] Massimo Walter Rivolta, Tamara Stampalija, Martin G Frasch, and Roberto Sassi. Theoretical value of deceleration capacity points to deceleration reserve of fetal heart rate. *IEEE Transactions on Biomedical Engineering*, 67(4):1176–1185, 2019.

[36] Edoardo Spairani, Giulio Steyde, Luca Subitoni, Giovanni Magenes, and Maria G. Signorini. A semi-supervised deep learning approach to automate the identification of fetal behavioral states in fetal heart rate tracings. 2024.

[37] Fred Shaffer and Jay P Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in public health*, 5:290215, 2017.

[38] Maria G Signorini, Giovanni Magenes, Sergio Cerutti, and Domenico Arduini. Linear and nonlinear parameters for the analysisof fetal heart rate signal from cardiotoco-

graphic recordings. *IEEE Transactions on Biomedical Engineering*, 50(3):365–374, 2003.

[39] Steven M Pincus, Igor M Gladstone, and Richard A Ehrenkranz. A regularity statistic for medical data analysis. *Journal of clinical monitoring*, 7:335–345, 1991.

[40] Manuela Ferrario, Maria G Signorini, Giovanni Magenes, and Sergio 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*, 53(1):119–125, 2005.

[41] Steven M Pincus. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6):2297–2301, 1991.

[42] Joshua S Richman and J Randall Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American journal of physiology-heart and circulatory physiology*, 278(6):H2039–H2049, 2000.

[43] Madalena Costa, Ary L Goldberger, and C-K Peng. Multiscale entropy analysis of biological signals. *Physical review E*, 71(2):021906, 2005.

[44] Boris P Kovatchev, Leon S Farhy, Hanqing Cao, M Pamela Griffin, Douglas E Lake, and J Randall Moorman. Sample asymmetry analysis of heart rate characteristics with application to neonatal sepsis and systemic inflammatory response syndrome. *Pediatric research*, 54(6):892–898, 2003.

[45] Abraham Lempel and Jacob Ziv. On the complexity of finite sequences. *IEEE Transactions on information theory*, 22(1):75–81, 1976.

[46] P Malcus, AA Baschat, A Lempel, J Ziv, MG Signorini, G Magenes, S Cerutti, D Arduini, GW Lawson, R Belcher, et al. Comparison between fetal heart rate standard parameters and complexity indexes for the identification of severe intrauterine growth restriction. *Methods of information in medicine*, 46(02):186–190, 2007.

[47] Manuela Ferrario, Maria G Signorini, and Giovanni Magenes. Complexity analysis of the fetal heart rate variability: early identification of severe intrauterine growth-restricted fetuses. *Medical & biological engineering & computing*, 47:911–919, 2009.

[48] Axel Bauer, Jan W Kantelhardt, Armin Bunde, Petra Barthel, Raphael Schneider, Marek Malik, and Georg Schmidt. Phase-rectified signal averaging detects quasi-periodicities in non-stationary data. *Physica A: Statistical Mechanics and its Applications*, 364:423–434, 2006.

[49] Andrea Fanelli, Giovanni Magenes, Marta Campanile, and Maria G Signorini. Quantitative assessment of fetal well-being through ctg recordings: a new parameter based on phase-rectified signal average. *IEEE Journal of Biomedical and Health Informatics*, 17(5):959–966, 2013.

[50] Giulio Steyde, Beniamino Daniele, Edoardo Spairani, Giovanni Magenes, and Maria Gabriella Signorini. Influence of gestational diabetes on fetal heart rate in antepartum cardiotocographic recordings. In *2022 Computing in Cardiology (CinC)*, volume 498, pages 1–4. IEEE, 2022.

[51] Silvia M Lobmaier, Nico Mensing van Charante, Enrico Ferrazzi, Dino A Giussani, Caroline J Shaw, Alexander Müller, Javier U Ortiz, Eva Ostermayer, Bernhard Haller, Federico Prefumo, et al. Phase-rectified signal averaging method to predict perinatal outcome in infants with very preterm fetal growth restriction-a secondary analysis of truffle-trial. *American journal of obstetrics and gynecology*, 215(5):630–e1, 2016.

[52] Giulio Steyde, Edoardo Spairani, Giovanni Magenes, and Maria G Signorini. Fetal heart rate spectral analysis in raw signals and prsa-derived curve: normal and pathological fetuses discrimination. *Medical & Biological Engineering & Computing*, 62(2): 437–447, 2024.

[53] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[54] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[55] Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A Becker, and Catherine Mooney. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*, 11(11):5088, 2021.

[56] Urja Pawar, Donna O'Shea, Susan Rea, and Ruairi O'Reilly. Incorporating explainable artificial intelligence (xai) to aid the understanding of machine learning in the healthcare domain. In *AICS*, pages 169–180, 2020.

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

[58] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[59] David Simmons. Diabetes and obesity in pregnancy. *Best practice & research Clinical obstetrics & gynaecology*, 25(1):25–36, 2011.

[60] Susan Y. Chu, William M. Callaghan, Shin Y. Kim, Christopher H. Schmid, Joseph Lau, Lucinda J. England, and Patricia M. Dietz. Maternal Obesity and Risk of Gestational Diabetes Mellitus. *Diabetes Care*, 30(8):2070–2076, 08 2007.

[61] Keenan E Yanit, Jonathan M Snowden, Yvonne W Cheng, and Aaron B Caughey. The impact of chronic hypertension and pregestational diabetes on pregnancy outcomes. *American journal of obstetrics and gynecology*, 207(4):333–e1, 2012.

[62] Antonietta Colatrella, Valentina Loguercio, Luca Mattei, Massimo Trappolini, Camilla Festa, Michela Stoppo, and Angela Napoli. Hypertension in diabetic pregnancy: impact and long-term outlook. *Best practice & research Clinical endocrinology & metabolism*, 24(4):635–651, 2010.

List of Figures

List of Tables

Acknowledgements

Ringrazio la Professoressa Maria Gabriella Signorini per l'opportunità di questo progetto di tesi, Giulio Steyde per l'enorme aiuto e pazienza e per essere stato un supporto costante, e infine la Dottoressa Marta Campanile per la disponibilità dimostrata.