Master thesis : Development of a new framework for (semi-) automated antepartum electronic fetal monitoring analysis and Intra-Uterine Growth Restriction detection

Auteur : Servais, Youri
Promoteur(s) : Sacré, Pierre
Faculté : Faculté des Sciences appliquées
Diplôme : Master en ingénieur civil biomédical, à finalité spécialisée
Année académique : 2021-2022
URI/URL : http://hdl.handle.net/2268.2/16338

University of Liège - School of Engineering and Computer Science

Development of a new framework for (semi-) automated antepartum electronic fetal monitoring analysis and Intra-Uterine Growth Restriction detection

Master Thesis with the aim of obtaining the degree of
Master in Biomedical Engineering

Author: **Youri Servais**

Student ID: 20173986
Advisor: Prof. Pierre Sacré
Co-advisors: Julien Penders, Maria Gabriella Signorini
Academic Year: 2021-22

Abstract

In this value based world, life and health are probably the most essential things that should be left out of inequalities and injustice. Unfortunately, these differences arise even before the birth due to hazard, but also abilities to monitor the maternal pregnancy. As an example of pathology affecting the fetus, Intra Uterine Growth Restriction (IUGR) is a fetal condition defined as the abnormal rate of growth and causing fetal or neonatal morbidity and mortality. Currently, clinicians can only suspect IUGR condition by estimating the birthweight with Ultra-Sound imaging. This prediction is only based on the weight estimationof the fetus but is not related to real fetus well-being. Thus, IUGR conditon can only be confirmed at birth. Hence, it would be interesting to have an additional tool related to fetal well-being allowing clinician to diagnose it during the pregnancy.

Cardiotocography (CTG) is one of the most used technique to assess fetal well-being during pregnancy. In this project, we will use CTG signals and more specifically Fetal Heart Rate (FHR) signal to predict the fetal condition before labour (antepartum). The work will focus on IUGR pathology and build a framework allowing us to detect it. Based on the literature, a set of features interesting for the analysis was determined allowing us to characterise the FHR signal of each subject. These parameters were then analysed over 3 datasets coming from different sources (*Politecnico di Milano*, *Bloomlife* and open-source data from *Data in Brief* [44]).

Finally, adjusted open-source data from 120 subjects (60 Healthy/ 60 IUGR) was used to implement and train 4 classification models and evaluate them. The selected model is a Bagged Ensemble composed of 5 decision trees. Analysis and optimization of the model were made on the training/validation dataset to improve its performance. As a final result, the model achieves to reach a global accuracy of 87% and a 96% sensibility on the validation dataset and succeed to classify 18 subjects over 20 on our test set with a 100% sensibility.

As a perspective, a bigger annotated signals dataset could be used to implement and train a strong model based directly on raw FHR signals. This prediction model could be used to offer an easily accessible tool to detect IUGR during pregnancy and distinguish them from physiological Small for Gestational Age fetuses.

Keywords: Fetal Heart Rate Monitoring, Intra Uterine Growth Restriction, Signal Processing, Physiology-based Parameters, Feature Analysis, Machine learning model

Acknowledgements

Firstly, I would like to express my gratefulness to Bloomlife and more especially to Julien Penders to have given me the opportunity to work on this inspiring project. I would also like to thank Pr Sacré for his important advice and remarks on the project.

"La ringrazio molto" to Politecnico di Milano to have welcomed me during the first quadrimester and inspired me a lot. More specifically, *grazie* to Pr Signorini that helped me to find the end-point of my thesis and allow me to work on it troughout the precious data given.

In addition, I would also like to emphasise the opportunity that I had to work with the wonderful Bloomlife team including Quentin, Eva, Michiel and my intern friend Pierre. Thanks to them for all the meetings and the time given to me and the project. But even more also for beeing able to handle me during lunchbreak and teambuildings. Special thanks to Eva to for her precious follow up of the work and her multiple reviews. This work would not be the same without her. Thanks also to Fréderique Laloux for the language review and remarks that helped me to put correctly in words all the things I had in my mind.

It would not be correct to forget to thank all the people around me that were big supports from the beginning to the end. "Eskerrik asko" to my girlfriend that helped from close to far in my work and was there for me even in the worse moments all along the work. Thanks also to my friends for having been there to motivate me with a couple of beers when I needed it. Finally, thanks to my family and even more my parents to trust me and push me in all situations in order to be able do dream big.

Contents

1 | CTG monitoring in details

To understand better the signals on which we will work in this project, we will first have an overview of CTG monitoring. In this section, a focus on Cardiotographic signals will be made. We will first explain of what is composed a CTG signal (FHR, MHR and TOCO) and how it can be measured. Then, we will focus on the current clinical use, list the different features that can be analysed in the Fetal Heart Rate (FHR) signal and explain how it is used for diagnosis.

1.1. CTG signal explanation

Cardiotocography (CTG) can be defined as : "A graphic record of the Fetal Heart Rate and uterine contractions through an ultrasound device placed on the maternal abdomen or through a fetal scalp electrode. The 'toco', registers the uterine contractions through a second transducer placed on the uterine fundus." [18] CTG is most commonly used in the third trimester, aims to monitor fetal well-being and allow early detection of fetal distress. Indeed, after analysing if the signal is "reactive" or "non-reactive", the clinician may indicate the need for further investigations and potential intervention. [48]

CTG monitoring can either be measured externally or internally (using scalp electrode inside the neck of the womb). Since we will use external CTG signals in the frame of this work, we will only develop how external measurements are made in practice. For external measurement, the equipment used to monitor the baby's heart is placed on the tummy (abdomen) of the mother. An elastic belt is placed around the mother's abdomen. It has two round, flat about the size of a tennis ball which make contact with the skin. One of these plates measures the baby's heart rate (FHR). The other assesses the pressure on the tummy (TOCO). (cfr figure 1.1)

In this way, it is able to show when each contraction happens and an estimates how strong it is. The CTG sensors are connected to a machine which interprets the signal coming from the plates. The baby's heart rate can be heard as a beating or pulsing sound which the machine produces. The machine provides then a printout or an electronic record which shows the baby's heart rate over a certain length of time. [51]

The other plate on the CTG measures how tense the mother's tummy (abdomen) is. This measurement is used to show when the uterus is contracting.

Figure 1.1: Image of CTG monitoring set up. The left electrode is used to sense the Fetal Heart Rate whereas the right electrode measures the uterine contraction (TOCO)

1.1.1. FHR : Fetal Heart Rate signal

The FHR is usually both displayed on the machine and recorded on the CTG paper. It is measured using a Doppler ultrasound signal. To make these measurements, the Doppler sensor uses Ultrasound waves (>20kHz) with Doppler theory. Indeed, by measuring the frequency shift between emitted signal and reflected signal, we can measure the speed of the organ relative to the probe. This gives us information on fluid or organ movement and so on heart beat.

The ultra-sound wave transducer is made of a piezoelectric crystal in which a High frequency electrical current is injected. In this way, the crystal will change shape according to the current injected in it (emission of wave) but also when subjected to the different wave pressure that will change the polarity of the crystal and so the electric current (reception). Therefore, the piezoelectric transducer will be responsible of both emission and reception of the Ultra-sound waves. [18]
The frequency shift due to the heart movement is called the Doppler effect and can be computed with the following formula:

$$f_R = \frac{2f_o}{c} V cos(\theta) \tag{1.1}$$

where f_R is the measured change in frequency [Hz], f_o is the emitted ultrasound frequency [Hz], c the speed of sound in the tissue (m/s), V the velocity of the reflecting target [m/s] (in this case: the fetal heart) and θ the angle with the ultrasound beam. [6].

By measuring the frequency shift, it is possible to measure the velocity of the reflecting target giving information on the movement of the heart and its beat. A schematic of how FHR measurement using Doppler theory can be seen in the figure 1.2.

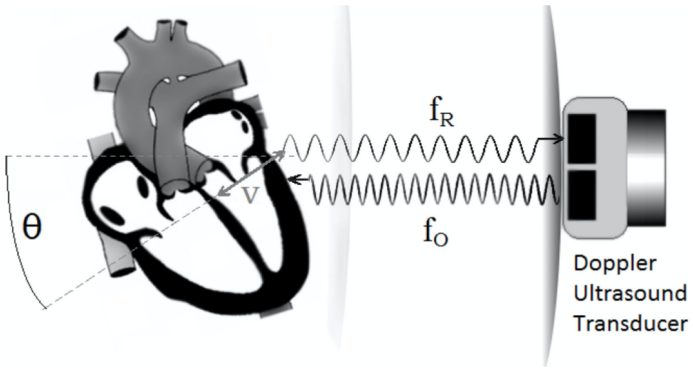


Figure 1.2: Representation of how the Doppler Ultrasound transducer is measuring the FHR. The transducer is placed on the maternal abdomen and emits ultrasound waves at the frequency f_O and receives the reflected wave with the frequency f_R. The frequency shift is computed according the formula 1.1. Figure from Cardiotocography and Beyond report[18] .

1.1.2. TOCO : Measure of the Uterine Activity (tocograph)

As explained earlier, the "TOCO" is placed on the maternal anterior abdominal wall over the fundus of the uterus and held by the elastic belt (cfr figure 1.1) [17]. It monitors the frequency and the length of the uterine contraction but not the strength. The amplitude of the signal is related to the change in shape and tone of the abdominal wall but do not reflect the strength of the contraction. The change of the abdominal wall during the uterine contraction creates a pressure wave that is recorded by the tocograph.

Sometimes, other factors can also change the shape of the abdominal wall. That's why a CTG monitoring is often accompanied by the mother's annotation of her contractions and/or a fetal movement appreciation. It is asked to the mother to push a specific button when feeling contractions. This allows the clinician to compare her annotation with the tocograph.

1.1.3. CTG Display and Recording example

As explained earlier, CTG monitoring can record and display several different signals :

- The Fetal Heart Rate (**FHR)**
- The Uterine activity (**TOCO)**
- The Fetal movement measurement
- The maternal annotations
- The Maternal Heart Rate (**MHR**)

An example of the CTG display and paper printout recording is showed in the figure 1.3

The figure 1.4 shows an example of a CTG monitoring composed of the FHR (A), the TOCO (D),

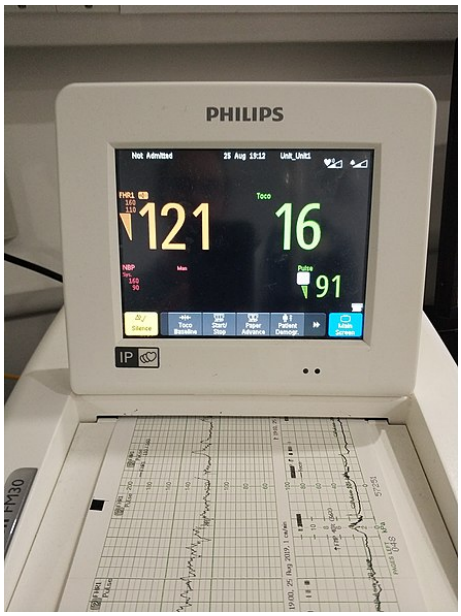


Figure 1.3: Display of a cardiotocograph. The FHR [bpm] is shown in orange, uterine contractions are represented in green (TOCO) [mmHg], and the small green numbers (lower right) represent the mother's heartbeat [bpm]. Below the printout of the recording is shown. [8]

the maternal annotations (B) and the fetal movement measurement (C).

The fetal movement is measured from the same Doppler signal than the FHR. It is obtained using a bandpass filtering, since it is generally associated with a lower bandwidth than the fetal heart wall movements. Indeed, a movement of 1-3cm/s will be reflected at 20-80Hz range if we have ultrasound of 2MHz [27]. The CTG shown here is a typical antepartum CTG (not in labour). We will work with this kind of signals (electronically recorded) in this project and more specifically with FHR signals.

1.2. Current use of CTG signal by clinicians

In case of low-risk delivery, CTG monitoring is not usually needed. Its use varies from a country to another but also from a doctor to another. Indeed, in Belgium in the case of low-risk pregnancy, there is not any mandatory CTG monitoring and its use belongs to gynecologist preference.

On the other hand, in case of a high risk pregnancy, a continual CTG monitoring is advised. This situation can be due to many different reasons such as :

- Maternal illnesses : Gestational diabetes, hypertension, asthma
- Obstetric complications : Multiple or post-term gestation, **IUGR**, previous caesarian, etc
- Other risks factors : Smoking, drugs, etc

A set of guidelines was defined by the International Federation of Gynecology and Obstetrics

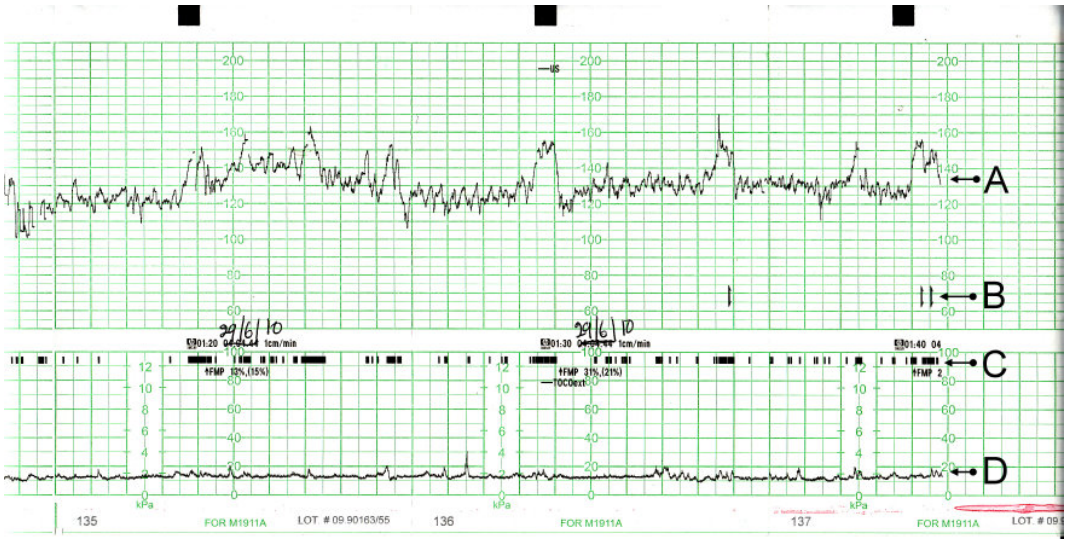


Figure 1.4: Typical CTG output printout for a woman not in labour. A: Fetal heart rate (FHR) [bpm]; B: Indicator showing movements felt by mother (triggered by pressing a button); C: Fetal movement; D: Uterine contractions (TOCO) [mmHg]. Here the CTG is displayed following the US convention [1cm square/min] [8]

(FIGO) for intrapartum fetal monitoring. [4]. In this report, it is said that most experts believe that continuous CTG monitoring should be considered in the case of a risk of fetal hypoxia and acidosis. Continuous CTG monitoring is then recommended when abnormalities are detected during earlier intermitent auscultation. It gives also a set of diagnosis indication for CTG signal analysis with respect to different features. Those features will be explained in more detail in the next subsection.

1.2.1. CTG signal features and diagnosis

In this section, we will explain in what consists a non-stress test and detail the different features used and their analysis as stated in the FIGO guidelines [4]. For a deeper analysis, one can also refers to the Handbook for CTG Interpretation [18].

Non-Stress tests are screening test used in antepartum to assess fetal well-being. The CTG to monitor fetal heart rate is combined with the pressure in the maternal abdomen. The clinician studies this signal and then terms it as "reactive" and "non-reactive".

First of all, defining the potential risk of the pregnancy (High or Low-risk) is an important step since it will give the context to the CTG reading and might also influence the contextual decision threshold for intervention. The CTG signal analysis can be structured into different steps : the evaluation of different CTG features followed by the diagnostic of the clinician.

1. **Baseline:**

   The Baseline is estimated in time segments of 10 minutes. It corresponds to the mean level

of the most horizontal and less oscilatory FHR segments. It is expressed in beats per seconds A normal baseline for the fetus will have a value in the range of 110 to 160bpm. If the baseline value is above 160bpm for more than 10 min then **Tachycardia** can be diagnosed whereas if it remains under 110bpm for more than 10 min we will be in the case of a **Bradycardia** [4].

2. **Variability :**
   Variability is "the oscillations in the FHR signal evaluated as the average bandwidth amplitude of the signal in 1-minute segments"[4].
   A normal variability is defined for a bandwith between 5 and 25bpm. If the variability has a bandwidth amplitude below 5bpm for more than 50min of for more than 3 min during deceleration (cfr deceleration section) the variability is defined as **reduced**. If it exceeds 25bpm for more than 30min then it is defined as increased variability also said saltatory pattern (associated with fetal hypoxia).

3. **Accelerations :**
   Accelerations are "increases in FHR above the baseline, of more than 15bpm in amplitude, and lasting more than 15 sec but less than 10 min" [4]. A signal with presence of accelerations will be defined as a "reactive" signal. It could also be said that before 32 weeks, the amplitude and duration of accelerations can be lower (10bpm - 10sec). Accelerations coincide with fetal movements and show a good behaviour of the fetus Autonomic Nervous System and so a sign of a responsive fetus without hypoxia or acidosis. Finally, it is also noted that accelerations do not often occur when the fetus is in deep sleep, knowing this the absence of accelerations is if uncertain significance if the CTG shows normal pattern otherwise. Therefore, it is often asked to the mother during non-stress tests to drink orange juice (blood glucose level increase) or to change position in order to wake the fetus up.

4. **Decelerations :**
   Decelerations are defined as "decreases in the FHR below the baseline, of more than 15 bpm in amplitude and lasting more than 15 seconds." There are different types of decelerations : early decelerations, variable decelerations, late decelerations and prolonged decelerations. All of them have different characteristics and potential diagnosis (more detailed explanation can be found in the FIGO guidelines [4] and in the figure 1.5).

CTG classification :

After the evaluation of the different features listed below, the clinician should classify it into 3 different classes : normal, suspicious or pathological. This overall impression is defined by how many features were reassuring or abnormal. The FIGO guidelines for classification can be seen in this figure 1.5.

An example of 2 different CTG signals from 2 different subjects can be seen in the figure 1.6. The first recording in the upper part of the signal can be classified as a normal recording since the baseline is around 140bpm, it shows a few accelerations and no big decelerations. The second

CTG classification

2015 revised FIGO guidelines on intrapartum fetal monitoring

	Normal	Suspicious	Pathological
Baseline	110-160 bpm	Lacking at least one characteristic of normality, but with no pathological features	< 100 bpm
Variability	5-25 bpm		Reduced variability. Increased variability. Sinusoidal pattern.
Decelerations	No repetitive* decelerations		Repetitive* late or prolonged decelerations for > 30 min (or > 20 min if reduced variability). Deceleration > 5 min
Interpretation	No hypoxia/acidosis	Low probability of hypoxia/acidosis	High probability of hypoxia/acidosis
Clinical management	No intervention necessary to improve fetal oxygenation state	Action to correct reversible causes if identified, close monitoring or adjunctive methods	Immediate action to correct reversible causes, adjunctive methods, or if this is not possible expedite delivery. In acute situations immediate delivery should be accomplished

*Decelerations are repetitive when associated with > 50% contractions.
Absence of accelerations in labour is of uncertain significance.

Figure 1.5: FIGO guidelines on intrapartum fetal monitoring for classification. Figure from FIGO guidelines if 2015 [4]

recording in the downer part of the figure shows a big deceleration that lasts more than 5 minutes showing potential phase of hypoxia and a baseline quite high (around 160 bpm). This recording will then be classified as a pathological CTG and the clinicians will follow more closely and make more tests to assess fetal well-being.

1.2.2. Limitations

Even if cardiotocagraphy can be an useful tool, it also shows also several limitations. One can see that the overall clinician impression for classification is subject to intra- and interobserver disagreement. [33]. The current use of CTG monitoring by clinician is based on a global overview of the CTG trace, accelerations and deceleration. A lot of misinterpretations continue to cause a significant amount of clinical issues such as perinatal deaths. Indeed, the CTG signal analysis is actually based on pattern recognition and morphological identification of ongoing decelerations classified as : early, variable, late and prolonged deceleration subjected to observer variability.

NHLSA's "10 years of Maternity Claims" [34] highlighted a number of CTG misinterpretations leading to stillbirth, Hypoxic-ischaemic encephalopathy (HIE) or cerebral palsy. The study states that in ten years covered by the study, 300 claims involving alleged CTG misinterpretation were reported to the NHSLA and in the 170 claims analysed, 148 presented CTG misinterpretation. A recent report from Marzbanrad in 2019 claims that current challenges of CTG include the lack of

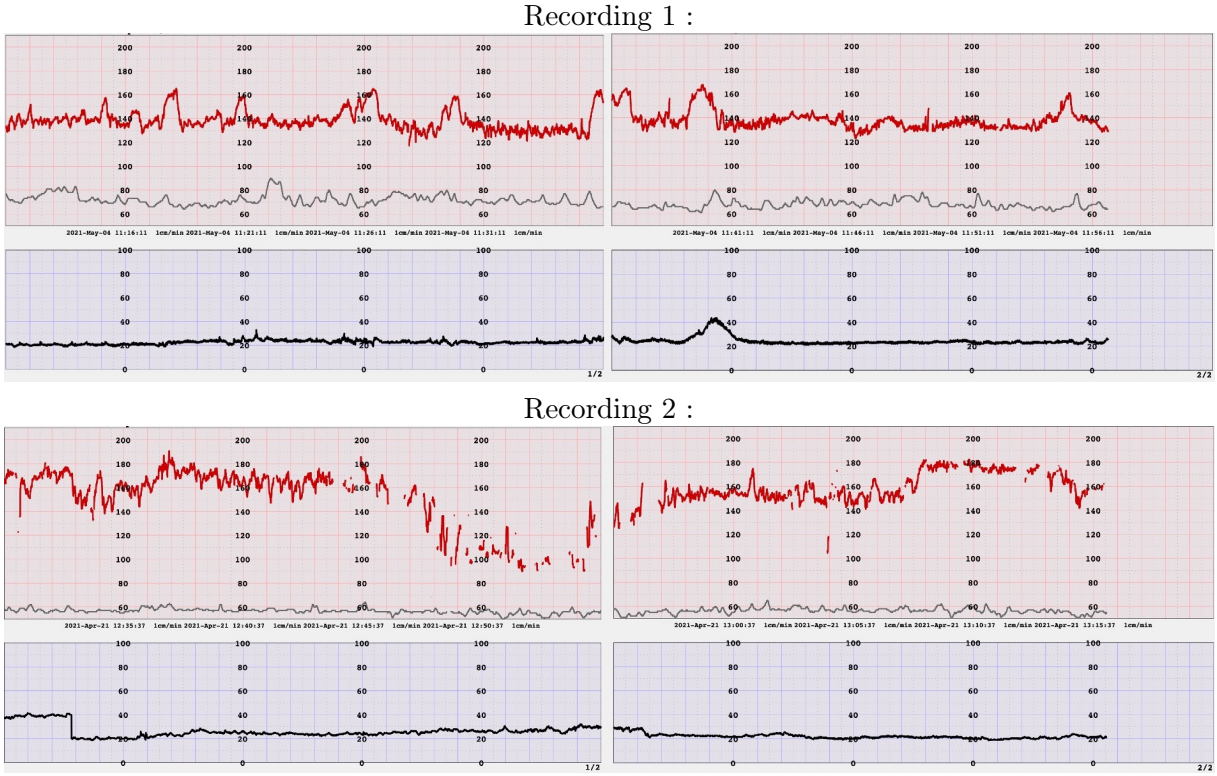


Figure 1.6: CTG output digitally recorded by Bloomlife. The Red line represents the FHR. The grey line shows the MHR and the black line in the under part is the TOCO signal. Every square is 1min of recording (EU convention). The Heart rate signals are expressed in bpm and the Toco in arbitrary unit. Recording 1 (upper) shows a normal recording whereas Recording 2 is pathological.

efficient signal quality metrics, insufficient signal processing for extraction of FHR and even more the lack of appropriate clinical decision support for CTG. [33]

It can also be remarked that the guidelines showed previously are intrapartum guidelines (during labour) whereas our work will focus on antepartum CTG. Even if the analysis is more or less similar, not any specific guidelines exist for antepartum recordings and the diagnosis is subject to the interpretation of the clinician. In the next section we will have an overview about the pathology of interest for this work : Intra-uterine Growth restriction. We will also see the current usual diagnostic process and its limitations.

2 | Intra Uterine Growth Restriction (IUGR)

In this chapter, Intra-Uterine Growth Restriction (IUGR) pathology will be studied. First of all, an overview about what is exactly the pathology, its causes and consequences will be made. Then, the process currently in practice for diagnostic will be explained and finally its limitations will be seen.

The chapter will be ended by an explanation of the role and the goal of the work for the IUGR detection. The approach and the framework of the project will be explained. The current state of the art will also be set. Finally, the point of view of clinicians about this project will be also reported.

2.1. Pathology

IUGR has been defined as "the rate of fetal growth that is below normal in light of thegrowth potential of a specific infant as per the race and gender of the fetus". It has also been described as "a deviation from or a reduction in an expected fetal growth pattern and is usually the result of innate reduced growth potential or because of multiple adverse effects on the fetus". [41]

An IUGR is a clinical definition and applies to neonates born with clinical features of malnutrition and in-utero growth retardation, irrespective of their birth weight percentile. Small for gestational age (SGA) can also be found in the literature. The difference between the 2 terms is that SGA determine neonates whose birth weight is less than the 10th percentile for that particular gestational age or two standard deviations below the population norms on the growth charts whereas IUGR applies to neonates born with clinical features of malnutrition and in-utero growth retardation, irrespective of their birth weight percentile. Using this definition, all IUGR infants will be SGA, but not all SGA infants will be IUGR.

2.1.1. Causes

IUGR reflects an abnormal adaptive fetal growth in a deleterious environment and affects 10–15 % of all pregnancies worldwide. IUGR may result from maternal, placental or fetal origin. Maternal

malnutrition before and during pregnancy represents the most prevalent non-genetic or placental cause.

According to a study of A. Wolmann [52] from the University Children's Hospital, Growth Research Center in Tübingen, Germany, the part responsible of the growth restriction between environmental, genetic and unknown factors is relatively equivalent. Among the etiologic factors, a third is due to genetic variables and two third due to environmental factors. A list of the different factors related to the IUGR conditions can be found in the table 2.1. A global overview of the different conditions is made.

Conditions associated with IUGR

Maternal	Fetal	Placental
Medical complications:	*Genetic:*	*Malformations:*
Hypertension, preeclampsia		
Severe chronic infections (infl. bowel disease, malaria, etc.)	Chromosomal abnormalities	Cardiovascular defects
Hypoxia (asthma, bronchiectasis,...)	Autosomal trisomies, monosomies, deletions	Gastrointestinal defects
Other severe diseases (diabetes, collagen disease, etc.)	Errors of metabolism (inborn)	Genitourinary defects
Uterus abnormalities		Skeletal dysplasias, etc.
Environmental factors	*Infections:*	*Metabolism, hormones:*
Smoking		
Alcohol	Viral (TORCH)	Growth Hormone variant
Drugs (antimetabolites, anticoagulants, anticonvulsants)	Bacterial (syphilis)	Placental lactogen
Narcotics	Protozoal (malaria, toxoplasma)	Insulin
High altitude		Steroids
Low socioeconomic status		Growth factors
Other conditions:	*Malformations:*	
Ethnicity:		
Prepregnancy weight, maternal height	Cardiovascular defects	
Pregnancy weight gain	Gastrointestinal defects	
Prior low-birth-weight infant	Genitourinary defects	
Low maternal age	Skeletal dysplasias, etc.	
Reproductive technologies		

Table 2.1: list of different conditions associated with IUGR from the study of Intrauterine Growth Restriction by *H A. Wollman.* [52]

As we can see, risk conditions can be maternal, fetal or even more specifically placental. On the maternal side, it can come from different medical complications. The most important one is hypertension. Indeed, severe, pregnancy-induced hypertension reduces birth weight by approximately 10% causing often IUGR fetuses. [28] Other diseases and infections can of course also have an impact on the growth of the fetus. Even if the cause can be from a long list of pathologies, in at least 40% of the IUGR, no underlying pathologies are found. [52]

Among the preventable, environmental causes of IUGR, smoking of the mother during pregnancy is by far the most important one, which is responsible for more than one third of all IUGR newborns. Alcohol and drugs also have an impact. Another important impact is the socio-economic status and the maternal nutrition.
Fetal and placental causes are quantitatively less important. Fetal infections is related to less than 10% of the IUGR cases. Genetic causes have also to be taken into account. Indeed, 40% of chromosomally abnormal infants have IUGR and the risk of an IUGR fetus having major congenital anomalies increases to 8%.

There is a strong correlation between fetal growth restriction and placental dysfunction. The growth restriction has an impact on the placenta on different levels: mechanically due to morphological effect but also chemically by the secretion and transport of hormones. Indeed, it regulates the transport between maternal and fetal circulation. [52] That's why any malformations such as Mosaïcism or chorioangiome will be a risk factor. Let's finally say that even if these conditions are often related to IUGR pregnancies, more or less 50% of the fetus in growth restriction do not present any maternal risk factor for the pregnancy.

2.1.2. Consequences

Fetal growth is the result of complex processes regulated by genetic factors and utero-placenta nutrition that can be affected in the case of a IUGR fetus. This will have consequences on the development of the fetus and will have neonatal and perinatal but also long-term consequences on the new-born. Indeed, IUGR increases significantly perinatal morbidity and mortality with survivors having a high likelihood of cardiovascular, metabolic and neurological disorders. IUGR affects the offspring even over several generations which is known as fetal programming.

Intrauterine Growth restriction pathology have an incidence of between 3% and 7% and is associated with 8 fold increased risk of stillbirth compared to non-IUGR [33]. Due to the fetus compression and non-optimal placenta nutrition and transport, IUGR fetus will be subjected to hypoxia and acidosis increasing morbidity and mortality. Let's also note that more 25 than % of stillborn are diagnosed as IUGR. It is why early detection of this pathology is critical to prevent perinatal morbidity and mortality.

Some neonatal complications are also caused such as a diminution of Apgar score and acidosis. Agpar score is a way to evaluate the health of a newborn through a five criteria evaluation based on : activity (tone), pulse, grimace, appearance, and respiration. A more detailed explanation of Agpar score can be found in the reference [35]. It also increases respiratory distress or increases ventricular hemorrhage. Those complications are function of the degree of prematurity and of the maternal pathology, But also metabolic and hematological disturbances, and disrupted thermoregulation.

Moreover, it is also well known that poor growth in utero presents relevantly more risks to develop cardiovasular pathology, type 2 diabetes when adults but also obesity, hypertension, dyslipidemia, and insulin resistance. Early onset growth delay and prematurity also increase the probability of neurological sequelae and motor or cognitive delay. [30] Since IUGR is also genetically caused, previous IUGR fetuses are more sucsceptible to have an IUGR pregnancy. So IUGR can continue from generation to generation.

Figure 2.1: neonatal, perinatal and long-term consequences of Intra Uterine Growth Restiction (IUGR).

2.2. Diagnostic

2.2.1. Currently in practice

Based on epidemiological definitions, the main measurable feature of IUGR is a low birth weight below the third percentile (standardized for gestational age and for gender) or below the tenth percentile (standardized for gestational age and for gender). Let's remind that IUGR is not synonymous with SGA. Indeed, a distinction has to be made between a normal growth with a small fetus due to the ethnicity, parity, sex and other parameters and an abnormal growth with increased risk of perinatal morbidity and mortality (IUGR). Indeed, normal SGA growth represents 70% of the cases whereas IUGR (30%) can be splitted into 2 groups : 15% due to infections and chrosomic anomaly and 15% due to vascular utero-placeta insufficiency (cfr subsection causes).

The diagnostic can be divided in different sections. The first one consists in identifying risk factors such as maternal medical and obstetrical antecedents, tabaco, drugs, alcool , socio-economic conditions,etc. These factors can be found in more details in the conditions associated with IUGR in the table 2.1 and more specifically in the maternal section. If a pregnancy is defined as a risky one, it will be followed closely by repeated clinical exams.

Currently, the diagnostic is made by the means of echo-imaging, estimating the fetal weight. In order to define the birthweight (BW), an ultrasound imaging parameter has to be chosen. One can know that it is impossible to measure precisely fetus weight in-utero. Parametric formula are then used to estimate the fetus weight with an error of between 8 and 15%. The most used is the one from *Hadlock et al.* [24] and is defined as :

$$log_{10}BW = 1.5622 - 0.01080HC + 0.04680AC + 0.171FL + 0.00034HC^2 - 0.003685AC.FL \tag{2.1}$$

Where BW is the birthweight estimation, HC is the cephalic perimeter, AC is the abdominal circomference and FL is the length of the femur. [24] Lets also note that the different formula are based on various parameters. The one with the best sensitivity for high risk population is the abdominal circumference and is therefore the one that is taken into account most of the time in practice.

Once the fetus weight is estimated, it is compared to a reference curve with respect to the gestational age. The threshold level varies with the pregnancy related conditions. For a "normal" pregnancy (without any risk factors detected), if the fetal weight is estimated as under the 10th percentile, IUGR will be suspected. If one of the biometric parameters is under P3, IUGR can be suspected. More generally, the fetus can be considered as IUGR for a pregnancy at risk if the weight is estimated as smaller than the 5th percentile and the 3rd percentile for pregnancies without any risk factors. An example of a reference curve for male and female fetuses are shown in the figure 2.2. Both the definition of the exact gestational age and the reference curve represent also a difficulty. Indeed, the reference curve might be subject to different parameters such as sex and ethnicity.

An important aspect is that 2 different types of growth restriction exist. The "type 1" also called "symmetrical" IUGR having an incidence of 20-25% and the "type 2" or assymetrical with 75-80 % of incidence. Type 1 affects the fetus earlier (before the 28th weeks) and are due to genetic causes, infection, etc. Type 2 happens when the fetal glycogen reserves are decreasing due to placental inefficiencies. Type 1 will affect both the weight and the birth length of the fetus whereas assymetrical have low birth weight and more or less normal birth length.

An illustration of the framework used by clinician to diagnose IUGR can be seen in the figure 2.3. In the next section, an overview on the different limitations arising from the current diagnosis process in place will be made.

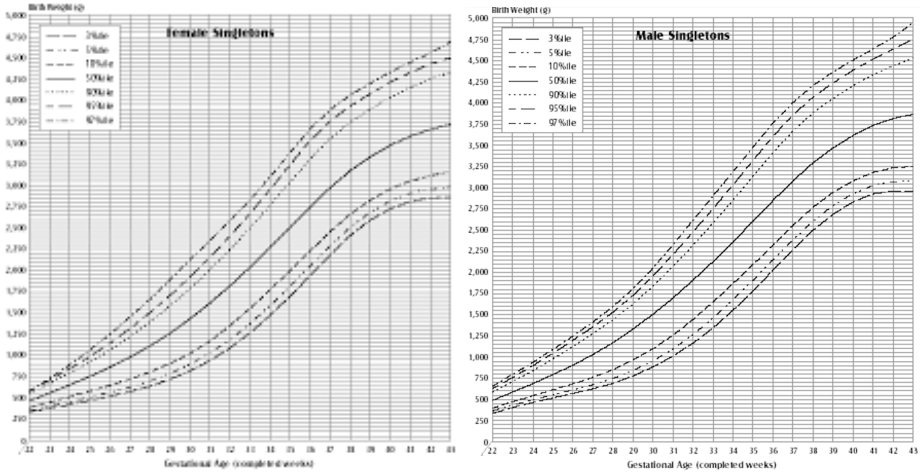


Figure 2.2: reference curves of Birth Weight (BW) [g] for GA in completed weeks of Canadian singletons. The left part of the figure shows female curves and the right part, the male curves. - Published by *Canadian perinatal system* [37]

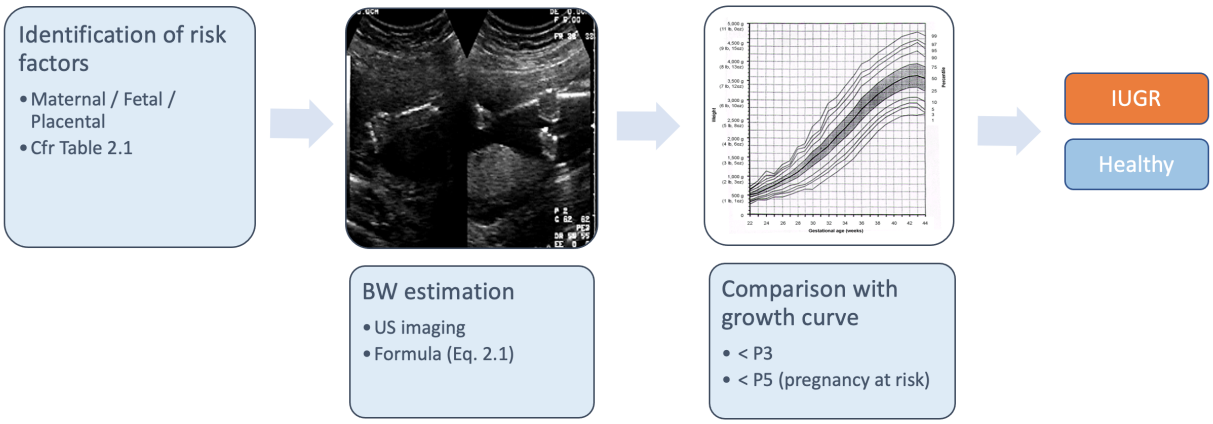


Figure 2.3: diagnostic framework currently followed by clinicians for IUGR detection.

2.2.2. Limitations

Now that we have seen how IUGR is currently diagnosed in practice, we will see its limitations.

First of all, the main problem remains in the distinction between a "normal" small growth (due to genetical factors such as ethnicity, sex, maternal and paternal height etc.) and abnormal growth leading to the increase of perinatal morbidity and mortality. Moreover, it is well known that there are babies with low birth weight and (nearly) normal birth length (asymmetric IUGR) as well as babies with a proportionate reduction of body weight and body length [52].

To assess a small for gestational age pregnancy, clinician has to compare with a normal intrauterine growth population. As we said previously, factors like altitude, racial characteristics, socioeconomic status and others have to be taken into account. This complicates seriously the diagnostic for the

clinicians and is not really taken into account in practice. Another inevitable difficulty is the assessment of the gestational age. The GA will have an impact on the analysis and could create a bias leading to over or underestimation of the reference weight (P10 , P5/3). To summarise, a baby classified as "normal" with an average birth weight would be classified as SGA (and vice-versa) on the basis of another standard showing high inter-observer variability.

In addition to the ultrasound estimation of weight, several measures might be done in the screening process. In addition to ultrasound biometry, tests include symphysis fundal height measurement, biophysical profile score, and multivessel Doppler studies to assess placentar circulation. Unfortunately, this is only the case in the context of a hospital with really good equipments. In the absence of sophisticated equipment in low-resource settings, IUGR detection is limited to the identification of maternal risk factors and the measurement of fundal height over time. [47] This measure is often not enough to diagnose really correctly IUGR and show really high inter-observer and intra-observer variability.

Current antenatal detection rates of IUGR are reported at 25 to 36%. Therefore, a preventative strategy to reduce stillbirths allows to improve the antenatal detection of fetal growth failure. Several studies have investigated other ways to detect IUGR. Some of them were based on the study of the fetal and placental cardiovascular system. Others were based on the heart rate variability. [47] [22] [43] In order to find a way to easily access the detection of growth restriction without need of high resources, we decided to focus on the fetal cardiac signal. Indeed, Fetal Heart Rate is one of the most low-cost and easy access source of information on the fetus well-being. Thus, CTG signals (cfr previous section on CTG signals) showing fetal heart rate through time will be used. This will also help us to assess the fetal wellbeing with Bloomlife device signals able to measure fetal heart rate aswell.

2.3. Focus of the work

In this section, we will explain the framework and the goal of our work. We will also have a look at the previous work already made in this field. Finally, the point of view of some clinicians about this project will be explained at the end of this chapter.

2.3.1. Role and Goal of the work

The work is focused on FHR signal analysis to give a prediction for IUGR subjects in antepartum. As we already explained, IUGR diagnostic is currently based on the estimation of the weight through echo-imaging metrics with parametric formula. Other tests can be made following it (such as a CTG monitoring or the echo-doppler) to assess the fetal well-being. Nevertheless, these tests can only be made in a context of well-equipped hospitals and clinics. When it's not the case, the follow-up is only based on the risk factors definition and the monitoring of the fundal height. Thus,

the idea of this work is to be able to acquire an indication with an easy and cheap signal to acquire (FHR) but also that Bloomlife device is able to obtain. Moreover, when a fetus is categorised as IUGR, a close monitoring is prescribed with recurrent CTG monitoring. So the goal of this work is to offer to clinician a metric helping in the analysis of the CTG signal.

As explain in the next sub-section, previous works tested parameters potentially interesting to use. The studies showed which parameters show difference between groups using statistical tests. On the other hand, the point of this work is to build a classifier using multiple parameters and able to support the growth restriction diagnosis for the clinician using only the FHR signal of the subject. In the next subsection, an overview of the previous work on distinction parameters is made and the parameters used for our algorithm will be explained more in detail in chapter 3.

Our algorithm should use raw FHR signals and give a classification index showing the probability that the fetus is IUGR or not. To do so, the first main step of this work will be to pre-process the raw FHR signals. This signals can present a lot of artefacts and be of bad quality. It is then important to avoid the algorithm to be influenced by this. With the pre-processed signal, a set of parameters characterising our signal will be built. Different types of parameters will be computed: variability (time-domain), frequency domain and complexity parameters. Finally, these parameters will be computed on FHR signals retrospectively annotated by clinicians ("IUGR" or "Healthy") and used as inputs to build and train a simple machine learning classification model predicting the state of the fetus.

2.3.2. State of the Art

In this subsection, an overview is made of the different works already published in this field that inspired us for this work. A lot of work on CTG signal processing exist but here the focus is made on those related to growth restriction. Most of the reports already done are signal processing and parameters computation of FHR signals to study the differences between IUGR and Heathy subjects.

In 2003, *M.G. Signorini et al.* published a paper on Linear and Non-linear parameters for the analysis of FHR signal from CTG [42]. It introduces frequency domain parameters as interesting for the CTG analysis and study it over 14 normal fetuses, 8 cases of gestational diabetes and 13 intrauterine growth retarded fetuses. The results showed that frequency parameters were able to separate normal to pathological fetuses constituting a first step to realize a new clinical classification system.

In 2009, *M. Ferrario et al.* suggested a indices for identification of IUGR using the computation of the Lempel-Ziv complexity (LZC) and Multiscale entropy (MSE) [23]. The results of the paper showed that these metrics could be useful to identify IUGR fetuses and separate them from physiological ones with a sensitivity of 77,8 %.

A new parameter based on Phase-Rectified Signal Average were proposed by *A. Fanelli et al.* in 2013 to quantitatively assess fetal well-being through CTG recording. Phase-rectified Signal

Averaged is a technique introduced by *Bauer et al.* in 2006 to study quasi-periodic oscillations in noisy and non stationary signals. [14] These parameters will be further explained in the next chapter. The parameters analysis was applied to 61 healthy and 61 IUGR fetuses signals acquired during non-stress tests. It showed that these parameters performed better than any other setting in this study in distinction of IUGR fetuses. This study were then used by *Signorini et al.* in 2014 [43] where they underlined its potential use with wearable technology. Let's also remark that this approach was also used by *T. Stampalija et al.* in 2015 [46] to compute acceleration and deceleration capacity. The study showed higher differences in very preterm than in preterm groups presenting differences emphasized in very preterm gestational age epochs.

Higher scale studies were published in 2017 by *Costa et al.* [13] and *Stoux et al.* [47] over 11687 fetuses from 25 to 40 weeks of pregnancy analysing short-term and long-term variability of Small for gestational age fetuses. It showed that SGA fetuses had lower long- and short-term variability with more pronounced differences between 28 and 35 weeks.

Finally, in 2019, a Dataset on linear and non-linear indices for discriminating healthy and IUGR fetuses were published by *Signorini et al.* [44]. This Open-source dataset is composed of 12 linear and non-linear features for 60 healthy and 60 IUGR subjects. They also worked on machine learning techniques and physiology based heart rate features using this dataset. They implemented a series of ML model and obtained as the best result a Random forest model with a classification accuracy of 85.5 % [45].

2.3.3. Point of view of clinicians

Since this work tends to help clinician in their diagnosis, it is relevant to know the medical point of view about this project. To do so, the study was presented to several doctors and professors that were then interviewed to give their comments.

According to *Pr Patrick Emonts*, professor of obstetric in the University of Liège and specialist in high risk pregnancies, the idea of the project is pertinent. "Indeed, the FHR is the principal parameter that the fetus is able to modify in order to adapt himself against constraints or difficulties of every kind". He also added that in contradiction with adults able to adapt their breathing frequency and thoracic cage amplitude, the fetus is not able to act on this adaptation ways and that it is then interesting to investigate the FHR modification to analyse adaptation that is first compensatory and progressively decompensated when the fetus is subjected to growth restriction.

Dr Sebastien Grandfils, also gave his point of view concerning this work. First of all, he found really interesting the idea and the methodology of the work. He confirmed that IUGR fetus has more mortality risks and it is relevant to diagnose them and then to induce the birth at the right moment in the following pregnancy monitoring. He explained that the first sign is the growth reduction and then diminution of movement due to nutriment deficiency. After that the metabolism tends to anaerobic. This induces acidosis that could lead to fetus heart to stop as a final result. Observations are often a modification of the cardiac rhythm with reduction of the number of accelerations and the

variability due to the decrease of sympathetic system activity. Finally, FHR shows also appearance of deceleration.

Moreover, he added that errors in diagnosis happen a lot due to a lack of specificity but also of sensitivity leading to IUGR misdiagnosed with an order of 15% of IUGR missed. These errors are due to the lack of precision of Ultrasound imaging and weight empirical formula. In addition to this, some physiologically small fetus (SGA) are often classified as IUGR whereas they are not pathological and that a big blur in the literature exists for this subject.

On the other side, he said that for early IUGR (less than 25 weeks), errors are more rare. The supplementary tool is more interesting later, but it could be relevant to use it to distinguish real IUGR from physiologically small fetus. However, the real concern would be after 32 weeks where the diagnosis accuracy is really not as good. This way, this tool could really help the clinicians in their diagnosis to be more sensible and specific. As a final remark, he also pointed that it would be of useful to differentiate IUGR from placental cause (Type 2) due to pathology or genetic causes (Type 1). Unfortunately, our dataset doesn't differentiate the two types of IUGR and it will then not be possible to separate them in the frame of this work. This could be potentially an interesting perspective and will be developed in the chapter 6.

Those comments globally show that this work could be interesting in practice. Indeed, it would allow clinicians to have an additional metric to increase their sensibility in IUGR detection. In the next chapters, we will start with features computation from FHR signals.

3 | Features implementation

In this chapter, we will implement algorithm to get the different features from the CTG signals. We will focus on the pre-processing of the CTG raw signals and then on the features implementation. We will divide the parameters into two categories :

- The "Standard parameters" that are the more common parameters sometimes used by the clinicians to analyse the signals.
- The "Non Standard parameters" that are parameters non specifically used in practice for CTG monitoring, but brings interesting information for the analysis and the classification.

In order to test the implementation extracting the parameters, a set of data given by Politecnico di Milano is used. A more detailed explanation of the dataset can be found in the Apendix A.

3.1. Pre-processing

First of all, raw CTG signals undergo a pre-processing step. Indeed, CTG signals are often affected by artefacts of any kind (clipping problems, movements, detection of the mHR instead of fHR, etc). It can also happen when the FHR signal is not correctly captured during periods of time. It is then important to pay attention to these issues before measuring parameters.

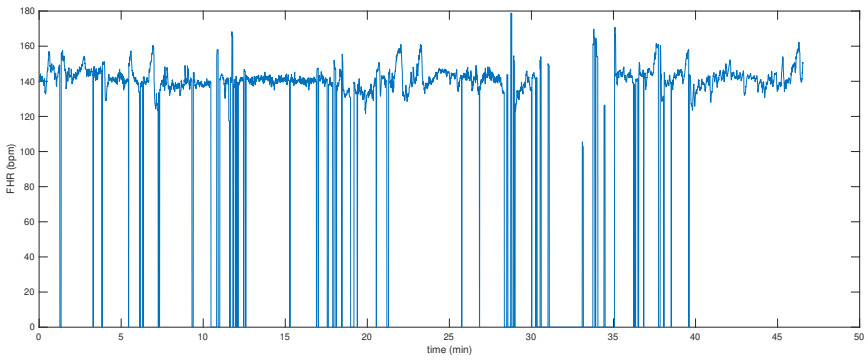


Figure 3.1: Example of a raw FHR signal [bpm]. In some cases, the signal goes to 0 (bad quality signal or signal not captured) or is subjected to artefacts.

In the dataset, the signal is expressed in 2 different ways (cfr Appendix A). The first one is the

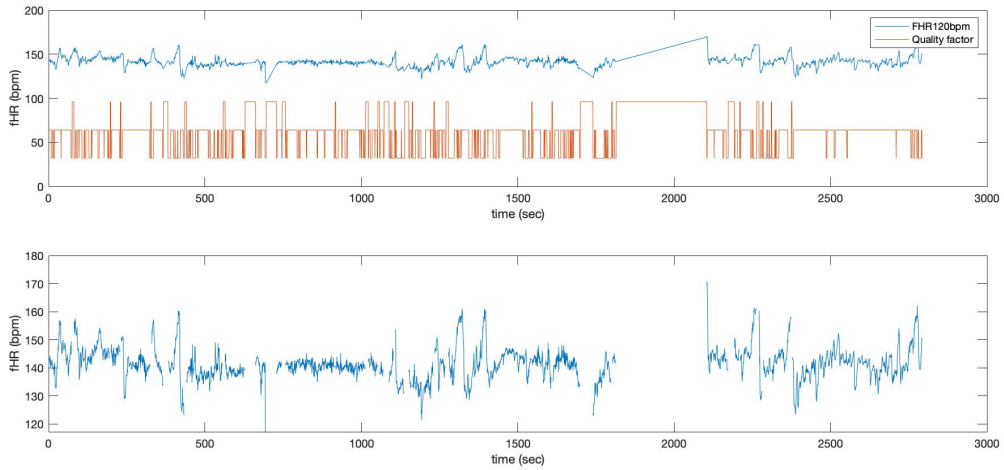


Figure 3.2: Removal of the bad quality segments of the signal: the upper figure shows the interpolated signal (FHR120bpm) [bpm] in blue and the quality factor in orange. The second one highlights the signal after removal of bad quality samples.

raw CTG data. In some parts of the signal, it can be seen that the FHR value goes to 0 (cfr figure 3.1). This means that the signal is not correctly acquired by the system and is then not trustable.

When measuring, the CTG associates to each sample of the FHR signals a specific quality index. The quality index is generally classified into 3 levels. In the case of the data used here, the signal is classified into the levels : 32, 64, 96. 32 is associated with a high-quality signal and 96 a bad quality signal. In practice, the recording only shows to the practician the part of the signal associated to the 2 highest levels of quality and discard the parts showing bad quality.

Another field in the dataset corresponds to the CTG data where the signals have been interpolated within time frames when the signal is missing, the interpolation is made taking a moving average taking the 5 samples before and after. Each sample is also associated with a quality factor. One can say that these interpolated parts cannot be trustable for the computation of some parameters and so should not be taken into account. In order to overcome this issue, a pre-processing algorithm is implemented to keep only the key signal. Let's note that different pre-processing steps are used depending on the parameters of interest. Indeed, some of them will need to use a signal pre-processed with only the good quality segments left whereas others use directly the raw signal. Let's also say that the processing step also depends on the length of the windows on which the signal is computed.

First of all, a pre-processing step is implemented to avoid taking into account bad quality signal in the computation of some of the parameters. A function called **FHRpreprocess.m** taking as input the raw FHR signal and giving as output the signal pre-processed is implemented. We decide to follow the approach often used in the literature such as *Signorini 2003* [42]. The preprocessing works as follow : first of all, the bad quality samples are erased (equals to NaN). Secondly, the

signal is sectioned in 3min segments ($180 * f_s$ samples). Then, each segments is analysed separately : If the segment in question contains more than 5 consecutive poor quality samples or if more than 5 % of all the samples are of bad quality, the subpart is not considered. Thus, the pre-processing function returns the signal fully preprocessed with NaN in bad quality samples and also with NaN vectors in the 3 minutes segments without a sufficient quality. The figure 3.3 illustrates the different steps of our pre-processing function.

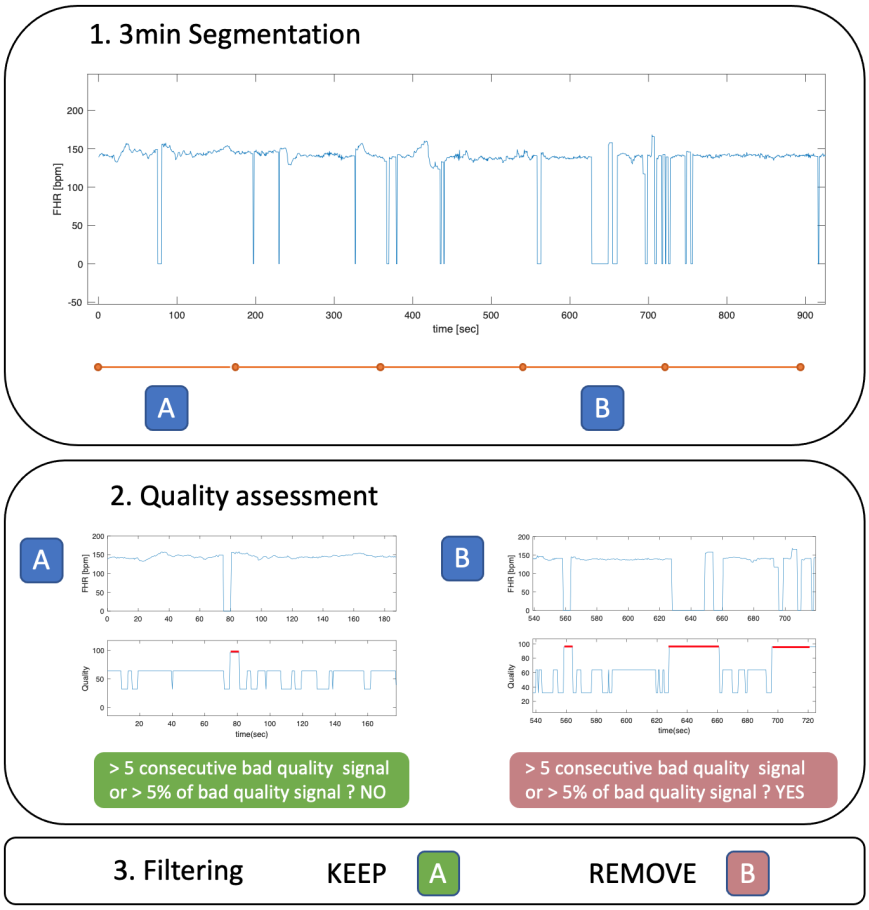


Figure 3.3: Illustration of the different steps of our pre-processing function for the computation of the parameters.1. The full signal is segmented in 3 min segments ($180 * f_s$ samples). 2. Each segment is analysed separately to check if it fills one of the 2 criteria: a) the segment contains a sequence of more than 5 consecutive bad quality samples. b) the segment contains more than 5 % of bad quality signal. 3. If the segment fill one of the 2 criteria, it is not taken into account and is replaced by a NaN vector.

Most of the parameters are not computed on all the duration of the CTG signals, but on smaller segments. The segmentation can be of 1 min or 3 min depending on the definition of the parameter. Parameters are then computed on all the segments and a global value is found by averaging all the values of signal sub-parts. To do so, functions are implemented to take as input the full signal and the sampling frequency associated and gives as output the signal segmented (functions

cutsignal1min and **cutsignal3min**). In addition to the segmentation of the signal, another pre-processing step is implemented. The pre-processing depends on the windows (and so the parameter concerned) as it will be further explained in the next section.

For the 1 min windowing, we erase (by equaling to NaN) all the bad quality samples of the signals (in this case, having a quality equals to 96).(cfr figure 3.2) After that, the filtering consists of keeping only the 1 min segments with at least a continuous good quality signals with a duration of minimum half of the segment (30 sec and 1min 30 sec for 3min). The ones that are not respecting the condition are removed from the computation of the parameters to only keep good quality segments of the signal. The same approach is made for the 3 min signals, but in this case the quality filtering is previously made in the pre-processing function *FHRpreprocess*. These 2 functions **cutsignal1min** and **cutsignal3min** give as output the signals segmented in the correct lentgh and with NaN vectors for the parts which don't have sufficient quality. The table 5.2 shows the pre-rocessing functions used to compute all the features. The features computation will be explained in detail in the next sections.

	FHRpreprocess	**cutsignal1min**	**cutsignal3min**
Baseline, Accelerations, Decelerations	X	X	X
STV, II, Delta	V	V	X
LTI	V	X	V
LF_pow, HF_pow, MF_pow, LF/(MF+HF)	V	X	V
ApEn, SampEn	V	X	X
LZC	X	X	X
AC, DC, AAC, ADC, APRS, DPRS	X	X	X

Table 3.1: Pre-processing functions used for the computation of the parameters.

Let's also say that the parameter computations don't take into account removed and bad parts of the signal (setted to NaN'). This approach helps us to discard noisy segments and avoid parts of the signal where the SNR is too low even for features while keeping a sufficient amount of signal to have a significant measure.

3.2. Standard parameters

In this section, we will study how to compute a set of standard parameters. These parameters are the most seen in practice. First of all, we will have a view about the classic parameters to which the clinicians are looking at when analysing the CTG : the baseline, number of accelerations and number of decelerations. After this, we will have a look at different metrics characterising the variability of the FHR: Short Term Variability (STV), Interval Index (II), Delta and Long Term

Irregularity (LTI). Finally, spectral parameters will also be analysed in this section.

A lot of standard features are actually time domain measures. That's why for some of the parameters such as STV, LTI or II, we use interbeat sequences instead of heart beat sequences converting bpm measures in ms. This is define as :

$$T(i) = \frac{60000}{S(i)}[ms]$$

with i = 1,..., N
T is the vector of interbeat sequences and S are the heart beat sequences [bpm].

Moreover, a lot of theories are based on signals sampled at 0.4Hz (every 2.5sec) and so the parameters are computed with a time series of 24 samples per minute. To have coherent measures, we define an undersampled time series by taking the average of consecutive FHR values of the signal :

$$T_{24}(i) = \frac{60000}{S_{24}(i)}[ms]$$

with i = 1,..., N$/(2.5 * fs)$

3.2.1. Baseline

As explained in chapter 1, the baseline consists of a running average of the heart rate. Physiscians construct this imaginary line to analyze accelerations and decelerations as deviation of the heartbeat from the baseline. When attempting to analyze the FHR automatically, the main problem is computation of the baseline against which all the other parameters are determined. The baseline is defined by the FIGO as "the mean level of the most horizontal, least oscillatory FHR segments" [15].

Multiple theories have been made in order to develop automatic analysis methods to define baseline of CTG tracings. The first attempt were made by *Dawes et al.* in 1982[20] A lot of other methods were published meanwhile such as the one from *Mantel et al.* that we will follow [32]. The algorithm is very complex, it is "constructed around two functional units, a digital filter and a trim function, which interact in an iterative process". More details about this algorithm can be found in the reference. [32]. For the baseline computation, we use the open-source toolbox from *Boudet et al.* [16], offering a re-implementation of a large amount of automatic analysis methods (AAMs) including one based on *Mantel et al* theory. (**aammantel.m**). In order to have a global unique parameter that our algorithm can use, we decide to compute the mean average of all the baseline curve. This gives us an indication of the "normal" heart rate of the foetus. An example of the baseline computation of a signal is shown in the figure 3.4.

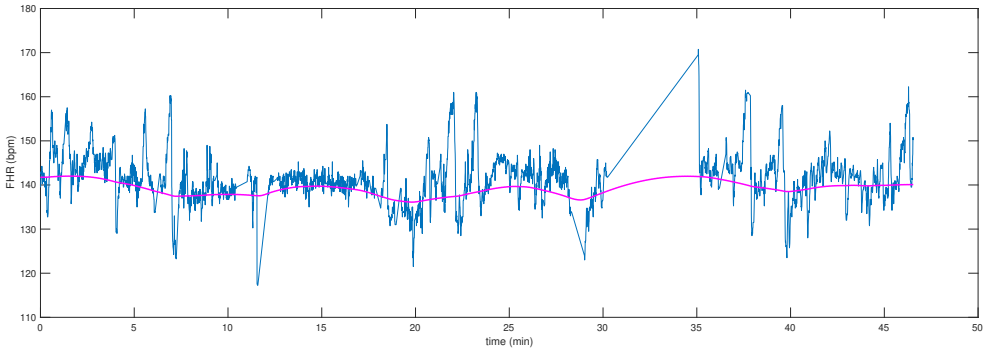


Figure 3.4: Example of baseline computation according to *Mantel et al.* theory. The baseline is represented in magenta whereas the FHR signal [bpm] (with interpolation in bad quality segments) is in blue. The mean value of the baseline is equal to 139.24 bpm.

3.2.2. Short Term Variability (STV)

Short Term Variabilty (STV) is a metric quantifying the variability over a short time scale. STV is evaluated over 1 min segments of the signal. Hence, to compute them we first use the function (**cutsignal1min.m**) previously implemented to cut the signal into 1 min segments. The data is then downsampled to get the T_{24}. Once this is done, following the definitions provided by *Arduini et al* [**?**], the STV for each 1 min segment is computed following this formula [43] :

$$STV = mean[|T_{24}(i+1) - T_{24}(i)|] = \frac{\sum_{i=1}^{23} |T_{24}(i+1) - T_{24}(i)|}{23} \quad i = 1, ...23 \tag{3.1}$$

A global value is then computed by averaging all the STV values of the 1 min signal segments to have a unique value characterising the whole signal (**computeSTV.m).**

3.2.3. Interval Index (II)

Interval Index (II) is another metric characterising the variability over short time scale. It is also computed over 1 min segments using undersampled interbeat signals T_{24}. Again using the definitions of *Arduini et al.* the II is defined over 1 min window as :

$$II = \frac{std[|T_{24}(i+1) - T_{24}(i)|]}{STV} \tag{3.2}$$

Again, the 1min values are averaged to get a unique II value for the signal.

3.2.4. Delta Index

Delta is simply the range of the signal in an interval of time. Here again, we use 1 min segments and undersampled interbeats T_{24} as :

$$Delta = max[T_{24}(i)] - min[T_{24}(i)] \ , \ i = 1, ...23 \tag{3.3}$$

The 1 min values are averaged to get a unique Delta value for the signal.

3.2.5. Long Term Irregularity (LTI)

The Long Term Irregularity (LTI) is a metric related to the variability of the signal. In this case, it quantifies irregularity over a longer time scale. LTI is often computed on 3 min signal segments. [23] Thus, we cut our signal into subparts of 3 min (using the **cut3minsignal.m** function). To compute LTI, we first compute interbeat T_{24} signal over 3 min (72 samples). LTI is then defined as the interquartile range of the modal $m_{24}(j)$ where :

$$m_{24}(j) = \sqrt{T_{24}^2(j+1) + T_{24}^2(j)}\ ,\ j = 1, ...71 \tag{3.4}$$

LTI is then the [0.25,0.75] range of the modal distribution. It is computed for each 3 min segment and then averaged to have a unique global value. To measure this feature, we implement the function **computeLTI.m** computing the LTI for each 3 min segment and the global LTI mean average value for the whole signal.

3.2.6. Power Spectral Analysis of fetal HRV

Another interesting set of features is the frequency parameters of the signal. It is well known in the literature that a relationship between the activity of neural cardiovascular control system and the frequency spectrum exists. [42] "Consistent link appears to exist between predominance of vagal or sympathetic activity and predominance of HF or LF oscillations, respectively : RR variability contains both of these rhythms and their relative powers appear to subserve a reciprocal relation like that commonly found in sympathovagal balance". [31].

Differently than for the spectral analysis of an adult that is generally divided in 3 contributions, we identify 4 different contributions in the case of the FHR spectral analysis :

- Very low frequency (VLF) [0-0.03Hz] : related to non-linear contributions and long period components of the signal.
- Low frequency (LF) [0.03-0.15Hz] : related to the sympathetic activity of ANS
- Movement frequency (MF) [0.15 - 0.5Hz] : depends on the fetal movement and maternal breathing. This spectral component is specific to the FHR spectrum analysis.
- High frequency (HF) [0.5 - 1Hz] : related to the fetal breathing

In our analysis, we will not consider the VLF part of the spectrum since literature has shown that it is not really relevant for this field. One can see that in order to catch all the spectrum of interest, the FHR signal should be sampled at least at a frequency of 2Hz with respect to the Nyquist theorem ($f_{nyq} = f_s/2$) cfr. Apendix A for the data.
After having computed the power spectral contributions in all the frequency ranges, we can com-

pute the $LF/(HF+MF)$ ratio. This metric corresponds to the LF/HF ratio computed for adults signal and quantify the autonomic balance between neural control mechanisms from different origins and is then related to the ANS activity (controlling the variability of heartbeats) [42].

In *Signorini et al.* [42] the computation of these parameters is made using an estimation of an autoregressive model (AR). They apply it on 3 min length FHR segments and then average the values of all the segments. Following the AR model theory, the signal is defined by the following formula [42] :

$$RR_{360}(n) = \sum_{i=1}^{p} a_i RR_{360}(n-i) + w_n \tag{3.5}$$

where w_n is defined as a white Gaussian noise : $w_n \sim WGN(0, \sigma^2)$, a_i are the model parameters and p is the model order. The parameter identification is made through estimation of autocorellation function and the optimal model order with respect to the Akaike criterion. The PSD of the autoregressive model is then defined as :

$$PSD(f) = \frac{\sigma^2 \Delta}{|1 - \sum_{k=1}^{p} a_k e^{-j2\pi k f \Delta}|^2} \tag{3.6}$$

More detailed information can be found in the paper of *Signorini et al, 2003* [42].

In the frame of this work, as it doesn't consist in the main part of the project, we decide to apply a non-parametric approach and implement it using direct estimation of the periodogram. PSD is applied on the 3 min windows with sufficient amount of quality signals following this formula :

$$PSD(f) = \frac{1}{NT} |DFT(x)|^2 \tag{3.7}$$

Where N is the number of samples (360 for 3min at 2Hz) and Direct Fourier Transform (DFT) is computed with the function *fft*.

Even-though this technique leads to a limited frequency resolution (limited by N, the number of samples) and is biased and non consistent, results are close to the one obtained by the AR model and good enough for our analysis. In figure 3.5 a scatter plot can be seen representing the values of all the 20 signals of our dataset and for each part of the spectrum. The x-axis shows the value from the dataset computed with the AR model and the y-axis represents the value computed with our direct approach. One can see that the values are following the reference line (equality) and are often centered around it with little differences between the x and y values.

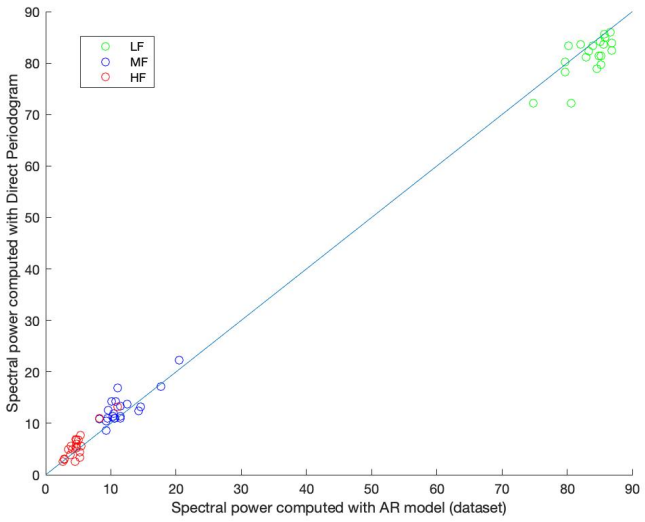


Figure 3.5: Scatter plot representing the PSD contributions of the signals from the dataset (x-axis) and computed with our direct approach (y-label). Green points represent Low Frequency contributions, Blue points represent Movement Frequency and Red ones High Frequency.

3.3. Non-Standard parameters

In addition to time and frequency domain parameters, we implement the computation of a set of non standard parameters. These features are non-linear features. We will first compute 2 different Entropy measurements. Then we will look at the Lempel-Ziv complexity and finally we will analyse the quasi-periodicities with Phase Rectified Signal Average analysis.

3.3.1. Entropy estimation: ApEn and SampEn

In this section, we analyze Multiscale Entropy (MSE) parameters and more specifically the Approximate Entropy and the Sample Entropy.

Approximate Entropy (ApEn) :
Approximate entropy was introduced in 1995 by Pincus SM [36]. ApEn is a statistic quantifying regularity and complexity. The idea is to analyse the regularity of patterns by comparing them to a specific pattern of length m. This with a specific tolerance r. Following the theory of Pincus et al [36], if we have N data points $\{u(i)\}$ and we define vector sequences representing m consecutive u values: $x_M(i) = [u(i), ..., u(i+m-1)]$and the distance $d[x(i), x(j)]$ between 2 vectors as the maximum difference in their scalar components. We can then compute :

$$C_i^m(r) = \frac{1}{N}\{j \leq n - m + 1 | d[x_m(i), x_m(j)] \leq r\} \tag{3.8}$$

For $i \leq N - m + 1$. $C_i^m(r)$ measures the number of j such that the distance is smaller than the

relative tolerance r. So it measures with a tolerance the regularity, or frequency, of patterns similar to a given pattern of a window length m. [36]. If we define now the function :

$$\Phi^m(r) = \frac{1}{(N-m+1)} \sum_{i=1}^{N-m+1} ln\ C_i^m(r) \tag{3.9}$$

We finally can define the Approximate Entropy as :

$$ApEn(m,r) = \Phi^{m+1} - \Phi^m \tag{3.10}$$

As explained by *Pincus et al.* "ApEn measures the likelihood that runs of patterns that are close to m observations remain close on next incremental comparisons". [36] The bigger probability to remain close, the smaller the ApEn values and conversely.

Following the results of *Ferrario et al.* [23] we set the window length to $m = 1$ and the tolerance $r = 0.1$.

Sample Entropy (SampEn) :
In 2000, Richman and Moorman introduced a new measure of entropy based on ApEn called the Sample Entropy. This new measure was established to overcome some limitations in consistency by removing the bias introduced by self-counts in the computation of ApEn. [39]. It has now largely been used in biomedical signal processing. Entropy parameters are calculated at different time scales in coarse-grained time series making them a basis of multiscale approach [43].

The idea is to find recurrent patterns at different time scales. It also uses 2 parameters m and r respectively the length of the specific comparison pattern and the relative tolerance (expressed as a percentage applied to the std), following the results of *Ferrario et al.* . The same values than previously are assigned to the window length : $m = 1$ and the tolerance: $r = 0.1$.

3.3.2. Lempel Ziv Complexity

Lempel and Ziv introduced this complexity measure in 1976. Lempel Ziv complexity (LZC) is defined as the minimum quantity of information needed to define a binary string[29]. LZC is used to quantify the rate of new patterns arising in a sequence of a binary value.

Since we are not working with strings of binary values but real time series (FHR signals), the first step to compute the LZC consists in converting signals into a symbolic string. To do so, several methods can be used. A first idea would be to use a method based on moving thresholds to code signal values by checking if x_n defined as the value of the signal at the sample n is smaller or larger than an average computed on a window of a specific length N. But this approach would lead conceptually to the same results than the Entropy estimation (cfr section 3.3.1) and would not bring any additional information.

Another idea introduced by J Szczepański in 2003 [49] for neural discharge analysis is to take into account changes in signal slope. This approach was then used by Ferrario to analyse FHR [23]. The coding is based on the sign of the slope of the signal (change of direction). Moreover, in order to avoid dependency on the quantization level and to limit the effects of noise, a factor p (expressed as a percentage of the current value) is introduced. If we consider the signal defined as a vector x, the encoding rule is defined as follow :

$$B_{n+1} = \begin{cases} 0, \text{ if } x_{n+1} \leq x_n + p * x_n \\ 1, \text{ if } x_{n+1} > x_n + p * x_n \end{cases} \tag{3.11}$$

Where B is the binary string on which the LZC is computed. Once we get the binary string of the signal, the open-source code from *Quang* implemented in 2012 based on Lempel Ziv paper (1976). [38] This code is used to compute the LZC of our binary string.

3.3.3. Phase Rectified Signal Average : AC, DC, AAC, ADC, APRS and DPRS

Phase Rectified Signal Analysis (PRSA) was presented by *Bauer et al.* in 2006. It is an efficient technique for the study of quasi-periodic oscillations in noisy, non-stationary signals. It allows the assessment of system dynamics without being influenced by phase reset and noise. [14] As explained in the section 1.2.1 increases and decreases of the FHR are controlled by the ANS in which the activity is directly correlated with the well-being of the fetus. The great advantage given by the PRSA curve is the fact that a long signal such as CTG recordings can be condensed in a single waveform, showing the average dynamic pattern[43]. It is for this reason that analysing the signal oscillations with PRSA could be interesting for assessing fetal well-being. In this subsection, we will first describe how the PRSA can be computed, we will see then the different parameters associated and how to get them.

PRSA:
Phase Rectified Signal Average is based on the definition of anchor points in the signal The anchor points help to align the signal oscillations, it is then followed by an averaging of the signal in a certain window around the anchor points.

As explained, the first step consists in the definition of anchor point. They are selected according to a certain condition on the signal properties. In our case, if we define a long time series (FHR) signal x_i, the conditions are the following :

$$\frac{1}{T}\sum_{j=0}^{T-1} x_{i+j} > \frac{1}{T}\sum_{j=1}^{T} x_{i-j} \tag{3.12}$$

in the case of increase events, or :

$$\frac{1}{T}\sum_{j=0}^{T-1} x_{i+j} < \frac{1}{T}\sum_{j=1}^{T} x_{i-j} \tag{3.13}$$

in the case of a decrease events. Let's remark that T actually sets an upper frequency limit for the periodicities that can be detected by PRSA [14].

Once anchors points are correctly defined, the next step is to define windows of length 2L centered around each anchor points. If i_v, v = 1,...M is the set of anchor points positions. We have for each anchor point a window composed of the following samples :

$$x_{i_v-L}, x_{i_v-L+1}, ..., x_{i_v}, ..., x_{i_v-L-2}, x_{i_v-L-1} \tag{3.14}$$

Where anchor points at the edge of the signal for which it is not possible to build a 2L length window are not taken into account. As explained in the paper, the parameter L must be larger than the period of the slowest oscillation that we want to detect.[14]

Finally, the last step consists to average all the signals of these anchor points windows. The PRSA curve is defined as :

$$PRSA(k) = \frac{1}{M}\sum_{v=1}^{M} x_{i_v+k} \tag{3.15}$$

for k = -L , ... 0, ..., L-1.
By computing this average, all the non-periodic components (such as noise, artefacts, etc) that are not synchronized with anchor points are canceled. This way, we only keep events that have a fixed phase relationship with anchor points. All the periodicities and quasi-periodicities are taken into account. An illustration from *Bauer et al.* [14] of the process of PRSA computation can be seen in the figure 3.6.

To compute the PRSA of our signal, we use the open-source implementation made by *M. Rivolta*. More details about the implementation can be found in the reference [40]. Following the results of *Fanelli et al*, we decide to use $L = 200$ and $T = 40$ samples to compute our PRSA. Those values showed the best performance in classifying IUGR and healthy subjects [22].
Now that we have computed the PRSA curve of our signal, it is useful to summarize the information of the later with global parameters. In the following, we will define some parameters that could help us in our classification.

Acceleration and Deceleration Capacity (AC and DC) :

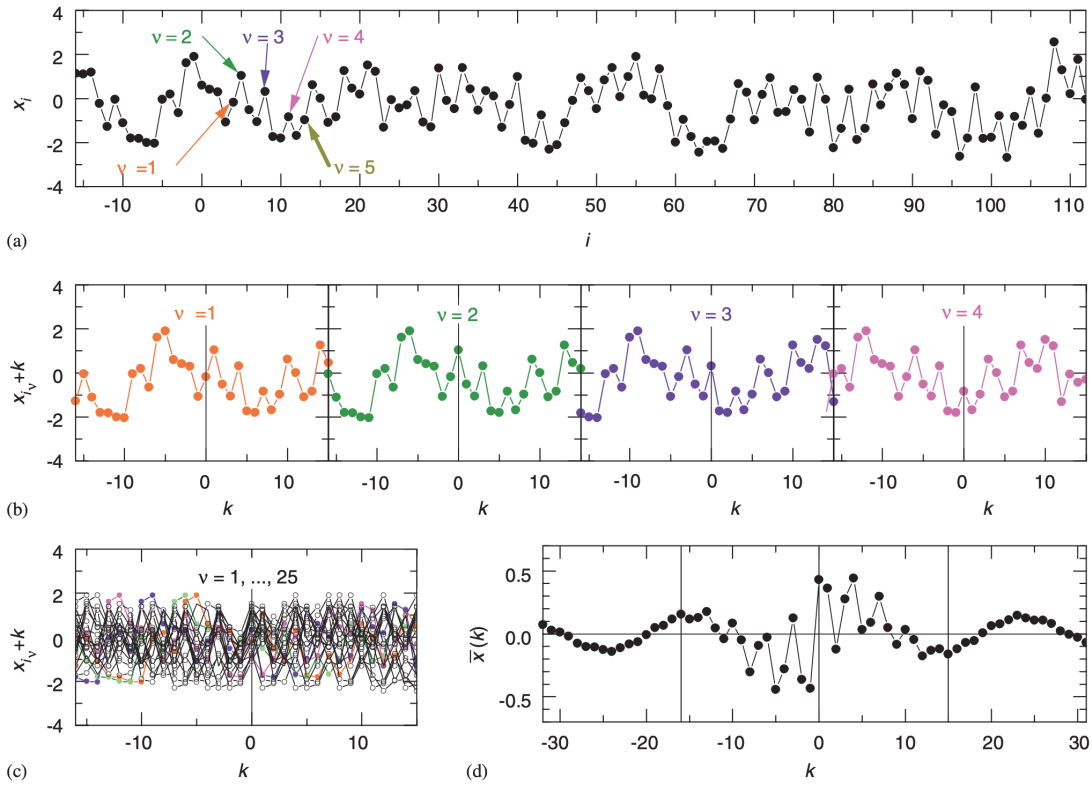


Figure 3.6: Illustration of the PRSA technique from *Bauer et al.* 2006 [14] : (a) Anchor points are selected from the original time series (x_i); here increase events are selected according to Eq. 3.12, corresponding to $T = 1$. (b) Windows (surroundings) of length 2L with $L = 16$ are defined around each anchor point; the points in each window are given by 3.14 and shown here for the first four anchor points. (c) The surroundings of many anchor points (all located in the centre) are shown on top of each other. (d) The PRSA curve $x(k)$ resulting from averaging over all surroundings is shown versus the offset k from the anchor points; the parameter L is increased to L = 32 in order to improve the visibility of the slow period.

Bauer et al. [14] defined the acceleration and deceleration capacity parameter (respectively AC and DC) as a first metric to characterize PRSA curve. If we consider $X(0)$ as the sample corresponding to the anchor point we have :

$$AC(DC) = \frac{\sum_{i=0}^{s-1} X(i) - \sum_{i=-s}^{-1} X(i)}{2s} \tag{3.16}$$

Where s is a scale factor. Following the paper of *Fanelli et al.* $s = 2$ is used. [22]

Average Acceleration and Deceleration Capacity (AAC and ADC) :

This parameter was used by *Huhn et al* [26], it is similar to AC but corresponds to the integral

measure of all periodic acceleration-related oscillations. [22] It is defined as :

$$AAC(T) = \frac{1}{2T}[\sum_{i=0}^{T-1} X(i) - \sum_{i=-T}^{-1} X(i)] \tag{3.17}$$

One can see that the AAC is in fact the AC with $s{=}T$. Here, $T = 40$ is used.

Acceleration and Deceleration phase rectified slope (APRS and DPRS) :

This parameter was introduced by *Fanelli et al* in 2013. It is defined as the slope of the PRSA curve computed in the anchor point. It is based on the fact that the diagnostic information is contained in the number and the temporal characteristic of increases and decreases in heart rate. We have :

$$APRS(DPRS) = \frac{\delta X(i)}{\delta i}|_{i_{AP}} \tag{3.18}$$

Where APRS is for increase events and DPRS decrease events. It describes both the increase (or decrease in amplitude) and its time length.

An example of PRSA curve computed on one of the FHR signals of the dataset is shown in the figure 3.7. In this curve, APRS is defined as the slope of the PRSA curve at the anchor point (sample 200). In the same way, we can compute DPRS value of the PRSA curve for decreasing events.

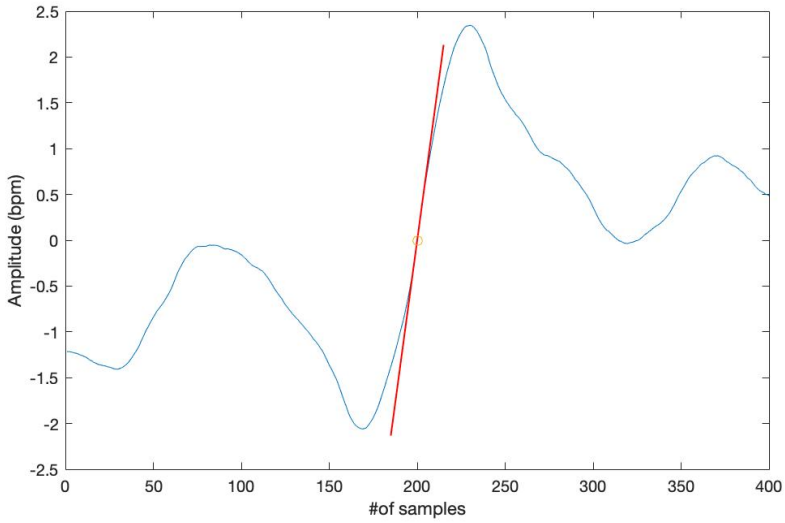


Figure 3.7: Phase Rectified Signal Average (PRSA) curve computed on a FHR recording (blue). The Acceleration Phase Rectified Slope is shown in red and the anchor point in orange. APRS is defined as the slope of the PRSA curve at the anchor point.

3.4. Features values and distribution in the dataset

In order to test our implementations, the dataset of Politecnico di Milano is used. As a final result, our algorithm takes as input the raw FHR signal and compute our set of parameters after adapted pre-processing steps. Since a lot of "machine" parameters are available for each subject, most of our parameters measurement were compared with the machine feature values. This helps us to have a first check on our implementation.

For each subject, a set of parameters is computed and saved. As a final result, a table is created with for each patient : the state (IUGR / Healthy) , the Gestational age (GA) and the set of parameters computed on the signal characterising each patient. An example can be seen in the table 3.2.

State	GA	baseline	baseline std	DELTA	II	STV	LTI	LF	MF	HF	LF/(HF+MF)
'Healthy'	37	139.24	1.5695	43.023	0.7893	4.94	18,77	85.93	8.79	5.27	6.0627
ApEn(1,0.1)	SampEn(1,0.1)	LZC(2,0)	AC	DC	AAC	ADC	APRS	DPRS			
1.4937	1.2835	0.8691	0.1467	-0.1599	1.5468	-1.6858	0.1455	-0.1576			

Table 3.2: Exemple of the table data obtained for one subject. The first column represents the state of the fetus (annotated retrospectively), the second one the Gestational Age (GA) when the CTG recording was made. The following columns (3-21) are the parameters computed by the algorithms over the raw FHR signal.

Once the parameters computation algorithms are fully implemented, we start to look at the results on our small dataset of 20 subjects (10 "healthy" and 10 "IUGR"). Boxplots is made for each parameter to compare the distribution of the parameter values over the 2 groups of interest.

First of all, mean baseline values and baseline standard deviation were analysed between the 2 groups. Boxplots found in figure 3.8 show that there is not really a lot of differences between the 2 groups for the baseline. The standard deviation highlights slightly higher values for the IUGR group but not with a significant difference.

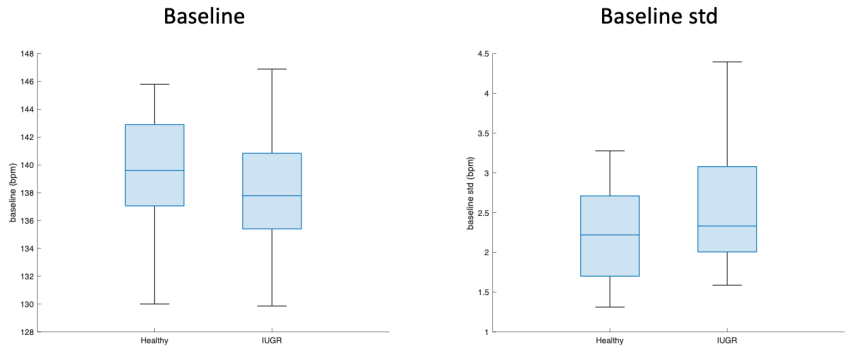


Figure 3.8: Boxplot of the parameters distribution over our 2 state groups. Left : the baseline [bpm] , Right : the baseline standard deviation [bpm]

Time domain variability parameters are then analysed. As we can see, the distributions show in general higher values for IUGR groups than for Healthy groups. Indeed, IUGR distribution has higher values of Delta, STV and LTI whereas II shows a wider distribution for IUGR subjects. IUGR subjects show STV median values around 7ms (compared to 5.5ms for Healthy) and LTI median values around 30ms (compared to 18ms for Healthy). These results are interesting because they show different tendency than what is normally found in the literature. Moreover, the parameters values were compared with the machine data so the difference should not come from our measuring algorithms. Of course those results are only on a population of 20 subjects and the values could depend on the Gestational Age (cfr section 4.1). That is why any conclusion cannot be made at this step.

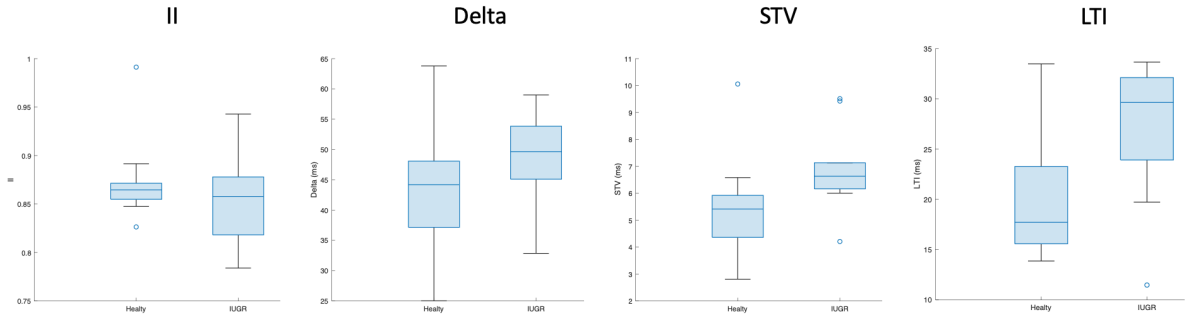


Figure 3.9: Boxplot of the time variability parameter distribution over our 2 state groups of: 1. Interval Index (II) 2. Delta (ms) 3. Short Term Variability (STV) (ms) 4. Long Term Irregulatity (LTI) (ms)

The analysis is followed by frequency parameters showed in the figure 3.10. We see that there is no significant difference again and that IUGR subjects look to have values distributed over a wider range. Still, one can see that the median value of the ratio $LF/(MF+HF)$ is around 4.5 for IUGR

whereas Healthy subjects have a median of 5.8.

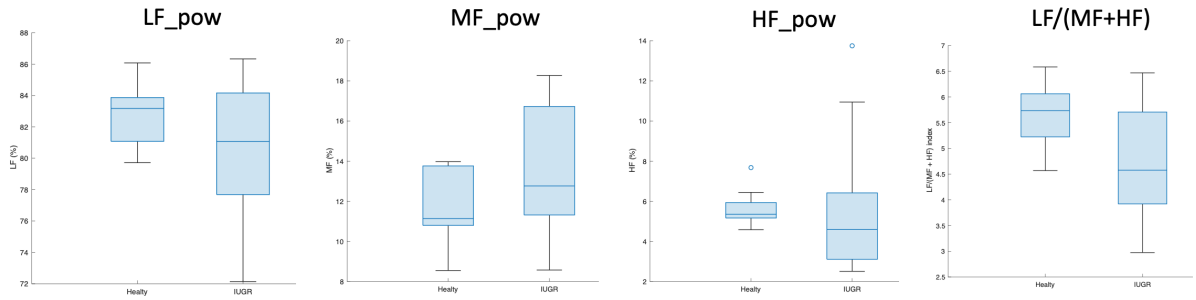


Figure 3.10: Boxplot of the frequency spectrum parameter distribution over our 2 state groups of : 1. Low Frequency power (LF_pow) (%) 2. Movement Frequency power (MF_pow) (%) 3. High Frequency power (HF_pow) (%) 4. LF/(MF+HF) ratio

For the complexity parameters, Lempel Ziv Complexity and Sample Entropy do not really show any differences whereas Aproximate Entropy shows lower values for IUGR but without any significant differences. Boxplots are shown in the figure 3.11.

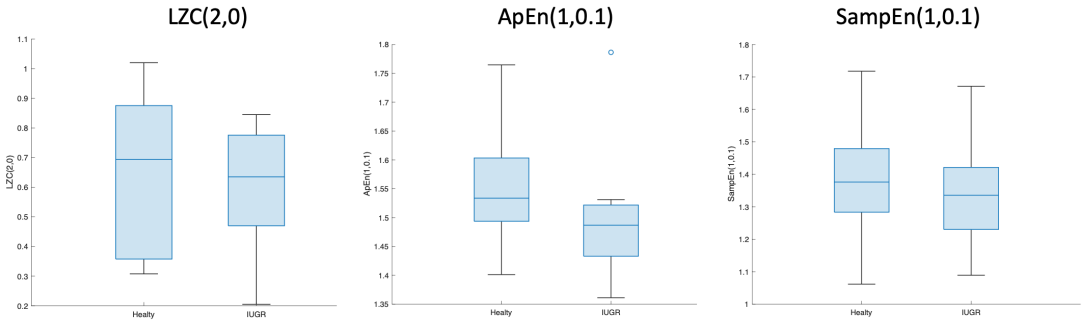


Figure 3.11: Boxplot of the compexity parameter distribution over our 2 state groups of : 1. LZC(2,0) 2. ApEn(1,0.1) 3. SampEn(1,0.1)

Finally, Phase Rectified Signal Average features distribution can be seen in the figure 3.12. Globally, one can see that parameters related to acceleration show distribution with higher values for IUGR and lower for deceleration parameters except ADC. We also see that deceleration related parameters show better distinction in their distribution (except for ADC). IUGR subjects tend to have lower (higher in amplitude) values for deceleration parameters than the Healthy population. In the same way, IUGR population looks to have in general higher values of acceleration parameters. This is relatively coherent with the fact that IUGR population shows more variability by having higher STV and LTI for instance.

Finally, one can say that the dataset used is really reduced in size and potentially subjected to

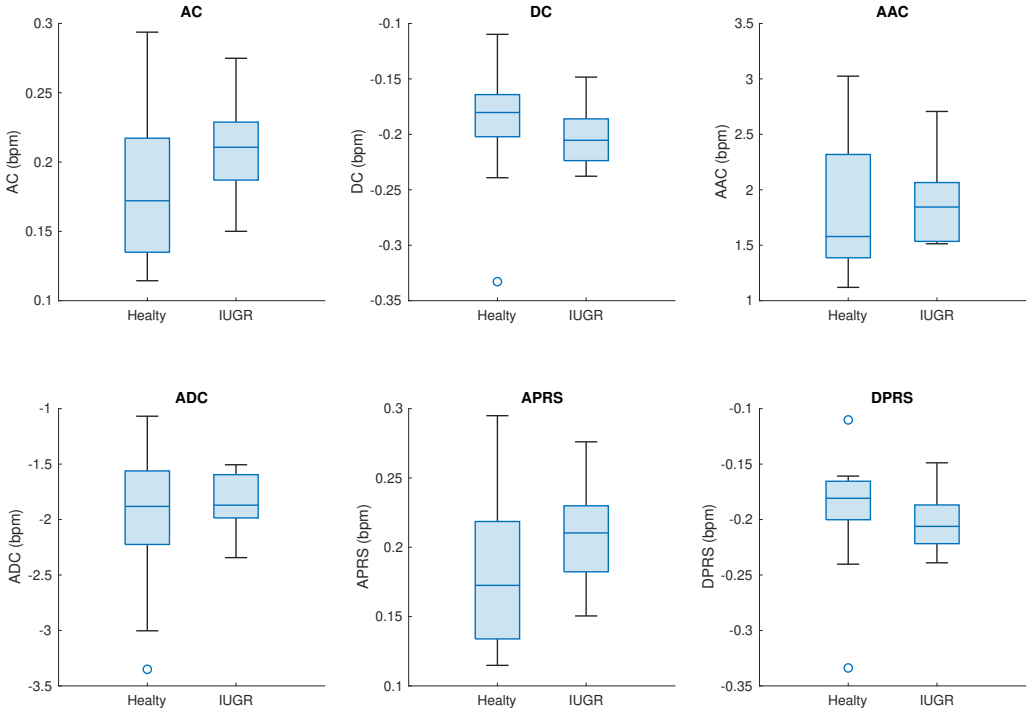


Figure 3.12: Distribution of the Phase Rectified Signal Average parameters of Polimi dataset for the 2 state groups : 1. Acceleration capacity (AC) [bpm] 2. Deceleration Capacity (DC) [bpm] 3. Average Acceleration Capacity (AAC) [bpm] 4. Average Deceleration Capacity (ADC) [bpm] 5. Acceleration Phase Rectified Slope (APRS) [bpm] 6. Deceleration Phase Rectified Slope (DPRS) [bpm]

noise, artefacts etc. Thus, it is complicated to make any conclusion on these results due to the reduced size of the dataset.

4 | Feature analysis

Now that we have implemented our computation algorithms to assess our set of parameters, a feature analysis will be made. Different datasets will be used in this chapter : the one given by Politecnico di Milano, the open-source dataset [44] and finally data from Bloomlife. The main characteristics are shown in the table 4.1 and more details can be found in the Appendix A.

	Polimi	**Bloomlife**	**Open-source**
Number of subjects	20	113	120
IUGR	10	12*	60
Healthy	10	101	60
Measurement System	Hewlett Packard	Avalon FM30	Hewlett Packard
f_s	2 Hz	4 Hz	2 Hz
FHR signal access	Full	Full	No
Machine measurements	Partially	No	Yes

* SGA, not retrospectively annotated IUGR

Table 4.1: Main characteristics of the datasets

Let's remind that our computation algorithms developed on Polimi dataset. The signals were acquired by a Hewlett Packard CTG. This dataset gives us full access to the FHR signal and some additional information about machine parameter measurements such as variability and spectral parameters. This helped us to construct our algorithm. Unfortunately this dataset contains only 20 subjects.

In addition to this, access to Bloomlife pilot data allows to get 113 more recordings. Unfortunately, this data is not retrospectively annotated by clinician to diagnose IUGR. The only information we have is if the fetus is considered as small for his gestational age or not. Let's also say that in this case the CTG machine was an Avalon FM30 and the sampling frequency of the signals is 4Hz.

Finally, we will use the parameters of the Open-source dataset published in 2020 by *Signorini et al.* giving access to a set of 12 linear and non-linear indices extracted from Fetal Heart Rate (FHR) traces acquired through Hewlett Packard CTG. Unfortunately, no access to the raw FHR signals is given. Thus, our algorithm cannot be used on this data. The populations consist of 60 healthy

and 60 IUGR fetuses retrospectively annotated by clinicians.

In this chapter, these datasets will be used together to analyse our features. First of all, the potential dependency of our parameters to the gestational age will be studied to avoid any bias in our data input of our future algorithm due to when the recording is made. After this, a comparison between the distribution will be made in order to see if some of the parameters are not influenced by the measurement system used. Finally, a dataset selection will be made for the input of our prediction algorithm.

4.1. Parameter dependence on Gestational age (GA)

In this section, we will study the potential impact on our different parameters that could have the Gestational Age when the recording is made. Indeed, it could be logical that some parameters such as the variability change during the pregnancy along the development of the fetus. For instance, a study made in 2017 by *C. Amorim-Costa* shows that the GA has an impact on the variability and more specifically on the STV and the LTI. The study demonstrates that over 11687 recordings "similar trends throughout gestation occurred : decrease in baseline, and increase in long- and short-term variability" [13]. Since we don't want the GA to bias our predictions, we decide to study the dependence of each parameter to the gestational age. The result of the algorithm should not be influenced by the moment in the pregnancy in which the recording is made. Meanwhile, if a parameter is dependent on the gestational age, it's value should not be taken into account in the same way at 28 weeks or 38 weeks. That's why if the GA of the recordings shows dependence effects on a certain parameter it should be rectified in order to have an algorithm equally performing over the spectrum of GA.

To do so the correlation of each set of parameters with the GA will be studied to see the ones showing dependency. For this subset of parameters a modelisation of a linear regression characterising the distribution over the GA will be done. The value will be adjusted with respect to the linear regression computed. After that, data will be checked to see if the GA dependency of the adjusted parameters is been correctly removed.

Let's remark that this analysis is only made for parameters available in the Open-source dataset. Indeed, the amount of annotated data using only Bloomlife and Polimi datasets would be too small and therefore the dependency would not be really relevant. Hence, the following parameters will not be analysed in this part : **Baseline, SampEn, AC, DC, AAC, ADC**.

4.1.1. Correlation between Gestational Age and Parameters values

The first step of analysis consists to determine which features show correlation with the GA. To do so, we decide to use the Spearman's correlation to assess the monotonicity of the relation between the GA and the parameter. Spearman's rank coefficient is a non-parametric measure of rank corre-

lation. It measures the strength and direction of the variables relationship using a non-parametric correlation statistic. This way, it assesses how well the relationship between 2 sets of values can be described with a monotonic function.

Spearman's correlation is equivalent to calculate the Pearson correlation coefficient on the ranked data. For a set of size n, the n values X_i , Y_i are converted to ranks $R(X_i), R(Y_i)$ and the spearman's correlation coefficient r_s is computed as :

$$r_s = \frac{cov(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} \tag{4.1}$$

Where the formula of the Pearson correlation is used with the ranked variables. $cov(R(X), R(Y))$ is the covariance of the rank variables, $\sigma_{R(X)}$and $\sigma_{R(Y)}$ are the standard deviations of the rank variables. [11]

Spearman's correlation coefficient is computed for each parameter using the 120 subjects (60 healthy // 60 IUGR) of the Open-source dataset. The results are shown in the table 4.2.

Spearman's Correlation coefficient r_s

	Overall	Healthy	IUGR
Delta	0.3344	-0.1177	0.2798
II	0.0291	0.0502	-0.0474
STV	0.3903	-0.1204	0.4231
LTI	0.2577	0.1196	0.1196
LF	0.13002	-0.03904	0.0075
MF	-0.0119	0.0435	0.1192
HF	-0.1911	0.0167	-0.1002
LF/(MF+HF)	0.15502	-0.0869	-0.0476
ApEn	0.2241	0.0352	0.1729
LZC	0.1773	-0.2348	0.0961
APRS	0.4189	0.0189	0.3356
DPRS	-0.4862	-0.0446	-0.4185

Table 4.2: Spearman's Correlation coefficient r_s between GA and parameters of interest computed with the Open-source dataset. The first column represents the correlation on the overall population, the second with only Healthy subjects and the third one with only the IUGR subjects. Pink cells are the ones showing a clear dependency with the GA whereas orange ones are parameters presenting a moderate dependency to investigate.

In the table, the values are significantly different for the 2 sub-populations whereas the study from *Costa et. al* over a large population showed that both SGA (and so IUGR) and normal population have similar trends.
One can think that these different results are due to reduced size of our dataset. Another really

important point is that the healthy population recording are distributed over only 2 gestation weeks (34 and 35) whereas the IUGR population has a larger spectrum (from 28 to 39) of gestation weeks in the dataset. The GA distribution in the Open-source dataset is shown in the figure. 4.1 on the left. It can be explained by the fact that in Healthy pregnancies, only one CTG monitoring is made around 34, 35 weeks whereas IUGR pregnancies are usually at risk so CTG are prescribed earlier and will be part of the deeper follow-up to assess the fetus well-being.

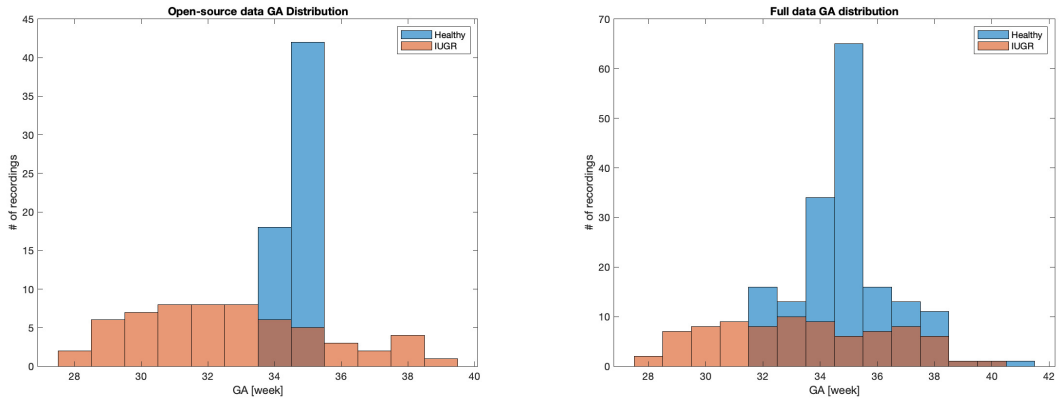


Figure 4.1: Histogram of the distribution of the Gestational age [week] in our dataset. The distribution for Healthy subjects is represented in blue and IUGR in orange. The left figure shows the distribution only for *Open-source* dataset. The right figure highlists the distribution for all the datasets.

To overcome this issue, we decide to study the GA dependency over an extended dataset including additional data from **Politecnico di Milano** and **Bloomlife**. The new GA distribution of extended data set is shown on the figure 4.1 on the right. Since CTG monitoring is usually not prescribed before 32 weeks in a non-risk pregnancy, healthy data is now distributed from the 32nd week to 38th. It allows us to have a better estimation of the dependency also for the healthy subjects and have a more robust regression adjustement. The table 4.3 shows the Spearman's correlation r_s of the extended dataset with the p value associated. Let's remind that correlation with a $p \leq 0.05$ presents a strong evidence against the null hypothesis which is here the independence over the GA and are therefore dependent parameters. [3]

As seen in the table, several parameters show p value smaller than 0.05 and are therefore dependent to gestational age : the variability parameters **Delta**, **STV**, **LTI** but also the non-linear parameters **ApEn**, **APRS**, **DPRS**. In the following section, regression for each parameter will be studied followed by an adjustment in order to remove its dependency to the gestational age.

4.1.2. Adjustment by linear regression

Now that we know which of our parameters are dependent to the GA, we try to find a linear model that is fitting the GA/parameter relationship. After several tries and checks, a robust linear model is chosen. Robust regression is an alternative to least squares regression when data are contami-

	r_s	p
Delta	0.26683	2.22587e-05
II	-0.06051	0.34458
STV	0.25051	7.10433e-05
LTI	0.15312	0.01538
LF	-0.08805	0.16512
MF	0.11539	0.06853
HF	0.04164	0.51221
LF/(MF+HF)	-0.05459	0.39012
ApEn	0.19024	0.00268
LZC	-0.09511	0.13213
APRS	0.25169	5.32278e-05
DPRS	-0.28385	4.69160e-06

Table 4.3: Spearman's correlation coefficient of the extended dataset, r_s is the correlation coefficient and p the probability that the null hypothesis (parameters independent from GA) is true. Pink cells show $p < 0.05$ and so dependency between the parameter and GA.

nated with outliers or influential observations. Least squares estimates for regression models are highly sensitive to outliers pulling the least square fit too far in their direction. It is current to find outliers in biomedical signals and even more in the parameters we will use as seen in the previous chapter. By using a normal least-square fitting, they would receive too much weight compared to non-outlier data leading to distorted estimates of the regression coefficients. Robust regression down-weights the influence of outliers, therefore we decide to use this technique to compute a less outlier-sensitive regression modeling our GA dependency. [2]

Robust fitting can be computed with different weight functions. The matlab implementation is used here (**robustfit.m**) with the default "Bisquare"' fitting function defined as :

$$w = (|r| < 1).*(1 - r^2).^2 \tag{4.2}$$

where

$$r = \frac{resid.}{tune * s * \sqrt{1-h}} \tag{4.3}$$

and the tuning constant is equal to $tune = 4,685$, $resid$ is the vector of residuals from the previous iteration, h is the vector of leverage values from a least-squares fit and s is an estimation of the standard deviation of the error term given by $s = MAD/0.6745$. MAD is the median absolute deviation of the residuals from their median. The constant 0.6745 makes the estimate unbiased for the normal distribution.

The robust linear regression was computed for the 6 parameters showing a dependency over the GA (cfr previous section). The distribution of the parameters with respect to the GA and the robust regression computed can be seen in figure 4.2.

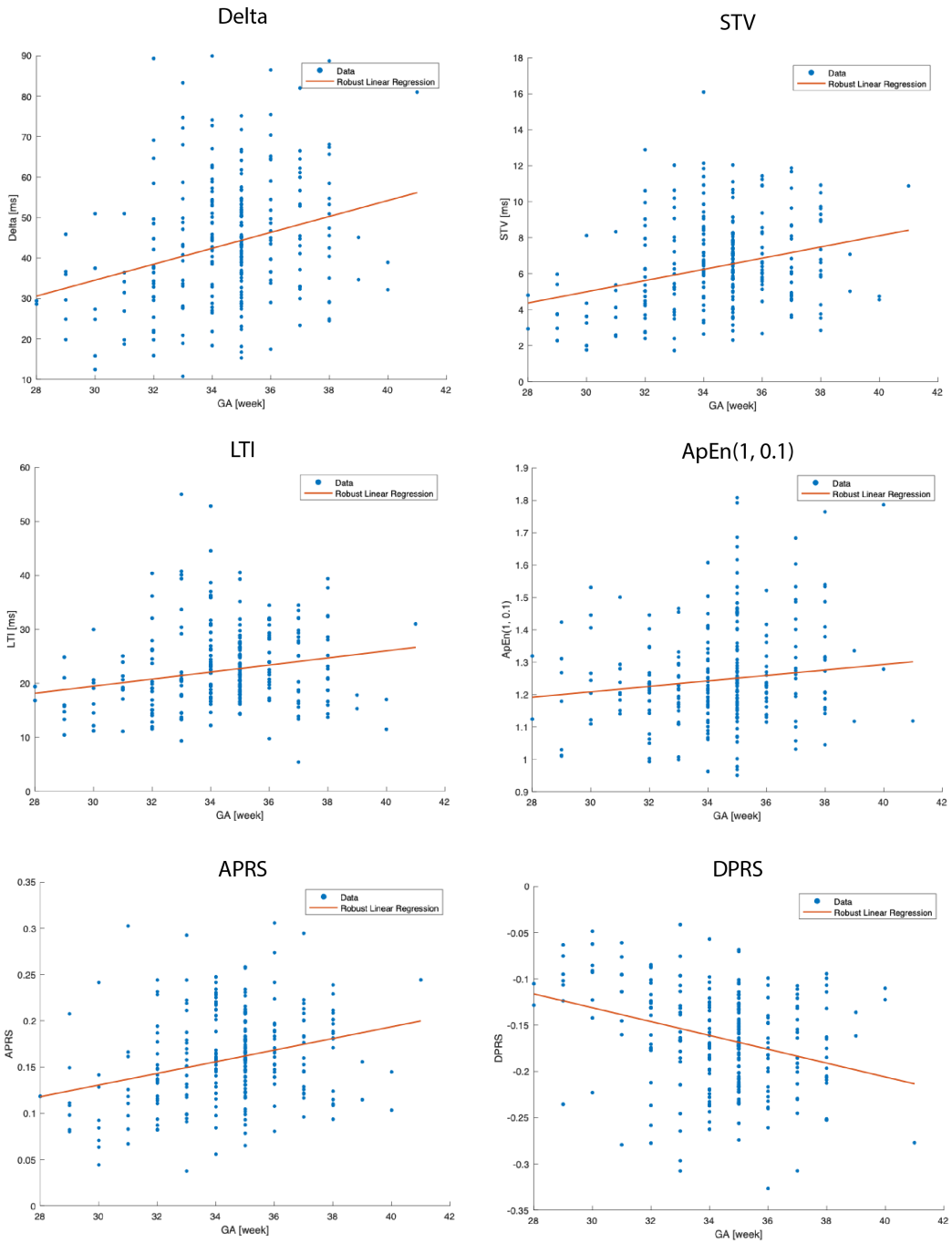


Figure 4.2: Scatterplots showing the distribution of the features Delta, STV, LTI , ApEn(1,0.1), APRS and DPRS with respect to the Gestation Age [week] (blue dots) and the Robust linear regression model of the features dependency on GA computed (red line).

Now that we have computed the linear regression values, we can adjust our subset of parameter in order to remove dependency to the GA. Parameters raw residuals of our dataset are computed

with respect to the regression model defined as :

$$y_i = r_i - \hat{r_i} \tag{4.4}$$

where r_i is the dataset value and $\hat{r_i}$ is the predicted value of our regression model :

$$\hat{r_i} = b_1 + b_2 * GA_i \tag{4.5}$$

where b is the vector of the 2 coefficient of our linear model. The coefficients are shown in the table 4.4.

	Delta	**STV**	**LTI**	**ApEn**	**APRS**	**DPRS**
b_1	-24.5147	-4.3550	-0.1279	0.9553	-0.0582	0.0934
b_2	1.9680	0.3116	0.6533	0.0084	0.0063	-0.0075

Table 4.4: Robust linear regression coefficients b_1, b_2 for adjustment of data. (Eq. 4.5)

In order to verify if our adjustment has correctly been done, we recompute the Spearman's correlation coefficient between the parameters value and the GA. The value can be seen in the table . Of course, the non-adjusted parameter shows the same ρ and p values but we can see that the GA dependency is removed for the 6 adjusted parameters ($p > 0.05$).

	ρ	p
Delta	-0.014636	0.819325
II	-0.060513	0.344586
STV	-0.030394	0.635218
LTI	-0.047280	0.456727
LF	-0.088059	0.165117
MF	0.115392	0.068537
HF	0.041641	0.512211
LF/(MF+HF)	-0.054586	0.390117
ApEn	0.068913	0.280651
LZC	-0.095109	0.132138
APRS	-0.039622	0.531240
DPRS	0.036976	0.559039

Table 4.5: Spearman's correlation coefficient between features values and GA after adjustement by Robust linear regression. None of the p values are <0.05 showing independence of all the parameters over GA

4.2. Features distribution and differences in datasets

Now that we are working with different datasets, it is interesting to compare the distributions between each other and check the differences and inconsistency. The first part will compare the distributions of the parameters in each dataset. A deeper analysis will then be made on the effect of the measurement system (CTG machine) on the signal and therefore the parameters computation.

4.2.1. Features distributions

Now that we have implemented our algorithms to compute the different parameter values and correctly adjust them with respect to the gestational age, an analysis of their different distributions will be made. To get the parameters value for each FHR signal, our pre-processing and computation algorithms is applied. On the other hand we only have access to the parameter value for the open-source data and not the original signals. A final table with all the parameter values for each subject is created. Boxplot of the parameters for each dataset are made using all the values computed or taken from datasets.

Time domain parameters :

First of all, the time domain parameters: **Delta, STV, II** and **LTI** are compared. All of them are variability related. Boxplots are shown in the figure 4.3. Concerning both Delta and STV values, it can be seen that Bloomlife's dataset contains a bit more of higher values whereas Polimi dataset looks to follow the same distribution. This is more or less the same for the Long-term variability even if one can see that the range of values looks bigger in the computed values compared to the open-source values. This could potentially be explained by a stronger pre-processing for the open-source data or by less noise and artefact in the signals. For the Interval Index, the opposite can be seen with Bloomlife having slightly lower values. This can actually be explained by the fact is a metric inversely in which the computation is rationalized by the STV value. Finally, we can conclude that the differences between the distributions are not significant with respect to the dataset size.

Frequency domain parameters :

Frequency features value distributions are represented in the figure 4.4. In this case, a large difference between Bloomlife's data and other sources can be seen. We can see that the LF values of both Polimi and Open-source data are mainly around 75 and 90 whereas Bloomlife values hare often lower with even a mean of 58.38 %. Logically, the same observation can be done for Movement and High frequency range but in the opposite tendency with values higher than those of Polimi and Open-source data. Finally, the ratio $LF/(MF + HF)$ is obviously lower in general for Bloomlife data compared to the others with mainly values between 2-3 against 4-6. Again, Polimi data on

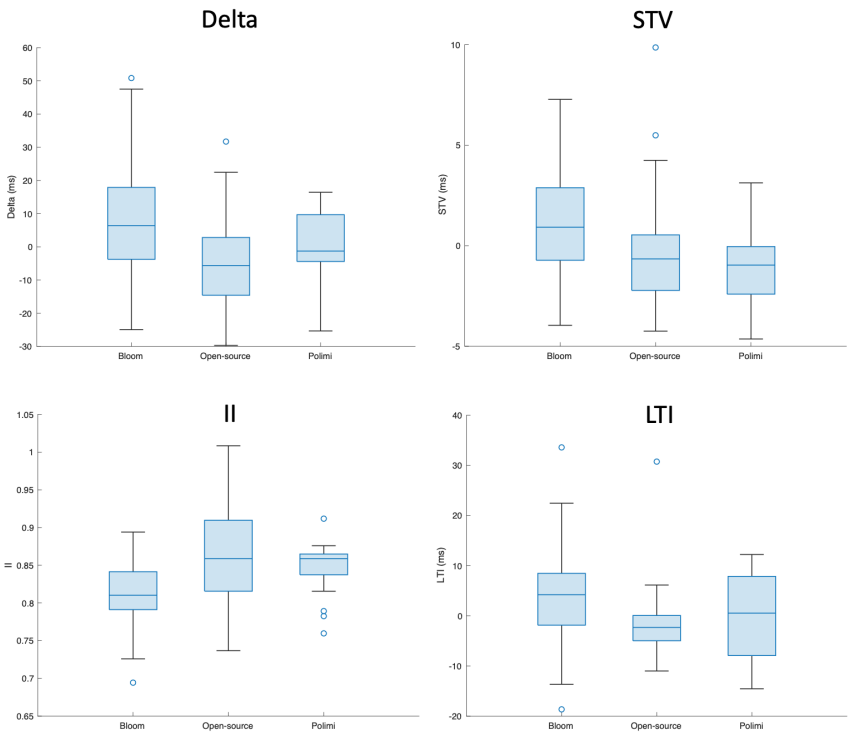


Figure 4.3: Distribution of time domain parameters in each dataset. Up-Left: Delta values (adjusted wrt to GA); Up-Right: Short Term variability (adjusted wrt to GA) ; Down-left: Interval index (non-adjusted wrt to GA) ; Down-right: Long Term Irregularity (adjusted wrt to GA))

the other side seems to be in the same range that the Open-source data.

This frequency features differences could potentially be explained by the difference in measurement device used to acquire data between the sources. This will be studied more deeply in the next subsection (4.2.2).

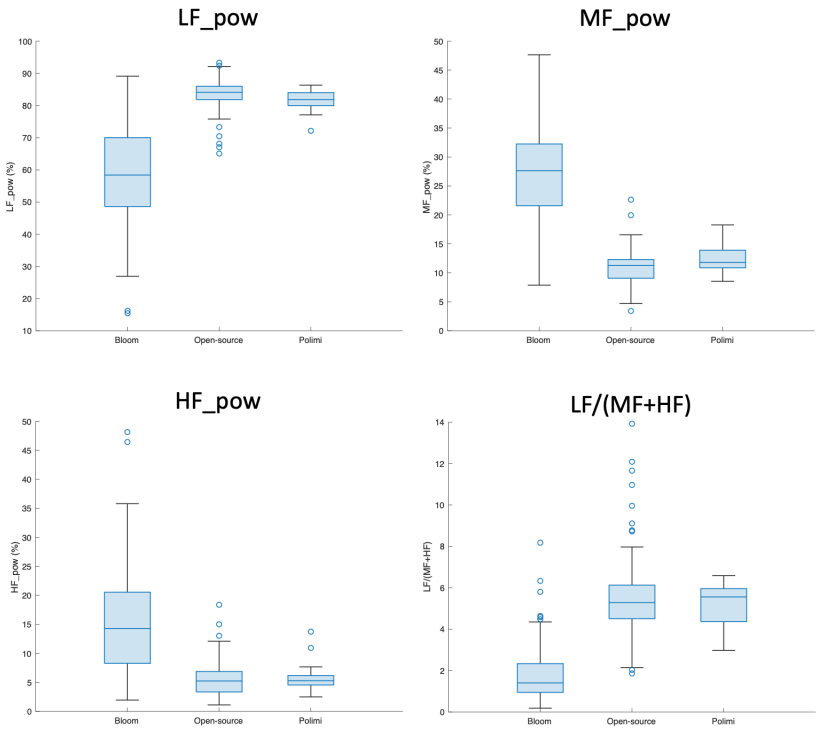


Figure 4.4: Distribution of frequency domain parameters in each dataset. Up-Left: Low frequency power; Up-Right: Movement frequency power ; Down-left: High frequency power ; Down-right: frequency power ratio $LF/(MF + HF)$

Complexity parameters:

The figure 4.5 shows the distribution of **ApEn, LZC** and **SampEn**. It can be seen that Approximate entropy has different distributions even if it is complicated to make any conclusion because of the small size of Polimi dataset, we can see that its Approximate Entropy values are globally high even if some of the Open source values are in the same range. These differences could come either from our implementation or from pre-processing step but one can say that Bloomlife values were computed using the same algorithms. Moreover, Polimi data is only a small dataset composed of 20 subjects, it is then difficult to make any conclusions.

On the other side we see that Bloomlife dataset (composed of 110 signals) is showing values a bit lower than the open source. This is actually the same for the Lempel Ziv complexity values and Sampling Entropy. This could be potentially due to the different measurement systems. This effect is studied in more details in the next section. Finally, sampling Entropy values are not accessible in the Open-source data. Since the Open-source dataset doesn't give access to raw FHR signals, the value are only computed for Polimi and Bloomlife datasets. In the same way than for Approximate Entropy, Polimi data shows globally higher SampEn values than Bloomlife data.

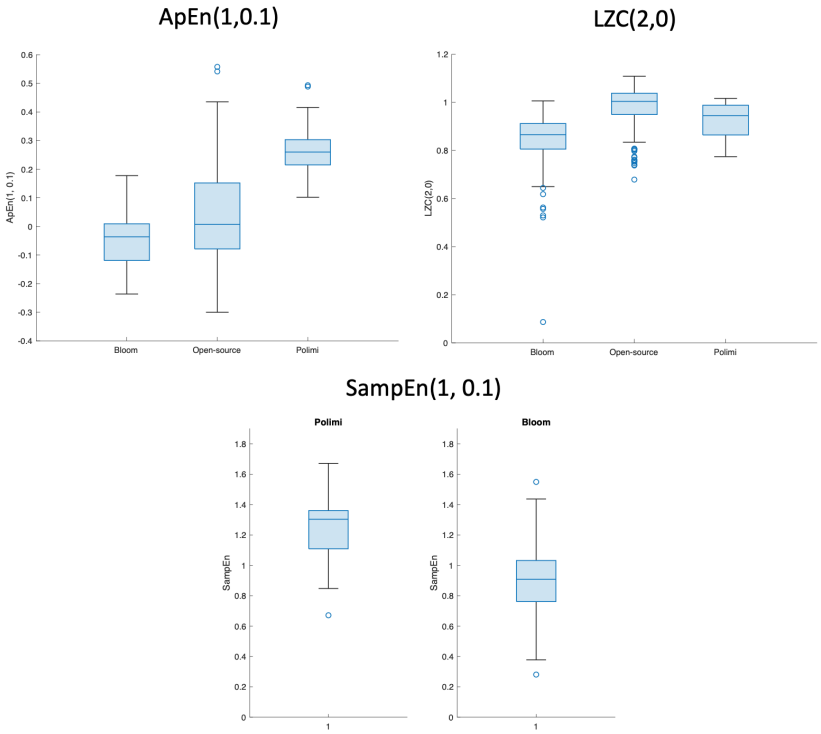


Figure 4.5: Distribution of complexity parameters in each dataset. Left: Approximate Entropy (ApEn(1, 0.1)), Right: Lempel-Ziv Complexity (LZC(2,0)), Down: Sample Entropy (SampEn(1,0.1))

Phase Rectified Slope Amplifier parameters :

In this case, the Open-Source data is only composed of Acceleration and Deceleration Phase Rectified Slope (APRS and DPRS) We check the distribution in the 3 datasets for those two parameters. We see that in this case all datasets have globally the same value range and there is no significant difference between the datasets. Since the other parameters are computed using the same process than the 2 Slope parameters (APRS and DPRS) but using a different final formula, it is not necessary to compare their values in the 3 datasets.

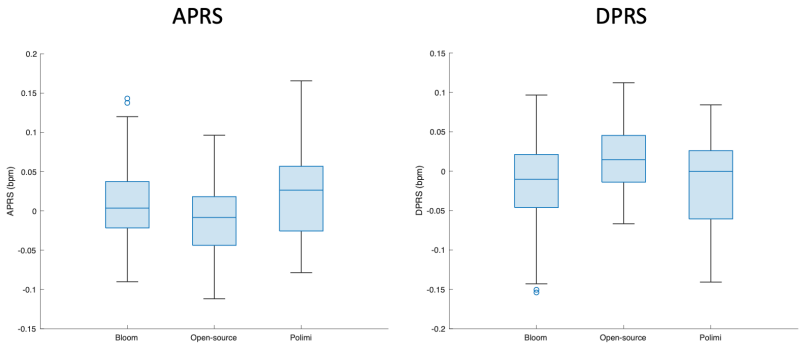


Figure 4.6: Distribution of Phase Rectified Signal Average parameters in each dataset. Left: Acceleration Phase Rectified Slope (APRS). Right: Deceleration Phase Rectified Slope (DPRS)

4.2.2. Effect of the measurement system (CTG)

In this section, an analysis of our results in datasets coming from different measurement systems will be done. We will analyse if there are significant differences between our features values from a dataset to another. To do so, we will analyse the values obtained from:

- Open-source data-set measured with **Hewlett Packard CTG fetal monitors**
- Polimi data-set measured with **Hewlett Packard CTG fetal monitors**
- Bloomlife data-set measured with **Avalon FM30**

To do so, an analysis of the features values distribution in each data-set separated is done. For this, the values from the Polimi and Bloomlife data are computed from the raw FHR signal using all the algorithms explained in chapter 3. Whereas for the Open-source the FHR signal is not available and only access to the parameters values is granted.

Frequency parameters :
One can see in the boxplots that a big difference exists between datasets for the frequency index **LF/(MF+HF)**. Indeed, the values of both the Open-source data and Polimi are around 4 to 8 where Bloomlife have mainly values below 4. This can be seen in the figure 4.7 showing a histogram of the frequency ratio distribution in each dataset.

It can be seen that the Polimi data follows the same tendency than Open-source data whereas the Bloom data looks completely different. In order to go further in the study, we made an ANOVA test analysis to see if the parameters value are dependent from the data source. All the data available were taken and were classified in groups of different sources : *"Open"* ; *"Polimi"*, *"Bloom"*. We use then the Matlab function **anova1.m** to perform the one-way anova. The one-way anova evaluates the impact of the datasets source and determines if there is a statistically significant difference between the means of our three groups. The null hypothesis H_0 is stated saying that each group

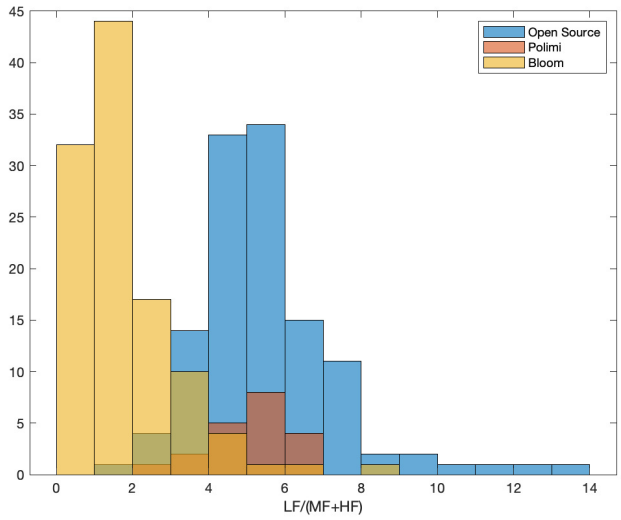


Figure 4.7: Histogram of the distribution of the parameter $LF/(MF+HF)$ in each dataset. Open-source dataset is represented in blue, Polimi in Orange and Bloomlife in yellow.

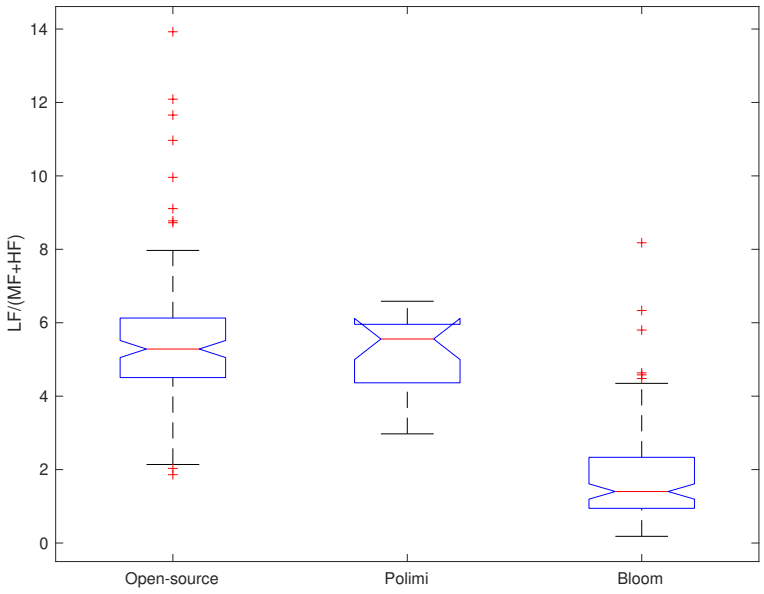
has the same mean and computes the F value as:

$$F = \frac{MST}{MSE} \tag{4.6}$$

Where MST is the mean sum of square due to treatments (difference in source) and MSE is the mean sum of squares due to error. It gives then a ratio comparing the variance between treatments (source groups) and within treatments. A high F value will increase the evidence of inconsistency of the null hypothesis H_0. As said before, the null hypothesis is evaluated stating that different datasets coming from different sources have the same mean.

The figure 4.8 shows the results of the one-way ANOVA test. $F = 162$ showing a significant variance between groups compared to within. The really small p value ($p = 9.7032 * 10^{-46}$) also shows that the null hypothesis cannot be confirmed. Therefore we can conclude that there is a significant difference between groups.
Another interesting point is that if we analyse the F values between the different groups pair, a significant difference happens between *Bloomlife* data and the others but not between *Polimi* and *Open-source* showing a high p value $p = 0.72088 >> 0.05$. This could be explained by the identical measurement system used in the 2 cases suggesting that the difference comes from the measurement device.

A further analysis is made in order to understand where this difference could come from. Thus, a frequency analysis of the raw FHR signals of the different sources is made by making a Power Spectrum Distribution (PSD) analysis. As explained before, the FHR signals are not available on the data from the Open-source. Since the measurement system for the open-source and the Polimi



Source	SS	df	MS	F	Prob>F
Groups	846.59	2	423.293	162.35	9.7032e-46
Error	643.98	247	2.607		
Total	1490.57	249			

Group A	Group B	Lower Limit	A-B	Upper Limit	P-value
{'Open-source'}	{'Polimi'}	-0.61335	0.30066	1.2147	0.72088
{'Open-source'}	{'Bloom' }	3.2472	3.7467	4.2463	0
{'Polimi' }	{'Bloom' }	2.5262	3.4461	4.366	2.0787e-18

Figure 4.8: Results of the ANOVA testing the dependence of the source of the data on the parameter $LF/(MF + HF)$. The first part (up) shows boxplots of the distribution of the parameter across each dataset. The table below shows then the global ANOVA results showing a $F = 162.35$ and $p = 9.7e - 46$. The last tabular shows the ANOVA values done pairwise between data source and their associated p values.

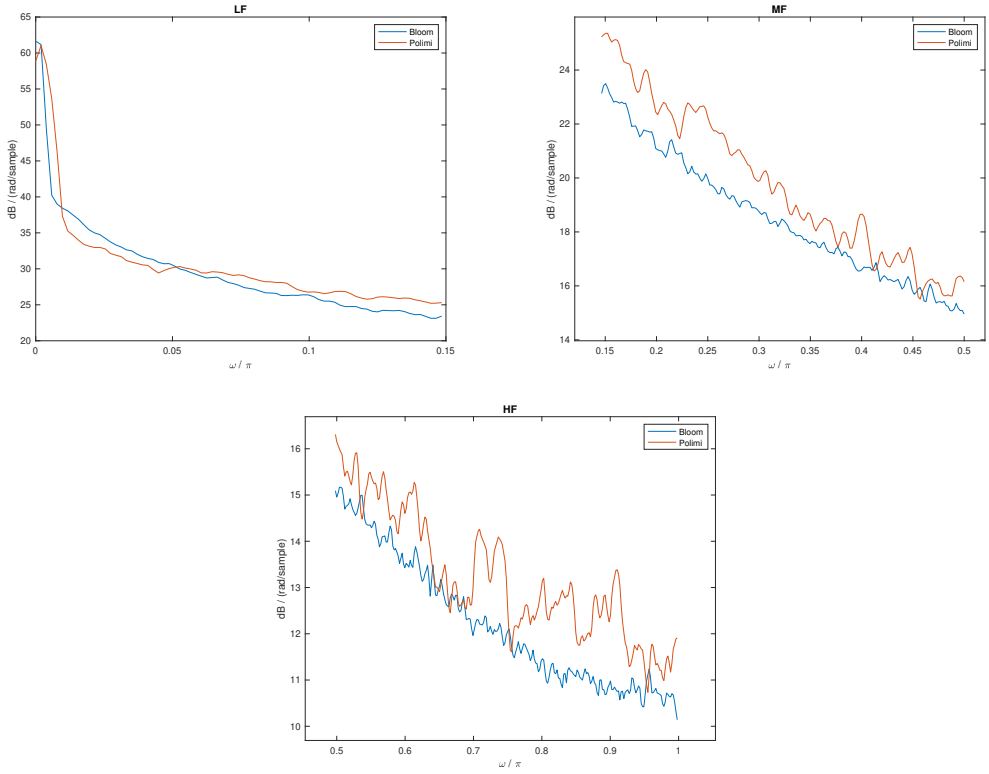


Figure 4.9: Power Spectrum Analysis of all our signals in the different frequency range of interest. PSD wer e computed with a Welch method using windows of 3min and no overlap. The PSD of all the signals from the same dataset were then averaged and represented (in db/(rad/sample)) in blue for Bloomlife data and in orange for Polimi data. Left: Low Frequency range [0.03 0.15]Hz , Righr: Movement Frequency range [0.15 0.5]Hz, Down: High Frequency range [0.5 1]Hz

data is the same and that the results show relatively similar values distribution, we decide to use the FHR signals from this datasets (in parallel with those of Bloomlife) to do our analysis. For the PSD analysis, *Welch* method is used to compute the power spectral analysis for each signal of the dataset. This choice is made to keep consistency with how our parameters are computed in practice. Thus, windows of 3 min ($3*180*f_s$ samples) without any overlap are used. The number of Direct Fourier Transform (DFT) points is setted to $nfft = 1024$ showing a good trade-off between significant PSD sampling and averaging over the spectrum to reduce the effect of artefact. The PSD of each signal is measured and an averaging of the spectrum density of all the signals from the same source is made in order to have a unique and global PSD for the dataset. This is made for both the data from Polimi and Bloomlife, we decide then to cut the PSD into the different frequency range of interest $LF = [0.03; 0.15]$, $MF = [0.15; 0.5]$ and $HF = [0.5; 1]$ Hz defined previously. Plot of the results can be seen in the figure 4.9. There is a difference in the average frequency spectrum between the 2 datasets. This can explain the difference in values in the frequency parameters.

On the other hand, even if the previous results shows a difference, it shows qualitatively an opposite trend compared to the parameters values computed. Indeed, Bloom data were showing lower

LF/(MF+HF) values than Polimi ones, suggesting this :

$$LF_{Bloom} << LF_{Polimi} \text{ or } MF_{Bloom} >> MF_{Polimi} \text{ or } HF_{Bloom} >> HF_{Polimi}$$

Whereas our results are showing the opposite tendency. An explanation could in fact come from artefacts of our signals since we previously took the raw FHR signals for our PSD analysis. In order to have a deeper understanding, we decide to make another PSD analysis based this time on the pre-processed signals. Our signals were pre-processed with our pre-processing function. After that the bad quality parts of the signals is removed, it is replaced by a linear interpolation in order to affect as they are as possible the frequency spectrum of the good part of the signal. Doing this, the effects of artefacts on the frequency spectrum are dampened. The PSD of all segments are averaged to get a unique PSD for the signal. In the same manner, the PSD's of all the signals from the same sources are averaged to get a global estimation of the frequency spectrum for each datasets. The results can be seen in the figure 4.10.

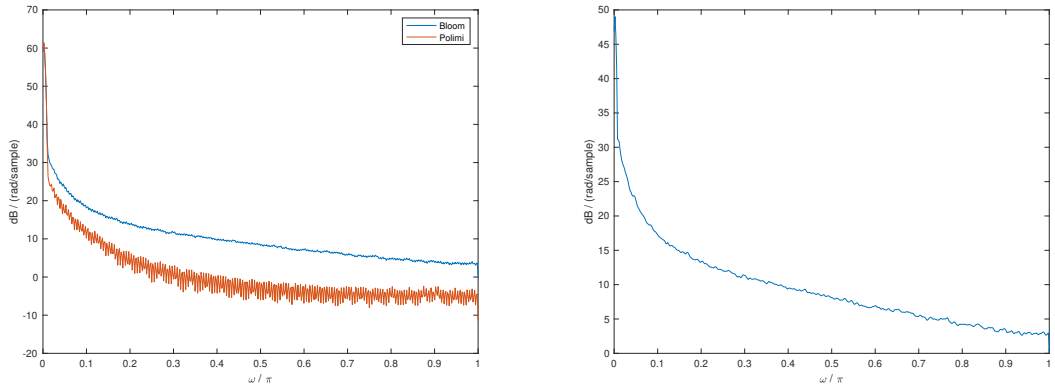


Figure 4.10: PSD analysis of Polimi and Bloomlife datasets with pre-processing removing bad signals and artefacts. Left figure shows the 2 averaged PSD in [dB], Bloomlife in blue and Polimi in orange. On the right, the figure shows the difference between the 2 PSD (Polimi - Bloom) in [dB]

As we can see, the average PSD of Bloomlife is higher in our frequency of interest than Polimi's one. What can also be seen is the difference between LF range and MF and HF ranges (much smaller) in the Polimi dataset which increase the ratio $LF/(MF + HF)$. To have a quantitative information, we also compute the sum of the values of the PSD in the frequency range and get the ratio value for the mean PSD. The values are shown in the table 4.6.

	LF	MF	HF	LF/(MF+HF)
Polimi	2080.2849	290.9006	97.4811	5.3562
Bloom	8093.1962	2798.4288	980.6627	2.1415

Table 4.6: Values of of the sum in the frequency ranges (LF,MF and HF) of the mean PSD and values of the ratio LF/(MF+HF) for the averaged PSD.

The values of the ratio are coherent with the distribution seen earlier in the dataset. It can be concluded that there is a difference in the frequency spectrum due to the measurement device.

Lempel Ziv complexity (LZC) :

Another parameter showing different range of values between datasets is the Lempel Ziv Complexity (LZC). As a reminder, LZC is used to quantify the rate of new pattern arising in a sequence of a binary values [29]. In this case binarization of the signals is done with the encoding rule based on the slope of the signal and defined in the equation 3.11. The complexity is then normalized as defined in the original paper and in the section 3.3.2. One can see that the complexity value can be dependent on how is acquired and sampled the signal. Indeed, our initial try was to compute the LZC value with the signal at its sampling frequency. Bloomlife's signal were then computed with a $f_s = 4Hz$ whereas in the Polimi and Open-source dataset are sampled at 2Hz. Hence, a significant difference was directly seen because of the length difference of our binary signals.

To overcome this issue, we decide to downsample signals coming from Bloomlife dataset by a mean average to get a 2Hz signal equivalent to the other datasets. Even after this, the values of Bloomlife dataset are significantly lower than the one from the Open-source data and those computed based on Polimi dataset signals. To understand a bit better from where could come this difference, we analyse in more details the raw signals on which parameters are computed. Example of signal are shown in the figure 4.11.

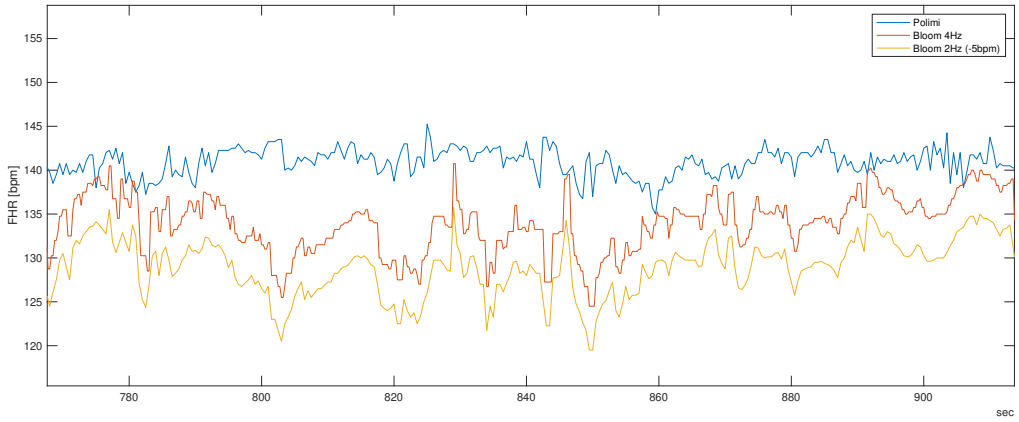
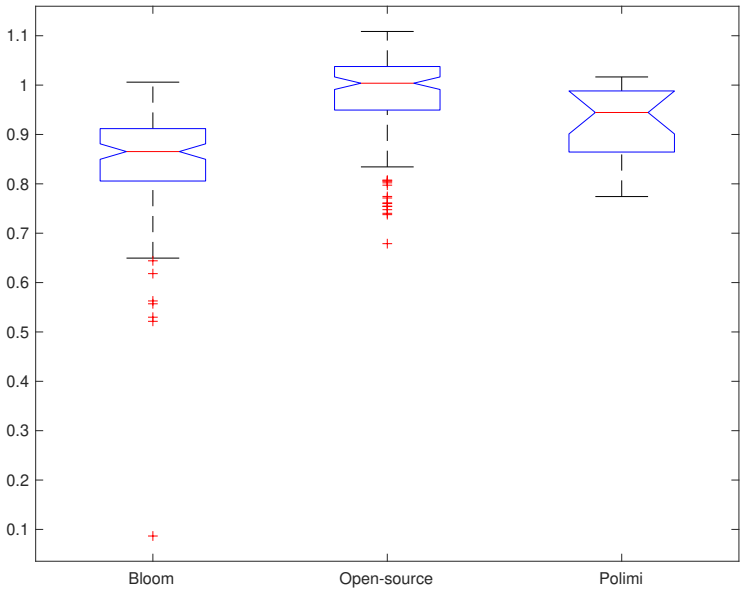


Figure 4.11: Examples of raw FHR signals coming from different datasets. In blue, a segment of a signal from Polimi, sampled at 2Hz. In red, a signal from Bloomlife's dataset sampled at 4Hz and in orange the same signal downsampled at 2Hz by mean averaging (-5bpm).

As we can see, even if the Bloomlife signal is sampled at a higher frequency (4Hz) the signal looks more jerky. This can come from the precision in the signal digitization. A potential explanation could be that the signal is coded digitally on a lower amount of bit, showing a signal with some steps

with some flat areas. Moreover, one can also see that the Polimi signal shows more small variations probably due to measurement noise. Since the Lempel Ziv complexity is computing complexity based on the change of the signal slope those variations and change in the discrete coding could have an impact the LZC values.

In order to confirm that we have a real distinction between the sources, we make an Anova test in the same way than for the frequency parameters. The results of the One-way anova test are shown in the figure 4.12.



Source	SS	df	MS	F	Prob>F
Groups	1.04818	2	0.52409	46.71	5.96195e-18
Error	2.79403	249	0.01122		
Total	3.84222	251			

Group A	Group B	Lower Limit	A-B	Upper Limit	P-value
{'Open-source'}	{'Polimi'}	-0.012863	0.047135	0.10713	0.15626
{'Open-source'}	{'Bloom' }	0.10148	0.13409	0.1667	2.1882e-22
{'Polimi' }	{'Bloom' }	0.026725	0.086952	0.14718	0.0020658

Figure 4.12: Results of the ANOVA testing the dependence of the source of the data on the parameter $LZC(2,0)$. The first part (up) shows boxplots of the distribution of the parameter across each dataset. The table below shows then the global ANOVA results showing a $F = 46.71$ and $p = 5.96e-18$. The last tabular shows the ANOVA values done pairwise between data source and their associated p values.

We can see that the F value (cfr eq. 4.6) is high ($F = 46.71$) meaning that the variance between groups is proportionally really higher than the variance within groups. In the same way, the p value is really small ($p = 5.96 * 10^{-}18 < 0.05$). We can then conclude that the difference between groups is significant. Again, we see that the group showing the higher difference is Bloom whereas

the p-value between Polimi and the Open-source data is not small enough to contradict the null Hypothesis H_0.

4.3. Dataset selection

In the last section, we have studied the impact of the measurement system on the signals and therefore the computed parameters. One-way anova tests showed us that the group showing the biggest difference in value is Bloomlife dataset. Knowing that significant differences exist between datasets, it is not possible to use all of them for our model. Moreover, deleting the parameters influenced by the measurement system would lead to a set of parameters too small to have interesting results for our model.

The initial idea of this work was to use raw FHR signal to extract parameters as input for a model. Unfortunately, full access to FHR signals are only given in Polimi data and Bloomlife data. Knowing that Bloomlife data is not correctly annotated by clinicians, we only have 20 signals retrospectively annotated as Healthy or IUGR. One can know that this population cannot be used to build an interesting model due to its reduced size.

Our choice finally went for the Open-source dataset.This one was correctly annotated by clinicians andits size (60 Healthy / 60 IUGR) allows us to build and train a simple model on it. Since we don't have access to the signals, the algorithm only uses the computed parameters found in the dataset. Our set of parameters is then slightly reduced due to the fact that Sampling Entropy, Acceleration/Deceleration Capacity and Averaged Acceleration/Deceleration Capacity are not available in this dataset.

In the next chapter, the open-source data will be used to build a classification algorithm. Let's also say that the data used as input is adjusted with respect to the gestational age as explained in the section 4.1. The prediction algorithm will be studied in details in the next chapter.

5 | Prediction Algorithm

5.1. Inputs and Outputs of the model

Before training our models, let's explain what will be the inputs with which our model works with and the outputs we tend to obtain.

Inputs :
As explained in the previous chapter, the inputs of our model will be taken from the open-source dataset from *Data in Brief* [44]. The recorded populations consist of two groups of fetuses: 60 healthy and 60 Intra Uterine Growth Restricted (IUGR) fetuses. The dataset is composed of 12 features value for each subject. The complete list containing parameters explained in chapter 3 is composed of :

- Time indices : **Delta, STV, LTI, II**
- Frequency indices : **LF_pow, MF_pow, HF_pow, LF/(MF+HF)**
- Non-linear indices : **LZC(2,0) ApEn(1,01), APRS, DPRS**

A more detailed explanation of the open-source dataset can be found in Appendix A or in the reference link [44].

Let's also remark that the features showing GA dependency is adjusted according to the process explained in chapter 4. The final data used takes the form shown in table 5.1. The first column is the class annotation that the model wants to predict (output) and that will be used for supervised learning. The second column is the Gestation Age of the fetus when the CTG monitoring was made. As explained earlier, this will not be an input to the model neither but will be used for the adjustment of other parameters.

State	GA	DELTA	II	STV	LTI	LF	MF	HF	LF/HF+MF	ApEn(1,0.1)	LZC(2,0)	APRS	DPRS
'Healthy'	34	14.85	0.92	2.18	-0.22	82.52	15.12	2.36	4.72	-0.002	1.043	0.056	-0.063
'IUGR'	31	-5.05	0.88	-1.74	1.23	87.22	8.52	4.26	6.82	-0.011	1.025	-0.026	0.0434

Table 5.1: Example of data input for our model. The first column is the restrospectively annotated State used for supervised training, the second column is the GA [weeks] used to adjust dependent parameters, the following columns are parameters values used as input.

The dataset is separated in 2 parts. The first part is the training and validation dataset and is composed of 100 subjects (50 Heathy/ 50 IUGR) and 20 subjects randomly chosen are removed to be

in the test set. This set will allow us to test our trained data on a completely independent data after beeing trained. Because of the small amount of data, we choose not to use hold-out validation.[50] Instead, we will use a 10-fold validation (90 subject for training and 10 for validation everytime).

Outputs:

Concerning the outputs, the model should classify the subject as "IUGR" or "Healthy". Even if the outcome is binary, the output should be used as an indication to be reviewed with additional tests and monitoring by clinicians. The model will be trained by the 'State' annotation made retrospectively by clinicians.

5.2. Potential models

In this section, we will investigate on which model can be used for IUGR prediction. To do so, a quick overview of the theory used by the model will be made. Then, a first sight of the performance of the models will be analysed in our training/validation dataset using 10-fold cross-validation.

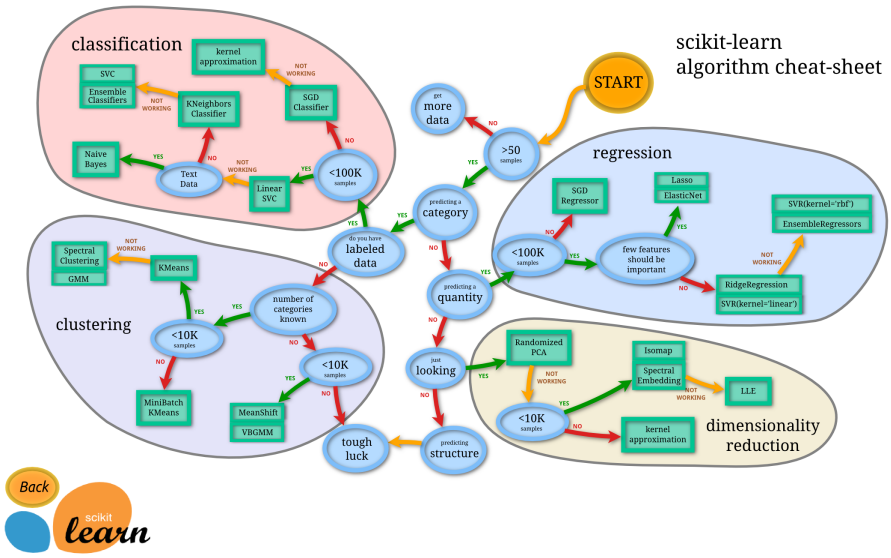


Figure 5.1: Algorithm cheat sheet from *Skicit learn* [7] Roadmap for our choices of classification models.

In order to select models suitable for IUGR prediction we follow the "algorithm cheat-sheet" roadmap from *Skicit learn* [7] (cfr figure 5.1). Knowing that we have 100 samples in our training/validation dataset, that the algorithm should predict a category and that we are in presence of labeled data. The first algorithm to test should be the **Linear Support Vector Machine**. Since we are not in the case of text-data, another option is the **K-Nearest-Neighbors** algorithm. After this, the performance of a **Decision Tree** will be made and finally an **Ensemble classifier** will be also tested. Afterwards, the models will be analyzed and compared to select the optimal one. Therefore, the following subsections will study the models :

- Linerar Support Vector Machine (SVM)

- K-Nearest Neighbors (KNN)
- Decision Tree
- Bagged Ensemble Trees

To assess the performance of our model, let's remind the basic metrics that we use. Those are shown in the figure 5.2.

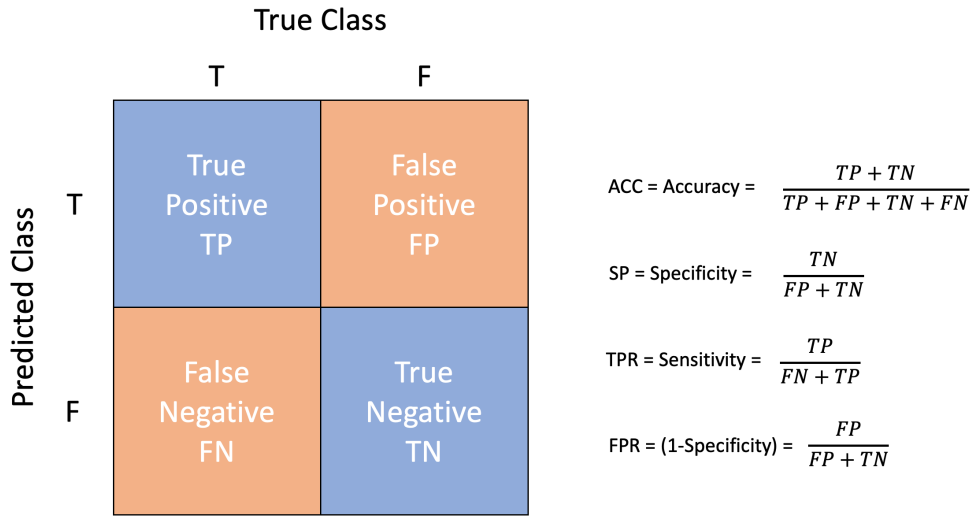


Figure 5.2: Confusion matrix and definition of the metrics associated : Accuracy (ACC), Specificity (SP), Sensitivity (TPR) and False Positive Rate (FPR).

We will have a global look on the different metrics trying to maximize TP and TN and minimize FP and FN. Since we are trying to build a model for screening, the focus is on minimizing FN as much as possible and so have the highest sensitivity possible. In the medical field in the case of an additional metric for diagnostic, FN should be the lowest possible. Having a FP a bit higher is less important because the docor will analyse further the data and will more easily be able to review the decision. [19]

5.2.1. Model 1: Linear Support Vector Machine (SVM)

Support vector Machine (SVM) is a supervised learning algorithm which can be used for classification and regression problems. Here a linear support vector classification (SVC) is analysed [5]. It is based on the construction of a hyperplane or set of hyperplanes in a high-dimensional space. The separation will be optimal as the margin gets higher. The margin is the distance of the hyperplane to the nearest training-data point of any class.

In this case, we train a Linear Support Vector machine with the training dataset. We use 10-fold cross-validation for our training and validation data. As a reminder, the data used is the one

explained in the section 5.1 and is composed of 100 subjects divided in 50 Healthy and 50 IUGR subjects. The 12 features listed on the previous section are used for each subject. For the model, we used a linear Kernel function, standardized data and since we deal with binary classification the multiclass method used is logically One-vs-One (the classification will splits the dataset into one dataset for each class versus every other class).
The model is trained by the Matlab classification learner app. After training, the classification algorithm shows a global accuracy of 78% averaged over the 10 folds. The validation cost over the 10 x10 subjects validation set (10-cross validation) is of 22 errors. The training time is around 1.49 sec with a prediction speed of more or less 1900 obs/sec. The confusion matrix of the validation dataset can be seen in the figure 5.3.

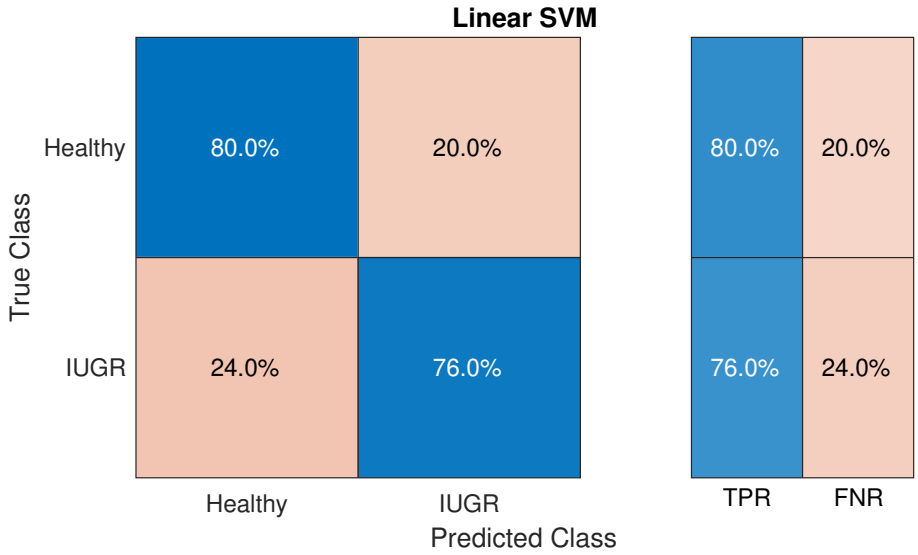


Figure 5.3: Confusion Matrix of the trained Linear SVM model over validation data.

We can see that 76% of IUGR patients are detected whereas 80% of Healthy patients are correctly categorised as Healthy. Therefore, 20 % (10/50) of Healthy subjects are classified as IUGR but the most important issue is that 24 % (12/50) of IUGR subjects are not correctly diagnosed and are misclasiffied as Healthy. This represents an issue because as a model used for screening, one would want to have as less as possible False negative and so have an algorithm with a higher sensitivity. Indeed, in the case of a "false positive" mis-classification, the fetus will be monitored with a higher focus. The diagnostic could then be reviewed later by the clinician expertise. On the other hand a "false negative" error will lead to think that the baby is not in growth restriction, the follow-up will then be less important and the error less easily reviewed. Hence, a FN error could have a real impact on the fetus health due to the lack of attention whereas it would not be the case in the opposite case.

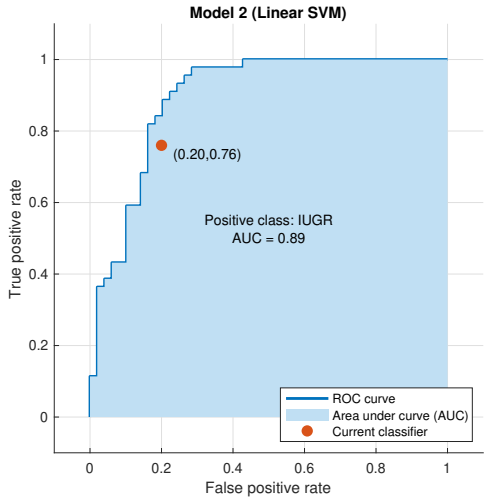


Figure 5.4: ROC curve (blue line) of the linear SVM model. The Area Under the Curve ($AUC = 0.89$) is the the sky blue area and our current classifier performance is the red dot.

It can be seen that the true positive rate is at 20 %. As explained earlier, an optimal model should have a high sensitivity (lowest false negative rate). In this case, we see that the model has more or less the same performance in specifity and sensitivity. An additional parameter tuning could consist in using a modified misclassification cost matrix with higher cost for false negative than for false positive. The Receiver Operating Characteristic curve (ROC) of the model is shown in the figure 5.4. ROC plots TP rate (sensitivity) versus FP rate (1-specificity) across varying cut-offs . "The curve corresponding to progressively greater discriminant capacity of diagnostic tests are located progressively closer to the upper lefthand corner in "ROC space". An ROC curve lying on the diagonal line reflects the performance of a diagnostic test that is no better than chance level. The area under the curve (AUC) summarizes the entire location of the ROC curve rather than depending on a specific operating point the AUC is an effective and combined measure of sensitivity and specificity that describes the inherent validity of diagnostic tests." [25] We wee that in this case, the classifier's True Positive rate could be increased to higher values at the expense of the False positive rate.

5.2.2. Model 2: K-Nearest-Neighbours

The second model investigates the performance of a K-Nearest Neighbors model. This model is a non-parametric classifier. It uses proximity to make classification. It identifies the nearest neighbors of a specific point, and assign a class label to that point using a determined distance metric.

Our model was trained with the Matlab Classification Learner app in the same manner than the SVM. Our 100 subjects (50/50) training/validation dataset is trained using 10-fold validation and used the 12 features. The model chosen is a Medium KNN, the number of neighbors of influence

is set to 10. This choice seems to be a good trade-off for the size of our training data (100) and the precision we want to achieve. The model uses Euclidean distance with equal distance weight. Finally, the data is standardized again.

Our trained model gives us a global validation accuracy of 76 % over the 10 folds. This means that over the 10x10 subjects validation misclassification occured 24 times. The prediction speed is around 750 obs/sec which shows slower performance than the SVM with also a training time of 2.658 sec.

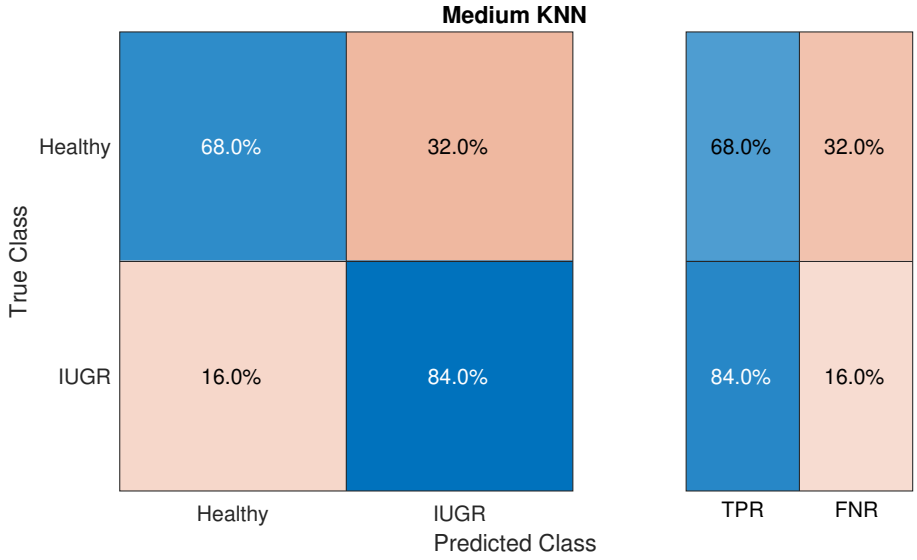


Figure 5.5: Confusion Matrix of the trained Medium KNN model over the validation data.

The confusion matrix of the Medium KNN trained model can be seen in the figure 5.5. We can see that even if the global accuracy is less good (76 %) than the SVM model, the KNN performs better for the IUGR case with 84% of IUGR detected and so 16% of False negative. The sensitivity of this model is then 84%, better than the SVM sensitivity.

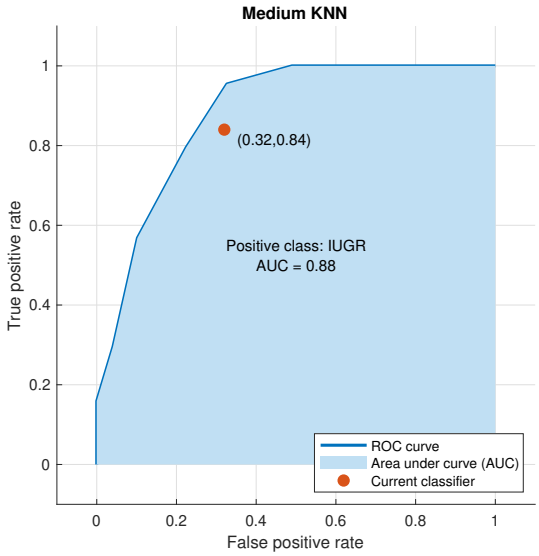


Figure 5.6: ROC curve (blue line) of our Medium K-Nearest Neighbours model. The Area Under the Curve ($AUC = 0.88$) is the the sky blue area and our current classifier performance is the red dot.

Analysing the ROC curve shown in the figure 5.6, one can see that our current classifier has a sensitivity and specificity respectively of 84% and 68%, and that the $AUC = 0.88$. The sensitivity could potentially be increased, but it would quickly lead to a specificity under 50%.

5.2.3. Model 3: Medium Decision Tree

The third model is a Decision tree. A Decision tree model is a supervised machine learning model in which the data is split according to criteria over parameters. It is composed of 3 basic entities: decision nodes, branches and leaves. Decision nodes are where the data is split, branches are the paths going from a node to another and the leaves are final outcomes. [10]

For our classification model, knowing that our training/validation data is only of 100 subjects and that we use 12 features, a Medium Decision tree is chosen with a maximum number of splits of 20 to avoid overfitting. The tree used the Gini's Impurity criterion measuring how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. To compute it, the probability p_i of an item with a specific label i is summed to be chosen and is multiplied by the sum of the probabilities to mis-classify it $\sum_{k \neq i} p_k = 1 - p_i$. The Gini impurity criterion will tend to 0 when all cases in the node fall into a single target category. [10]

The trained Decision Tree model shows good performance on the validation data. It outperformed the other models with a global accuracy of 91 %. Thus, it shows only 9 validation cost over the

10x10 subjects validation subsets (so less than 1/10 in average). In addition, the prediction speed stays good with around 1900 obs/sec and so a training time 1.9199 sec.

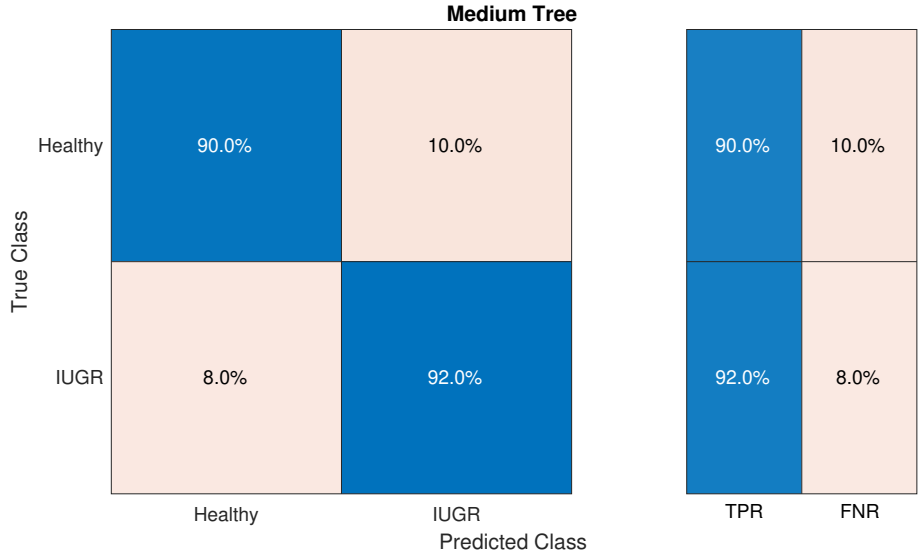


Figure 5.7: Confusion Matrix of the trained Medium Decision Tree model over validation data.

As we can see in the Confusion matrix in figure 5.7 the model has good results both for IUGR and Healthy subjects with 10% of false positive and only 8% of false negative. The ROC curve is shown in the figure 5.8. We see that the AUC is equal to the one of the SVM model, but that our Decision tree has an optimized sensitivity (92 %) and specificity (90 %). As we said previously, we want to have an optimal sensitivity in order to avoid False negative errors. In this case, only 8% IUGR are not-detected. A result significantly better compared to the 2 other models.

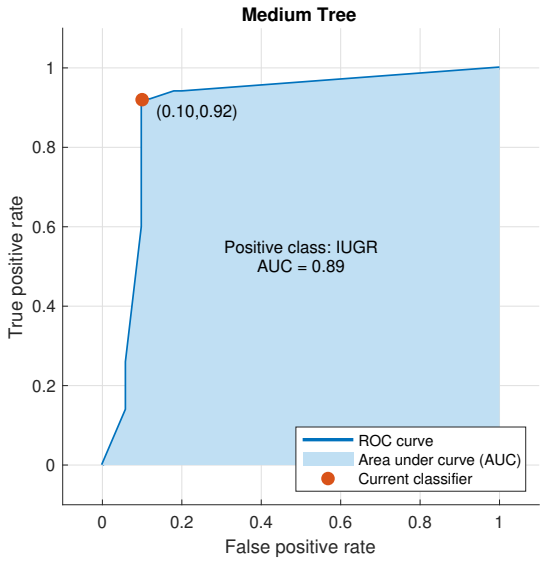


Figure 5.8: ROC curve (blue line) of our Medium K-Nearest Neighbours model. The Area Under the Curve ($AUC = 0.89$) is the the sky blue area and our current classifier performance is the red dot.

5.2.4. Model 4: Bagged Trees Ensemble

Finally, we train a Bagged ensemble of trees as suggested in the "algorithm cheat-sheet". A Bagged ensemble model, is an ensemble of weak models trained in parallel. In bagging, "a random sample of data in a training set is selected with replacement. After several data samples are generated, these weak models are then trained independently. " [21] After training, the average of prediction is taken to compute an estimation used for the final classification. This kind of models is used to reduce variance in an imperfect dataset. [21]

For this Ensemble model, we decide to use 5 different trees with a maximum of 10 splits. Since only 12 parameters are used (and that some of them are related) and that our training dataset is only of 100 subjects, these ensemble parameters seem to be a good trade off to have weak trees analysing differently the data without overfitting it. This relatively small number is also chosen such as we can still analyse the different trees individually. Again, the model is trained on the training dataset with 10 fold validation. Each tree is then trained separately and then bagged together. The confusion matrix of the final model can be seen in the figure 5.9.

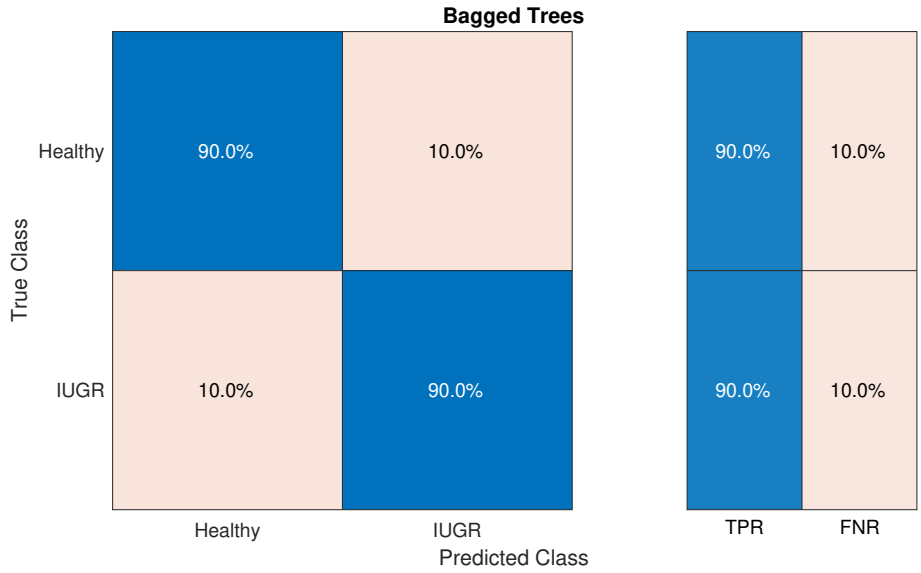


Figure 5.9: Confusion matrix of the trained Bagged Ensemble Trees model over validation data

The trained Ensemble model shows good performance on the validation data with results comparable to the simple decision model. The global accuracy reaches 90 %. Thus, it shows only 10 validation cost over the 10x10 subjects validation subsets (so less than 1/10 in average). On the other side, the prediction speed is lower with around 670 obs/sec and so a training time of 4.2212 sec.

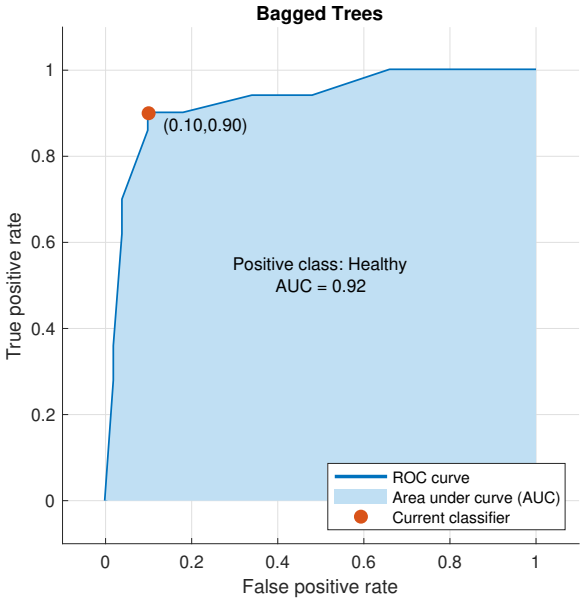


Figure 5.10: ROC curve (blue line) of our Bagged Ensemble model. The Area Under the Curve (AUC = 0.92) is the the sky blue area and our current classifier performance is the red dot

The ROC curve is shown in the figure 5.10. We can see that the model achieves to have a 90%

sensitivity and specificity and that the AUC is higher than for any other models with $AUC = 0.92$. We see that the model misclassify only 5 IUGR subjects (only one more than the previous model) which represents a good results for our application.

5.2.5. Model Selection

In this section, the different models we trained for the prediction algorithm will be compared. A model will be selected to be optimized and analysed in the next sections of the work.

As a reminder, the main purpose of this algorithm is to give an additional information for clinicians for the diagnostic of IUGR fetus. In this case, the application is more a screening process than a diagnosis process. Indeed, it would have more impact to classify an IUGR fetus as an Healthy fetus than misclassify a Healthy fetus as IUGR. The consequence of a false positive will be that the clinician will better monitor the pregnancy and probably make additional tests able to review the information given by the prediction algorithm whereas in the opposite case, the clinician could badly monitor a pregnancy at risk. Therefore, the focus is to decrease as much as possible the false negative rate. On the other side, the specificity shouldn't be too low neither otherwise the algorithm will diagnose too much healthy cases as IUGR and will lead to worry too much clinicians that will not trust the algorithm as a final result.

Let's also say that in this case, the speed is not a main limitation factor. Indeed, we don't need really fast decision in our cases. Since the signal measurement takes already dozens of minutes up to 1 hour. Even if a fast algorithm is always more interesting even more when will have to train larger amount of data, this is not a big decision factor in this case.

	Linear SVM	Medium KNN	Decision Tree	Bagged Ensemble
Accuracy	78%	76%	91%	90%
Prediction speed	1900 obs/sec	750 obs/sec	1900 obs/sec	670 obs/sec
TPR (sensitivity)	76%	84%	92%	90%
SP (specificity)	80%	68%	90%	90%
AUC	0.89	0.88	0.89	0.92

Table 5.2: Performance comparison of the 4 models trained on the Open-source training dataset (from *DatainBrief*)

The table 5.2 synthesizes the most important performance metrics of the models. As we can see, the Decision Tree and the Bagged ensemble outperform the 2 other models either in validation accuracy, sensibility or specificity. With a sensitivity of 90%, 5 IUGR subjects were misclassified in the case of the Ensemble compared to 4 for the Decision Tree. The Bagged ensemble has also a lower speed but as we already said, this is not an important factor for our application. Comparing the areas under the curve, the bagged ensemble has the best performance with a $AUC = 0.92$. The

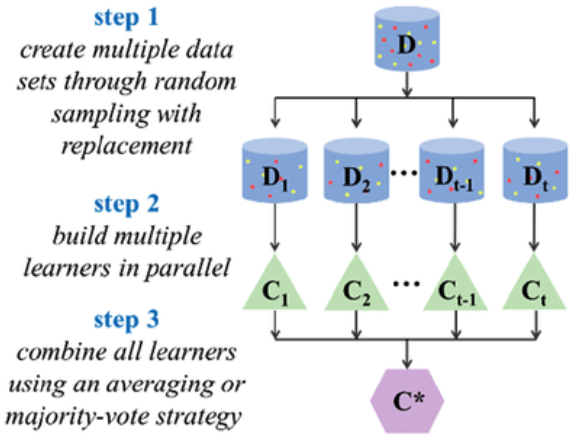


Figure 5.11: Illustration of the different steps of bagged ensemble algorithm training [53]

2 last models are comparable and the choice is then really difficult to do, but one can think that using a bagged ensemble will reduce variance and overfitting and would potentially perform well in the case of a larger training dataset.

In the next section, the bagged ensemble model will be analysed and optimized as much as possible in order to have the best model for our application.

5.3. Model Analysis

5.3.1. Bagged Ensemble model analysis

Before going in any further improvement and optimization of our model, the characteristics of our selected model will be analysed. As a reminder, our ensemble model is composed of 5 weak trees of maximum 10 splits afterwards bagged together. The idea is that our final classification prediction is function of the forecast of the 5 trees. Using a set of bagged weak trees helps not to overfit our dataset while keeping a good prediction thanks to the fact that a misclassification in a tree can be corrected by the correct classification of the others. Indeed, the combined estimator is usually better than any of the single base estimator because its variance is reduced.

Suppose a set of S of s signals. At each interaction of our model, a new training set S_i of s signal is sampled with replacement from the set S. Each training set is then used to train our model and a model M_i is created for each set D $<$ i. After this, each tree returns its prediction and the final bagged ensemble classifier counts the predictions to assign the final forecast as the one with the most votes. As a binary classification, our final prediction is the classification (IUGR or Healthy) that occurs the most time ($p > 0.5$) in all the predictions of our weak trees. [12] An illustration of the process of the algorithm can be seen in the figure 5.11.

In order to study a bit more which parameters influence our model and how. We compute the

predictor importance of each parameter.

We want to focus our modeling efforts on the predictor fields that matters the most and consider dropping or ignoring those that matters the least. To do so, predictor importance chart of our model is studied. In this chart, the relative importance of each predictor in estimating the model is given. The predictor importance shows the influence of each predictor in making a decision whether or not prediction is accurate. It is not related to predictor accuracy. The purpose of predictor importance is also interesting to study if two inputs are carrying the same information for the model. For example if 2 predictors such as LTI and STV are strongly related to the State prediction, then feature selection will say that both are important predictors but you might find that in fact only one of the 2 is really used in our trained model because they carry the same information. Since we have predictors that are parameters related to a same time, frequency or complexity concept, one can say that it is of interest to study the predictor importance of our model. [9]

The predictor importance for each parameter can be seen in the right side of the figure 5.12. As we can see the parameter with bigger importance is the Lempel Ziv Complexity, followed by the High Frequency power and then the Short Term Variability. This observation is interesting since each one is a different kind of parameter (complexity for LZC, frequential for HF_pow and variability for STV). In addition, we see that parameters with the less importance are LFMFHF, DELTA and APRS also from the 3 types. This can be explained by the fact that HF and LF are inversely correlated by definition and this is the same for APRS with DPRS. One can also think that Delta would also be correlated with the STV as a variability parameter over a short time frame (1min).

In addition to this, we also compute the Out-of-bagged predictor importance. As explained in mathworks documentation [1]. "Out-of-bag, predictor importance estimates by permutation measure how influential the predictor variables in the model are at predicting the response." The influence of a predictor is studied by permuting its value and see how it affects the model error. If the permutation has no influence on the error, then the predictor can be characterised as not influential for the model. Inversely, the influence of the predictor increases with the value of the increase in the model error due to the permutation. On the other side, if the predictor importance is negative, it would mean that the permutation even increases the predictor value and performance of the model.
As we can see, the LZC and the High Frequency power are again the 2 most important predictors. A difference is that LF/(MF+HF) shows also a high predictor importance. This could be explained by the fact that it is inversely correlated to HF_pow in the previous case whereas in this case the correlation has less importance. In this case, we see that the DELTA, STV and MF_pow have a negative predictor importance. This could mean that those predictors mislead our model and could potentially create additional errors.

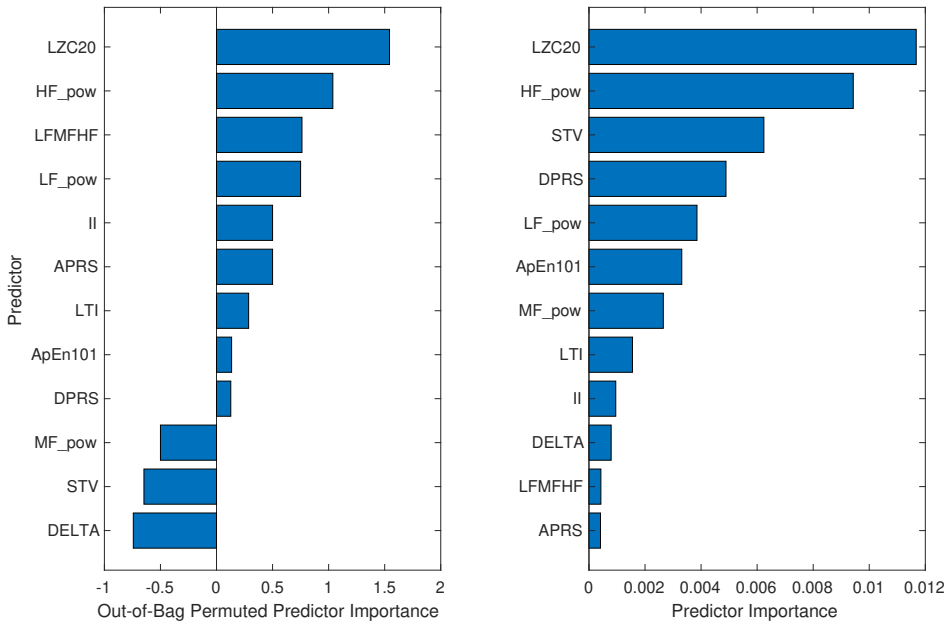


Figure 5.12: Out-of-Bag Predictor Importance (left) and Predictor Importance (right) of our Bagged Ensemble Model

5.3.2. Model Optimization and Implementation

Model optimization workflow :

Seeing the results of our predictor importance analysis, we decide to remove the parameters showing bad Predictor Importance. Hence, we train a model without the following parameters : **Delta, STV and MF_pow**. The new trained model shows really good results with an accuracy of 92% and a $AUC = 0.96$. On the other hand, this model fails to perform better in sensitivity (TPR = 90%) whereas this metric is our priority for the prediction model.

In order to increase the sensitivity and therefore decrease the false negative test. We decide to train the model with a new misclassification cost. Indeed, we set the misclassification cost of a IUGR fetus, beeing classified as a Healthy subject higher than misclassifying a Healthy subject. After several tests, we decide to set the misclassifiation cost t to 3 whereas the other is equal to 1.

		Predicted Class	
		Healthy	IUGR
True Class	Healthy	0	1
	IUGR	3	0

Figure 5.13: Misclassification cost matrix for the training of the final model

After training, the performance values of our model are bit lower than the ones of the previous model with a global validation accuracy of 87% and an $AUC = 0.95$ but the model shows really good sensitivity with only 4% of FN rate and so a TPR = 96%. Indeed, the model does not have as good performance in the specificity and so to detect Healthy patients with a 22% of FPR but as we said our priority stays to not miss the IUGR patients where the consequences could be really more dangerous. This model shows then interesting results for the IUGR detection application. The Confusion matrix of the model over the validation dataset can be seen in figure 5.14 and the ROC curve in the figure 5.15.

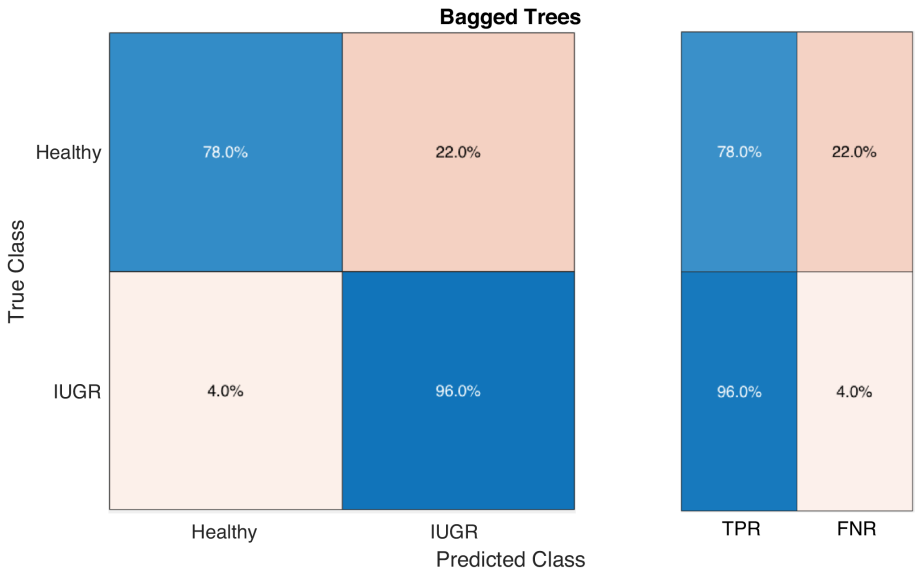


Figure 5.14: Confusion matrix of the final Bagged Ensemble model on our training/validation dataset

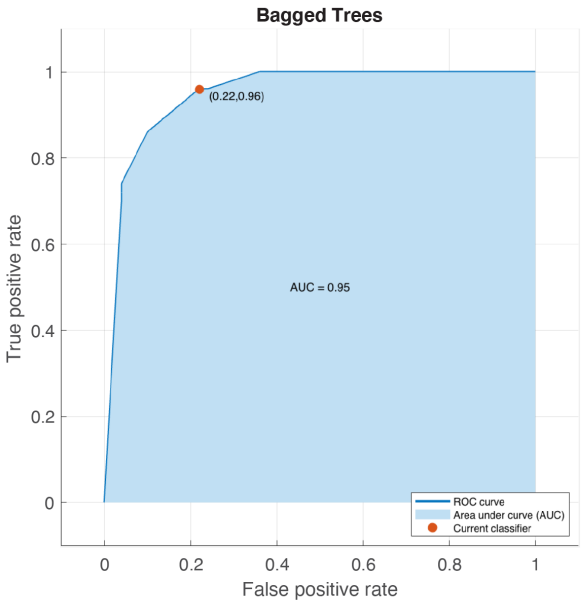


Figure 5.15: ROC curve of the final classification Bagged ensemble model on the training/validation dataset. The Area Under the Curve (AUC = 0.95) is the the sky blue area and our current classifier performance is the red dot

As we can see, even if this final model is not perfect in its performance, it shows relatively good results over the validation data for this application and with respect to the amount of data available. In the next section, we will study a bit more the details of the final model.

5.3.3. Final Model characteristics

In this section, the final model is analysed in more details. First of all, a look to the final predictor importance and the Out-of bag predictor importance will be made. A quick overview of the different weak decision trees will then follow in the analysis.

As explained earlier, the predictor importance shows the influence of each predictor in making a decision whether or not this forecast is accurate, whereas the Out-of bag predictor importance estimates by permutation how influential the predictor variables in the model are at predicting the response. Predictor importance of each features can be seen in the chart on the right of the figure 5.16. As we can see, the Lempel Ziv complexity stays the metric with the highest importance and is followed by the High Frequency power and the Approximate Entropy and then finally by the DPRS. Conversely, we see that the importance of the variability metrics (LTI and II) are quite low in our model. On the left side of the figure 5.16 we see that the frequency index $LF/(MF + HF)$ is the most important with the 2 complexity parameters and again that variability parameters are showing less importance in the classification.

In order to go further, we decide to plot the partial dependence of the sets of frequency and complexity parameters over the classification. Those plots show the proportionality of subjects

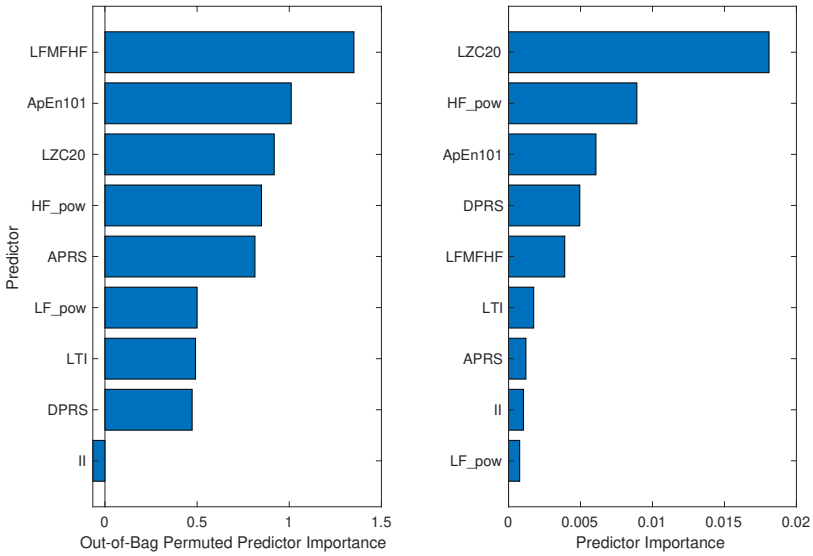


Figure 5.16: Out-of-bag Permuted Predictor (left) importance and Predictor Importance (right) of our final classification model

classified as IUGR with respect to the parameter values. The partial dependence plots can be seen in the figure 5.17.

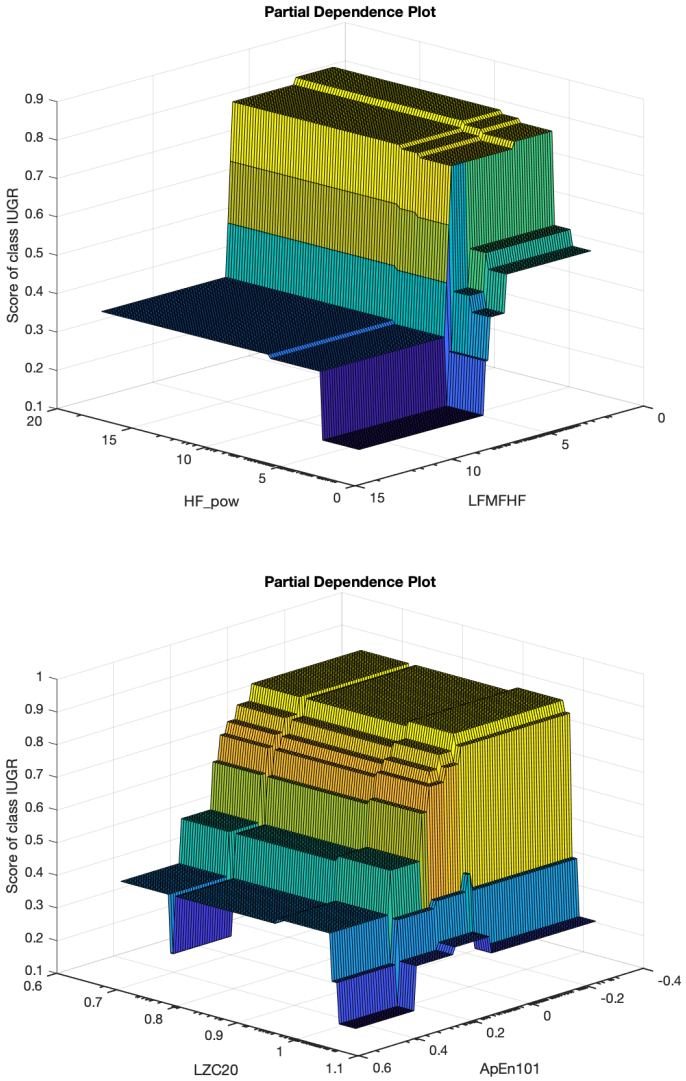


Figure 5.17: Partial dependence plots. The first plot shows the dependence of the frequency predictors HF_pow and $LF/(MF+HF)$. The second one shows the partial dependence of the complexity predictors $LZC(2,0)$ and $ApEn(1,0.1)$

It can be seen that both types have scores approaching 0.8 and show almost a monotonic relationship for both of the parameters. We also see that the slope of the limit is really steep for the parameter in the region around $LF/(MF+HF) = 5$ and $LZC20 = 1$ and that these parameters show higher dependence than the other parameter within the plot. This could explain why $LF/(MF+HF)$ have a higher Out of bag predictor importance and LZC a significantly higher predictor importance.

The 5 decision trees of our Bagged Ensemble can be seen in the Appendix B. Each decision tree is showed in different figures. We can see that most of the nodes in all the trees are frequency or complexity parameters. The highest node is either complexity or frequency parameters.

5.4. Final results and performance on Test set

Now that we have the final trained model, we can test it on the test dataset. Let's remind that the test data was defined randomly before any training. Therefore, this data is a set of 20 subjects totally independent of the training dataset. This way, testing the prediction of the model on this data will show us how our final model perform on new unknown data. This will give information on both the performance of the model and potential overfitting.

To test the model, test data is given in input for classification. This classification is then compared with the real retrospective annotation made by clinician. The confusion matrix on our 20 subjects test set is shown in the figure 5.18.

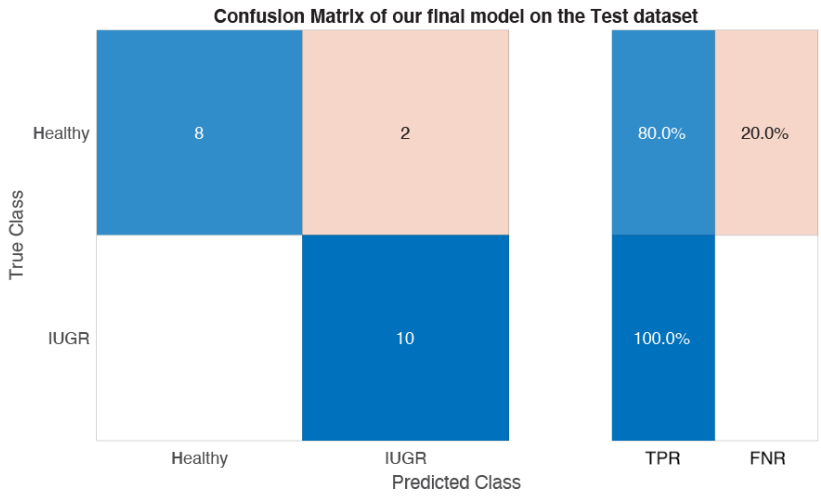


Figure 5.18: Confusion matrix of our final Bagged Ensemble model on our independent Test dataset

The results are shown in more details in the table 5.3. This table shows the prediction of the algorithm compared to the true state. The probabilities to be classified as Healthy or IUGR are also shown. These probabilities are computed according to the classification of each weak trees. Let's remind that the bagged ensemble model contains 5 decision trees and classify the subject according to the outcomes that is the most probable. An interesting point is that we see that for IUGR subjects, the model shows high probabilities for IUGR showing a high sensibility. Another interesting result is that the 2 misclassified subjects have a probability of 60% showing a lowest confidence in the classification.

The model shows really good results with a global accuracy of 90 % over the test data (only 2 misclassification over 20 subjects). Moreover, the results are coherent with the ones obtained on the cross-validation data. This shows that our simple Bagged Ensemble model is not showing overfitting over the training dataset with relatively same performance in sensitivity and specificity. Finally, the most promising result is that the model does not misclassify any IUGR subject and

then have a 100% sensitivity on the test set. Of course this result is only over 10 IUGR subjects and one could say that this should be taken into account. This will be developed a bit more into details in chapter 6 : Conclusion and Future developments.

True State	Prediction	Prob Healthy	Prob IUGR
Healthy	Healthy	0.8000	0.2000
Heathy	Heathy	1.0000	0
Healthy	Healthy	0.6000	0.4000
Healthy	IUGR	0.4000	0.6000
Heathy	Heathy	0.6000	0.4000
Healthy	Healthy	0.8000	0.2000
Healthy	Healthy	1.0000	0
Heathy	Heathy	0.8000	0.2000
Healthy	Healthy	0.8000	0.2000
Healthy	IUGR	0.4000	0.6000
IUGR	IUGR	0	1.0000
IUGR	IUGR	0.2000	0.8000
IUGR	IUGR	0	1.0000
IUGR	IUGR	0	1.0000
IUGR	IUGR	0	1.0000
IUGR	IUGR	0	1.0000
IUGR	IUGR	0	1.0000
IUGR	IUGR	0	1.0000
IUGR	IUGR	0	1.0000
IUGR	IUGR	0	1.0000

Table 5.3: Predictions and probability scores of the final model on the test set. The first column is the true retrospectively annotated state of the subject, the second is the prediction of our classification model. The third and 4th one are respectively the probability of one of the bagged ensemble tree to predict the subject to be Healthy or IUGR.

6 | Conclusions and future developments

6.1. Summary

Throughout this work, we have been interested in building a new framework for semi-automated analysis of CTG signals and use it to build a classifier using multiple parameters and able to support the Intra-Uterine Growth Restriction diagnosis for clinicians.

The first part of the work focused on CTG signals acquirement and specificity. We analysed the different signals presents in it, how to interpret them and which one could be of interest for our work. We also saw the limitations of the cardiotography signal acquirement.

After several discussions with clinicians and specialists in this field, we specialised our analysis into Intra-Uterine Growth Restriction. We studied the pathology, the causes and consequences. We saw among other things that the pathology is a documented cause of fetal and neonatal morbidity and mortality. After that, an overview of the steps followed by clinicians to diagnose it were seen. We found out several limitations in the current medical practice. Indeed, we saw that it fails to distinguish correctly SGA and true IUGR fetus. Moreover, the sensitivity and the specificity of the diagnostic do not have good results due to imperfect Ultrasound Imaging and empirical weight formula precision.

We focused our work on FHR signal analysis to give a prediction for IUGR subjects in antepartum. To do so, a set of parameters characterising signals of each subjects was implemented. This implementation was organized as follow :

- Pre-processing of raw FHR signals to remove too noisy and bad quality parts of the signal.
- Implementation of computation algorithm for Standard CTG parameters.

  Time domain variability parameters : Short-Term Variability (STV) , Interval Index (II), Delta, Long-Term Irregulatity (LTI).
  Frequency domain parameters : Low frequency power (LF_pow), Movement frequency power (MF_pow), High frequency power (HF_pow), frequency ratio index LF/(MF+HF).
- Implementation of computation algorithm for Non-Standard CTG parameters.

  Complexity parameters : Approximate Entropy (ApEn), Sample Entropy (SampEn), Lempel-Ziv Complexity (LZC).

Phase Rectified Signal Average parameters : Acceleration / Deceleration Capacity (AC / DC), Averaged Acceleration / Deceleration Capacity (AAC / ADC) , Acceleration/Deceleration Phase Rectified Slope (APRS/DPRS)

These features were used to build a table containing a set of parameters characterising each subject. This dataset allowed us to make a feature analysis on the data. We first analysed parameters dependency over Gestational Age. Dependency was found for the following parameters : Delta, STV, LTI, ApEn, APRS and DPRS. Adjustement by robust linear regression were then applied to these parameters in order to remove their GA dependency. After this, we compared the distribution of our adjusted parameters between the different datasets and analysed the potential impact of the measurement system used. A clear difference was found for the frequency parameters and for Lempel-Ziv Complexity.

Due to these measurement system differences, a dataset selection was made. Open-source data was finally preferred because of its bigger size compared to Polimi dataset and the fact that it has been correctly annotated by a retrospective study (which is not the case for Bloomlife data). In this dataset, the FHR signal are not published and only a defined set of parameters is accessible. Hence, the set of parameters was slightly reduced since Sampling Entropy, Acceleration/Deceleration Capacity and Averaged Acceleration/Deceleration Capacity are not available in Open-source dataset.

After that, we worked on the creation and training of a prediction model for IUGR detection. Several types of model were trained with our adjusted data. The models showed the following results :

- Linear Support Vector Machine (SVM) :
  Accuracy= 78% , TPR = 76% and SP = 80% , AUC = 0.89
- Medium K-Nearest Neighbours :
  Accuracy= 76% , TPR = 84% and SP = 68% , AUC = 0.88
- Decision Tree :
  Accuracy= 91% , TPR = 92% and SP = 90% , AUC = 0.89
- Bagged Ensemble Classifier :
  Accuracy= 90% , TPR = 90% , SP = 90%, AUC = 0.92

Finally, a deeper optimization was made for the Bagged Ensemble algorithm. After analysing the predictor importance, a feature selection was made by removing Delta, Short Term Variability and Movement frequency power in the model inputs. In addition to this, misclassification errors were modified to increase sensibility by improving the False negative error cost (IUGR subject classified as Healthy). Our final model has a global accuracy of 87% having a better accuracy on adjusted data than previous works (85.5%) by Signorini and reached 96% of sensitivity on our validation dataset. It shows also good results on the 20 subjects (10 Healthy / 10 IUGR) independent test set with a global accuracy of 90% (18/20) and a 100% sensibility (10/10).

6.2. Future developments

Limitations and potential improvements:

Several limitations in the work can be pointed. The first one comes from the pre-processing where a signal processing more similar to the one made by CTG machine could be implemented. To do so, a deeper study and analysis of the pre-processing step of the different CTG system should be made in order to finally have computation algorithm capable to work well in all the different devices.

The aim of this work was to build a framework for CTG semi-automated signal analysis and use it to give an additional metric to clinicians for IUGR detection. A set of parameters characterising the signals was obtained by our computation algorithms. Unfortunately, differences between dataset arised and the access to annotated data was limited. This issue lead our work to focus on a single dataset (Open-source) with only 120 subject failing in generality. An additional issue was that only pre-computed parameters were accessible accessible in the dataset forcing us to use only those parameters. Our model showed good results. However, it would have been interesting to have a classification process starting initially from the raw FHR signals. Different improvements could help to achieve this :

- Have access to a bigger set of data correctly annotated. Indeed, only 20 signals (from Polimi dataset) were correctly annotated in our data. This didn't allow us to train an efficient classification model starting from the FHR signals. A higher retrospectively annotated FHR signal dataset could help us to compute directly parameters from it and then use them as input to train our classification model.
- Have a better signal pre-processing before parameters computation. Indeed, we saw that the parameters were subjected to the measurement system difference leading to differences in the signals. A better pre-processing would potentially help us to use signals from different sources. To do so, a better understanding of the CTG signals acquirement should be acquired. In addition to this, stronger artefacts, re-sampling and noise processing functions should be implemented.
- Use only the parameters not subjected to the measurement system. This would reduce the amount of parameters included and therefore a bigger training dataset would be also needed in this case. Another solution would be to implement them in a different way such as implementing frequency parameters with an auto-regressive model as it has been done by *Signorini et al.* in 2003.[42]

In any case, a bigger dataset would be of interest to train a stronger model. Having a bigger dataset would also allow us to build a model with more decision trees without overfitting too much our data. Indeed, the actual model is composed of 5 weak trees of maximum 10 splits. A higher number of trees with more splits could potentially lead to a thinner prediction. Having more decision trees would also help us to have a more precised probability measure that could be used by the clinician.

Indeed, in addition to the global classification, it could be interesting for the clinician to have the IUGR classification probability in the prediction of the model in order to have an additional metric showing the confidence of the model in the prediction. This way, a clinician will not be as confident if the probability to be IUGR is of 60% compared to 90%. A metric like this will then increase the attention and the trust of the clinician for less confident predictions.

In addition to a bigger dataset, a larger set of parameter could also help the model to have a better prediction. First of all, the parameters computed by the algorithms that were not present in the open source dataset (SampEn and PRSA parameters except APRS and DPRS) but also baseline value and the number of accelerations and decelerations could also be parameters interesting for the model. Furthermore, as explained by *Dr Emonts*, growth restriction also decreases the movement of the fetus. The algorithms could then also use the signal relative to the movement of the fetus in addition to the FHR signal and use it as an input of our model.

Moreover, a deeper analysis of the trained classifier should be made in order to have a choice supported by statistical test. This would ensure us that the classifier choice is optimal. Additional test over other set of data could also be done in order to assess the performance on additional data coming from other source.

Potential application :

As explained throughout this work, a prediction model based on FHR signal could be used to help in the detection of IUGR in some cases :

- First of all, in some environments where Ultra-sound imaging is not easily accessible. FHR signal is a cheap and easy access test, this could help in the case where people don't have access to well-equipped hospital or clinic.
- Another application of this work would be also for the distinction of true pathological Intra Uterine Growth Restriction (IUGR) from physiological Small for Gestational Age (SGA) fetus. Indeed, current diagnostic is only based on the estimation of the weight by empirical formula based on metrics estimated by US imaging. The estimation is then followed by the comparison with a reference weight curve. This diagnostic actually detects SGA but is not sensible to the fact that the baby suffers from a growth restriction or is just simply physiologically small. In the same way, it could also be more sensible to detect baby showing normal weight but being in growth restriction (physiologically big fetus). Our model based on FHR parameters could add sensibility in the detection of IUGR. As suggested by *Dr Grandils* a further analysis would be to distinguish "Type 2" from "Type 1" IUGR and have a real detection of growth restriction due to placental inefficiencies that could be checked by a placental circulation echo-doppler check afterward.
- Another suggestion made by *Dr Grandfils* is to use it in a situation where clinicians could have some doubt about an IUGR fetus. In particular, later in the pregnancy (after 32 weeks)

where the US is not performing as well and the error in the weight estimation is higher. In those cases it is often complicated for clinicians to know if the fetus is in growth restriction and if the birth should be induced or not. In this case, the prediction model could add interesting information to know if the fetus as a nutriment deficiency or not and therefore had sensibility in the diagnostic.

Bibliography

[1] Out of bag predictor importance. URL https://nl.mathworks.com/help/stats/regressionbaggedensemble.oobpermutedpredictorimportance.html.

[2] 13.3 - robust regression methods: Stat 501. URL https://online.stat.psu.edu/stat501/lesson/13/13.3.

[3] Spearman's rank correlation coefficient rs and probability (p) value calculator. URL https://geographyfieldwork.com/SpearmansRankCalculator.html.

[4] FIGO consensus guidelines on intrapartum fetal monitoring: Cardiotocography - Ayres-de-Campos - 2015 - International Journal of Gynecology & Obstetrics - Wiley Online Library. URL https://obgyn.onlinelibrary.wiley.com/doi/10.1016/j.ijgo.2015.06.020.

[5] Support vector machine (svm) algorithm - javatpoint. URL https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm.

[6] Cardiotocography. Baby heartbeat monitor. Labour and delivery. URL https://patient.info/pregnancy/cardiotocography.

[7] Choosing the right estimator. URL https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html.

[8] Cardiotocography. URL https://www.wikiwand.com/en/Cardiotocography.

[9] Predictor importance, 2021. URL https://www.ibm.com/docs/en/spss-modeler/18.1.0?topic=SS3RA7_18.1.0%2Fmodeler_mainhelp_client_ddita%2Fclementine%2Fidh_common_predictor_importance.html.

[10] Decision tree learning, Jul 2022. URL https://en.wikipedia.org/wiki/Decision_tree_learning.

[11] Spearman's rank correlation coefficient, May 2022. URL https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient.

[12] Ensemble classifier: Data mining, Jan 2022. URL https://www.geeksforgeeks.org/ensemble-classifier-data-mining/.

[13] C. Amorim-Costa, D. A. de Campos, and J. Bernardes. Cardiotocographic parameters in small-for-gestational-age fetuses: How do they vary from normal at different gestational ages? a

study of 11687 fetuses from 25 to 40 weeks of pregnancy. *Journal of Obstetrics and Gynaecology Research*, 43(3):476–485, 2017. doi: 10.1111/jog.13235.

[14] A. Bauer, J. W. Kantelhardt, A. Bunde, P. Barthel, R. Schneider, M. Malik, and G. Schmidt. Phase-rectified signal averaging detects quasi-periodicities in non-stationary data. *Physica A: Statistical Mechanics and its Applications*, 364:423–434, May 2006. ISSN 03784371. doi: 10.1016/j.physa.2005.08.080. URL https://linkinghub.elsevier.com/retrieve/pii/S037843710501006X.

[15] S. Boudet, A. Houzé de l'Aulnoit, R. Demailly, L. Peyrodie, R. Beuscart, and D. Houzé de l'Aulnoit. Fetal heart rate baseline computation with a weighted median filter. *Computers in Biology and Medicine*, 114:103468, Nov. 2019. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2019.103468. URL https://www.sciencedirect.com/science/article/pii/S0010482519303403.

[16] S. Boudet, A. Houzé l'Aulnoit, R. Demailly, A. Delgranche, L. Peyrodie, R. Beuscart, and D. Houzé de l'Aulnoit. A fetal heart rate morphological analysis toolbox for MATLAB. *SoftwareX*, 11:100428, Jan. 2020. ISSN 2352-7110. doi: 10.1016/j.softx.2020.100428. URL https://www.sciencedirect.com/science/article/pii/S2352711018302498.

[17] E. Chandraharan. *Handbook of CTG Interpretation*.

[18] E. Chandraharan. *Handbook of CTG Interpretation*, volume 1. Cambridge University Press, 2017.

[19] P.-H. C. Chen, Y. Liu, and L. Peng. How to develop machine learning models for healthcare. 18(5):410–414. ISSN 1476-4660. doi: 10.1038/s41563-019-0345-0. URL https://doi.org/10.1038/s41563-019-0345-0.

[20] G. S. DAWES, C. R. HOUGHTON, and C. W. REDMAN. Baseline in human fetal heart-rate records. *BJOG: An International Journal of Obstetrics and Gynaecology*, 89(4):270–275, 1982. doi: 10.1111/j.1471-0528.1982.tb04695.x.

[21] I. C. Education. Bagging, May 2021. URL https://www.ibm.com/cloud/learn/bagging.

[22] A. Fanelli, G. Magenes, M. Campanile, and M. 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. doi: 10.1109/jbhi.2013.2268423.

[23] M. Ferrario, M. Signorini, and S. Cerutti. Complexity analysis of 24 hours heart rate variability time series. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 6:3956–9, Feb. 2004. doi: 10.1109/IEMBS.2004.1404105.

[24] F. P. Hadlock, R. B. Harrist, R. J. Carpenter, R. L. Deter, and S. K. Park. Sonographic estima-

tion of fetal weight. the value of femur length in addition to head and abdomen measurements. *Radiology*, 150(2):535–540, 1984. doi: 10.1148/radiology.150.2.6691115.

[25] K. Hajian-Tilaki. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation, 2013. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824/`.

[26] E. A. Huhn, S. Lobmaier, T. Fischer, R. Schneider, A. Bauer, K. T. Schneider, and G. Schmidt. New computerized fetal heart rate analysis for surveillance of intrauterine growth restriction. *Prenatal Diagnosis*, 31(5):509–514, 2011. doi: 10.1002/pd.2728.

[27] K. H. A. D. J. J. J. Wróbel, T. Kupka. Automated detection of fetal movements in doppler ultrasound signals versus maternal perception. *Journal of Medical Informatics Technologies. Vol.23*, pages 43–50, 2014.

[28] M. S. Kramer, M. Olivier, F. H. McLean, D. M. Willis, and R. H. Usher. Impact of intrauterine growth retardation and body proportionality on fetal and neonatal outcome. *Pediatrics*, 86 (5):707–713, 1990. doi: 10.1542/peds.86.5.707.

[29] A. Lempel and J. Ziv. On the complexity of finite sequences. *IEEE Transactions on Information Theory*, 22(1):75–81, 1976. doi: 10.1109/tit.1976.1055501.

[30] S. Longo, L. Bollani, L. Decembrino, A. Di Comite, M. Angelini, and M. Stronati. Short-term and long-term sequelae in intrauterine growth retardation (iugr). *The Journal of Maternal-Fetal amp; Neonatal Medicine*, 26(3):222–225, 2012. doi: 10.3109/14767058.2012.715006.

[31] A. Malliani, M. Pagani, F. Lombardi, and S. Cerutti. Cardiovascular neural regulation explored in the frequency domain. *Circulation*, 84(2):482–492, 1991. doi: 10.1161/01.cir.84.2.482.

[32] R. Mantel, H. P. van Geijn, F. J. 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, May 1990. ISSN 0020-7101. doi: 10.1016/0020-7101(90)90030-x.

[33] F. Marzbanrad, L. Stroux, and G. D. Clifford. Cardiotocography and beyond: A review of one-dimensional doppler ultrasound application in fetal monitoring, Aug 2018. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6237616/`.

[34] NHS Litigation Authority. *Ten years of maternity claims: An Analysis of NHS Litigation Authority Data*. 2012. ISBN 978-0-9565019-2-9. URL `http://www.nhsla.com/safety/Documents/Ten%20Years%20of%20Maternity%20Claims%20-%20An%20Analysis%20of%20the%20NHS%20LA%20Data%20-%20October%202012.pdf`. OCLC: 1179733161.

[35] C. on Obstetric Practice American Academy of Pediatrics, C. on Fetus, and Newborn. The apgar score, Oct 2015. URL `https://www.acog.org/en/clinical/clinical-guidance/committee-opinion/articles/2015/10/the-apgar-score`.

[36] S. Pincus. Approximate entropy (ApEn) as a complexity measure. *Chaos (Woodbury, N.Y.)*, 5(1):110–117, Mar. 1995. ISSN 1089-7682. doi: 10.1063/1.166092.

[37] Public Health Agency of Canada. Birth weight for gestational age, Jan 2004. URL `https://www.canada.ca/en/public-health/services/injury-prevention/health-surveillance-epidemiology-division/maternal-infant-health/birth-weight-gestational.html`.

[38] T. Quang. Calc_lz_complexity, 2022. URL `https://www.mathworks.com/matlabcentral/fileexchange/38211-calc_lz_complexity`.

[39] J. S. Richman and J. R. Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology. Heart and Circulatory Physiology*, 278(6):H2039–2049, June 2000. ISSN 0363-6135. doi: 10.1152/ajpheart.2000.278.6.H2039.

[40] M. W. Rivolta. Phase-Rectified Signal Averaging (PRSA), Aug. 2021. URL `https://github.com/MassimoWRivolta/PRSA`. original-date: 2019-07-22T15:32:51Z.

[41] D. Sharma, S. Shastri, and P. Sharma. Intrauterine growth restriction: Antenatal and postnatal aspects, 2016. URL `https://pubmed.ncbi.nlm.nih.gov/27441006/`.

[42] M. Signorini, G. Magenes, S. Cerutti, and D. Arduini. Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings. *IEEE Transactions on Biomedical Engineering*, 50(3):365–374, 2003. doi: 10.1109/tbme.2003.808824.

[43] M. G. Signorini and G. Magenes. Reliable nonlinear indices for fetal heart rate variability signal analysis. *2014 8th Conference of the European Study Group on Cardiovascular Oscillations (ESGCO)*, 2014. doi: 10.1109/esgco.2014.6847595.

[44] M. G. Signorini, N. Pini, A. Malovini, R. Bellazzi, and G. Magenes. Dataset on linear and non-linear indices for discriminating healthy and IUGR fetuses. *Data in Brief*, 29:105164, Apr. 2020. ISSN 2352-3409. doi: 10.1016/j.dib.2020.105164. URL `https://www.sciencedirect.com/science/article/pii/S2352340920300585`.

[45] P. N. M. A. B. R. a. M. G. Signorini, M. G. Integrating machine learning techniques and physiology based heart rate features for antepartum fetal monitoring. *Computer Methods and Programs in Biomedicine*, 185:105015, 2020. doi: 10.1016/j.cmpb.2019.105015.

[46] T. Stampalija, D. Casati, M. Montico, R. Sassi, M. W. Rivolta, V. Maggi, A. Bauer, and E. Ferrazzi. Parameters influence on acceleration and deceleration capacity based on transabdominal ECG in early fetal growth restriction at different gestational age epochs. *European Journal of Obstetrics, Gynecology, and Reproductive Biology*, 188:104–112, May 2015. ISSN 1872-7654. doi: 10.1016/j.ejogrb.2015.03.003.

[47] L. Stroux, C. W. Redman, A. Georgieva, S. J. Payne, and G. 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. doi: 10.1111/aogs.13228.

[48] D. V. Subramanian. How to read a ctg, 2011. URL https://geekymedics.com/how-to-read-a-ctg/.

[49] J. Szczepański, J. M. Amigó, E. Wajnryb, and M. V. Sanchez-Vives. Application of Lempel-Ziv complexity to the analysis of neural discharges. *Network (Bristol, England)*, 14(2):335–350, May 2003. ISSN 0954-898X.

[50] I. Walsh, D. Fishman, D. Garcia-Gasulla, T. Titma, G. Pollastri, J. Harrow, F. E. Psomopoulos, and S. C. E. Tosatto. Dome: Recommendations for supervised machine learning validation in biology, Jul 2021. URL https://www.nature.com/articles/s41592-021-01205-4#citeas.

[51] D. H. Willacy. Cardiotocography. baby heartbeat monitor. labour and delivery, 2021. URL https://patient.info/pregnancy/cardiotocography.

[52] H. A. Wollmann. Intrauterine growth restriction: Definition and etiology. *Hormone Research in Paediatrics*, 49(Suppl. 2):1–6, 1998. doi: 10.1159/000053079.

[53] X. Yang, Y. Wang, R. Byrne, G. Schneider, and S. Yang. Concepts of artificial intelligence for computer-assisted drug discovery. *Chemical Reviews*, 119(18):10520–10594, 2019. doi: 10.1021/acs.chemrev.8b00728.

A | Data Set

In this project, we worked with different datasets coming from different sources. This appendix explain the different datasets we worked.

A.1. Politecnico di Milano dataset

This dataset was given by *Politecnico di Milano* Institution and more especially by the research group of Pr *Maria Gabriella Signorini* working actively on antepartum CTG parameters. This dataset should stay confidential so it is not openly available. The signals have been acquired by a Hewlett Packard CTG machine with a sampling frequency f_s = 2Hz.

The dataset is composed for 10 IUGR subjects and 10 Healthy subjects of:

- patnum: the patient number
- FHR: The FHR signal vector (double) [bpm]
- QUALITA: The quality associated to each FHR sample. The quality can take 3 value: Good = 32 , Medium = 64, Bad = 96.
- TOCO: the tocograpgic signal assessig uterine contraction [mmHg]
- M1SO: the retrospective annotation: 0 = Healthy subject, 1 = IUGR subject
- eta = Age of the mother
- sett_gestazione = Week of gestation when the recording is made (GA)
- Distribution information of the parameters:
  STV , II, DELTA, LTI, FHRB (baseline) , APEN, LF, MF, HF, NUM_ACCEL_GRANDI (the number of big accelerations), NUM_ACCEL_PICCOLE (number of small accelleration) For these parameters, the value of the quartiles, the maximum, minimum, median and the mean of the distributions of the small computational segements are given.
- FHR120bpm = the FHR signal at $f_s = 2Hz$ interpolated in bad quality part by a mean averaged (5 samples).
- intdec = position of the potential decelerations in the signal
- FHR24bpm = FHR120bpm downsampled at 0.4Hz (24 samples /min)

A.2. Open-Source dataset

This dataset is the open-source dataset published in *Data in Brief* by *M.G Signorini et al.* [44]. The Dataset gives linear and non-linear indices for discriminating healthy and IUGR fetuses. It is composed of "12 linear and nonlinear indices computed at different time scales and extracted from Fetal Heart Rate (FHR) traces acquired through Hewlett Packard CTG fetal monitors (series 1351A), connected to a PC". [44]. The sampling frequency of the signal is $f_s = 2Hz$. The IUGR/Healthy state was retrospectively annotated by clinicians after birth.

The dataset is composed of parameters for 60 Healthy and 60 IUGR subjects. It contains 12 parameters for each subjects:

DELTA, II (Interval Indew), STV (Short Term Variability), LTI (Long Term Irregularity), LF_pow (Low Frequency power), MF_pow (Movement Frequency power), HF_pow (High Frequency power), LF/(HF+MF) (frequency index), ApEn(1,0.1) (Approximate Entropy), LZC(2,0) (Lempel-Ziv Complexity), APRS (Acceleration Phase Rectified Slope), DPRS (Deceleration Phase Rectified Slope).

It takes then the following form:

State	GA	DELTA	II	STV	LTI	LF	MF	HF	LF/HF+MF	ApEn(1,0.1)	LZC(2,0)	APRS	DPRS
'Healthy'	34	14.85	0.92	2.18	-0.22	82.52	15.12	2.36	4.72	-0.002	1.043	0.056	-0.063
'IUGR'	31	-5.05	0.88	-1.74	1.23	87.22	8.52	4.26	6.82	-0.011	1.025	-0.026	0.0434

A.3. Bloomlife dataset

Bloomlife data from their pilot study is also used. The signal were acquired by a Philips Avalon FM30 machine at a sampling frequency $f_s = 4$Hz. The digital signal was acquired by an additional Bloomlife device connecting the CTG machine to a PC and acquiring the signal digitally.

This data was unfortunately not annotated by clinician after birth to diagnose IUGR.The only information given is the classification made by clinician according to the last Ultra-sound imaging measurement. The subject are then only classified as 'Normal', 'Small', or 'Excessive'. No guarantee is given for the IUGR classification. Raw FHR signals were given for 113 subjects in which 12 were categorised as small for their gestational age (SGA). The GA of the recording is also given.

B | Appendix B: Final Classification Model

In this Appendix, the decision trees of our final bagged ensemble are shown. Triangles represent decision nodes, lines are branches and full circle are the leaves of the true representing the final classification.

B.1. Decision Trees of the Bagged Ensemble

B.1.1. Decision Tree 1:

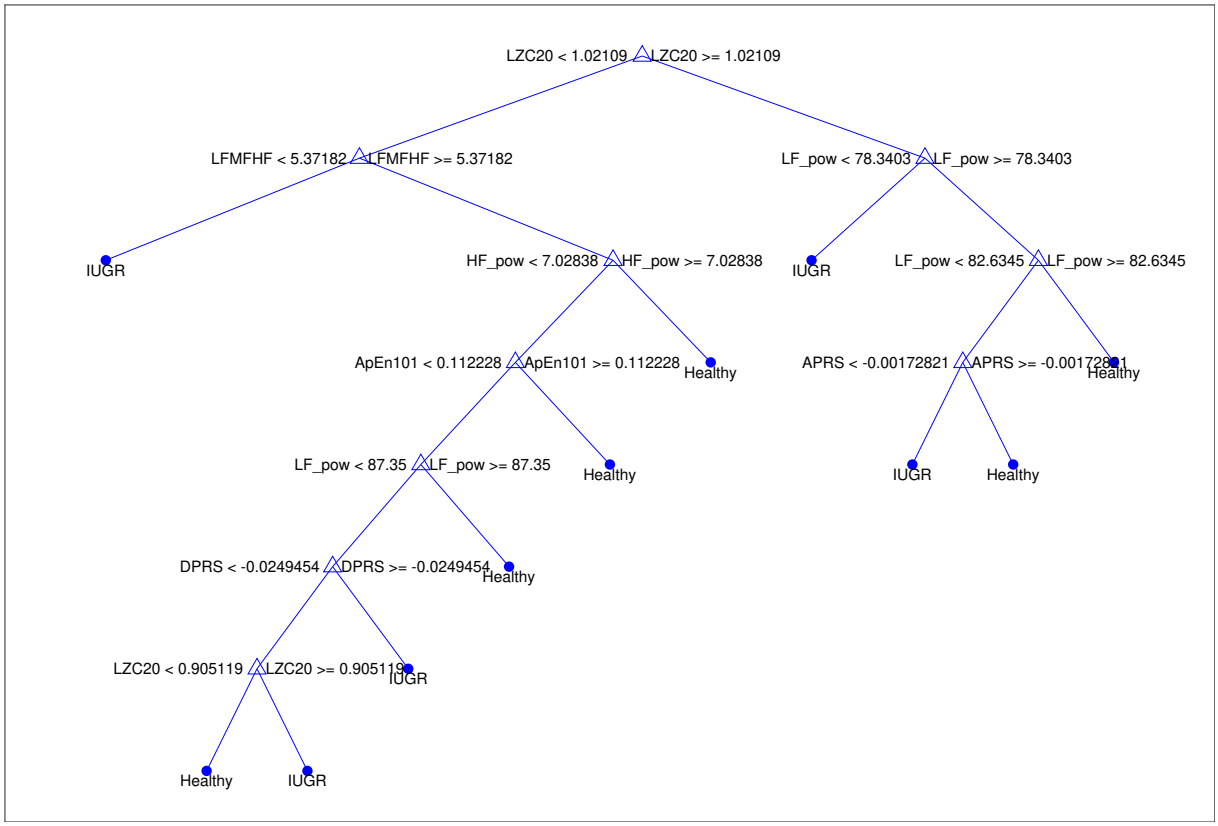


Figure B.1: 1rst weak decision tree of our Bagged Ensemble model

B.1.2. Decision Tree 2 :

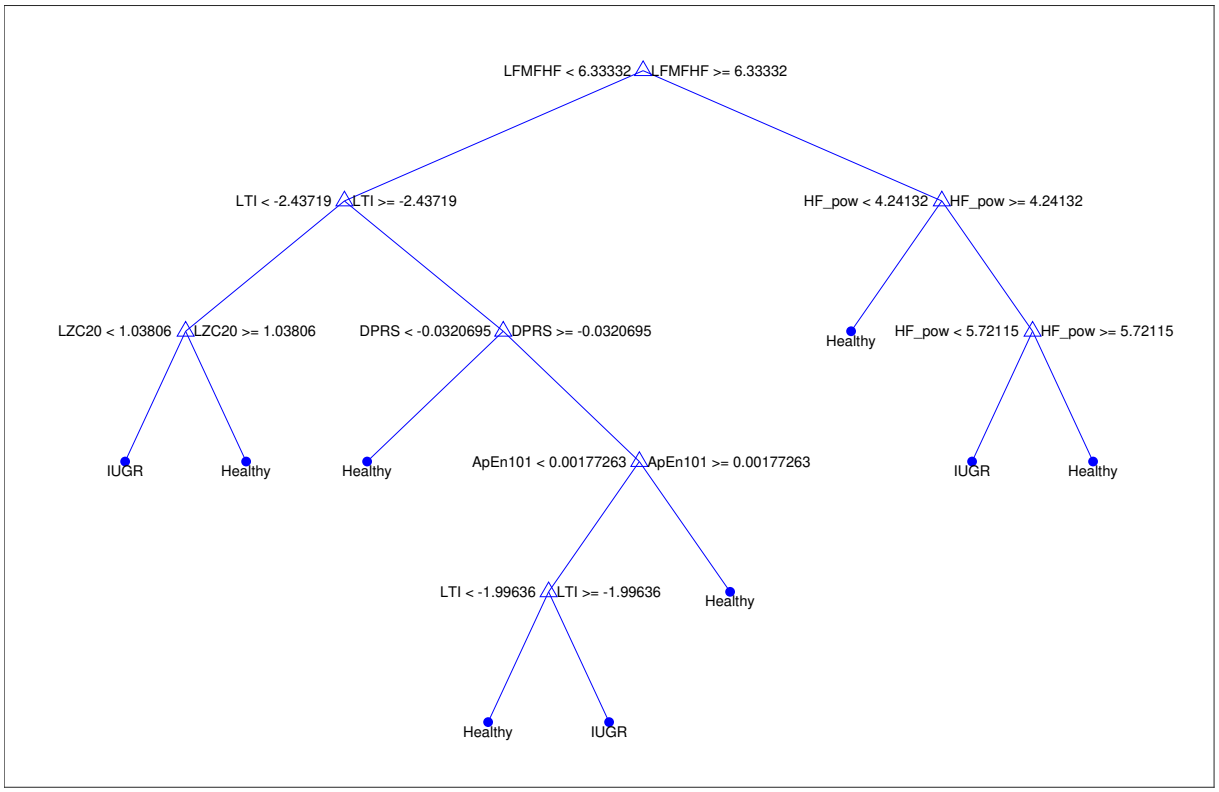


Figure B.2: 2nd weak decision tree of our Bagged Ensemble model

B.1.3. Decision Tree 3 :

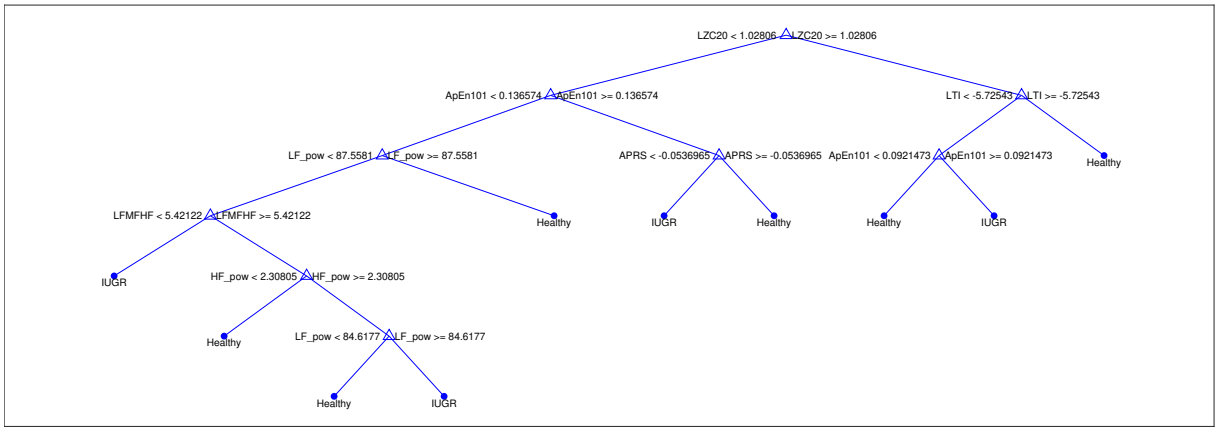


Figure B.3: 3rd weak decision tree of our Bagged Ensemble model

B.1.4. Decision Tree 4 :

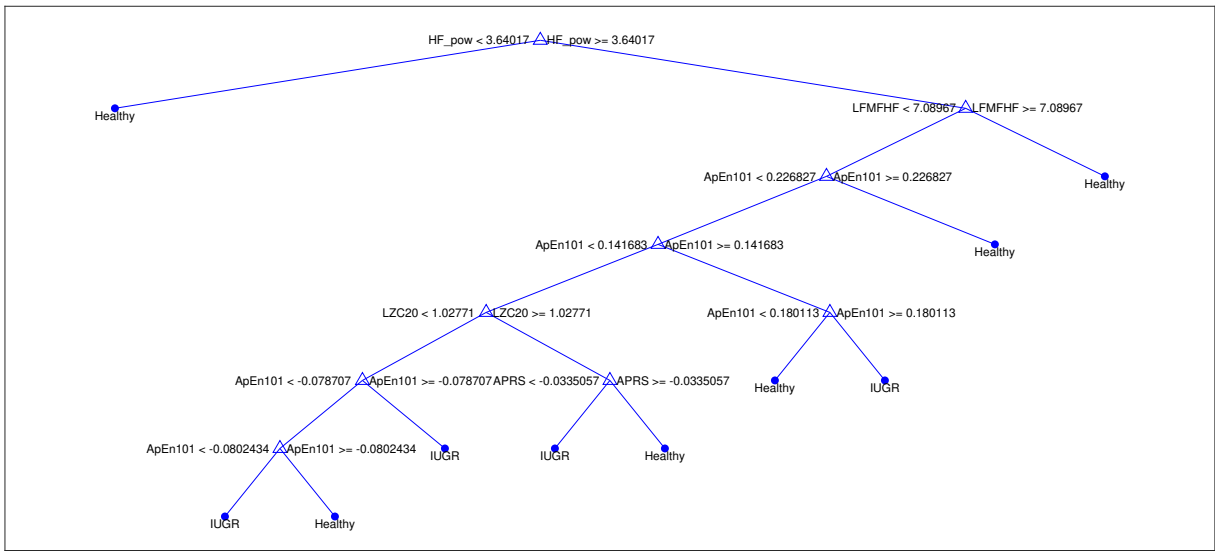


Figure B.4: 4th weak decision tree of our Bagged Ensemble model

B.1.5. Decision Tree 5 :

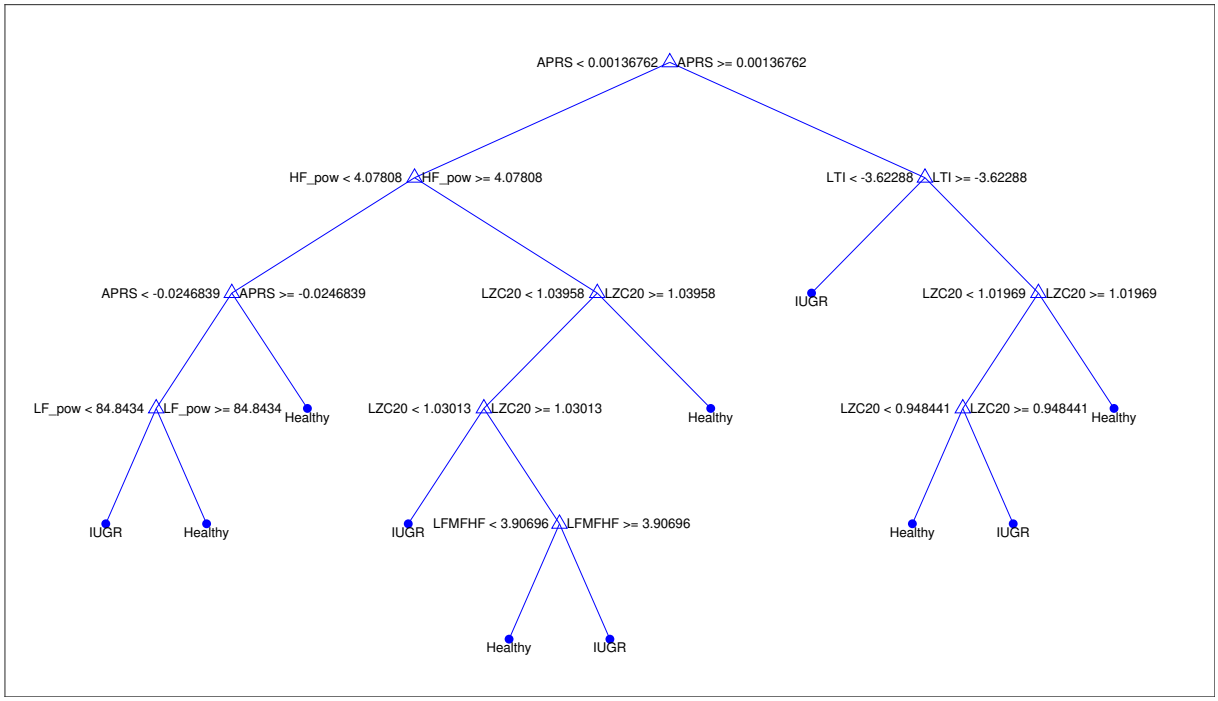


Figure B.5: 5th weak decision tree of our Bagged Ensemble model

List of Figures

List of Tables