UNIVERSITA' DEGLI STUDI DI PAVIA

FACOLTA' DI INGEGNERIA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

DOTTORATO DI RICERCA IN BIOINGEGNERIA E BIOINFORMATICA
XXXVI CICLO - 2023

ARTIFICIAL INTELLIGENCE SOLUTIONS FOR PROCESSING, ANALYZING AND CLASSIFYING FETAL HEART RATE VARIABILITY SERIES: A MULTI-PARAMETRIC APPROACH

PhD Thesis by

EDOARDO SPAIRANI

Advisor:
Prof. Giovanni Magenes

PhD Program Chair:
Prof. Silvana Quaglini

Sintesi.

Il presente elaborato raccoglie l'insieme delle ricerche condotte e dei risultati ottenuti nel corso del mio percorso di dottorato. Quest'ultimo si è svolto presso il laboratorio di Bioingegneria del dipartimento di Ingegneria Industriale e dell'Informazione dell'Università di Pavia, sotto la guida del Professor Giovanni Magenes.

Le attività svolte e raccolte nel presente elaborato rientrano nell'ambito del progetto PRIN ICT4MOMs, che mira a realizzare un sistema di monitoraggio intelligente del sistema madre-feto nel corso della gravidanza.

Il progetto è coordinato dalla professoressa Maria Gabriella Signorini del Politecnico di Milano.

Sebbene la gravidanza in sé non rappresenti un evento pericoloso per la vita del sistema madre-feto, l'antepartum rappresenta generalmente un periodo critico, che necessita, in quanto tale, di un attento monitoraggio per evitare esiti sfavorevoli.

In tale contesto si sviluppa il tema focale della presente tesi, che riguarda lo sviluppo e l'applicazione di soluzioni di intelligenza artificiale (AI) per l'elaborazione, l'analisi e la classificazione di serie di variabilità cardiaca (HRV), con una particolare attenzione al caso fetale.

I segnali di variabilità cardiaca fetale (FHR) rappresentano l'oggetto di studio della cardiotocografia (CTG), una pratica clinica non invasiva utilizzata per monitorare lo stato di benessere fetale nel corso della gravidanza.

Introdotta a partire dai primi anni '70, questa metodica ha avuto una notevole diffusione nella pratica clinica, tanto da rappresentare ad oggi lo standard di riferimento per il monitoraggio e la valutazione del benessere antenatale.

Diversi studi hanno dimostrato la validità della CTG in intrapartum, portando a concludere ragionevolmente che un pattern CTG “normale” sia da considerarsi un buon indicatore di benessere fetale, mentre un test “anormale” risulti in una scarsa predittività nei confronti della presenza di sofferenza fetale. Per contro, l’efficacia della CTG nell’identificazione della sofferenza fetale in antepartum presenta caratteristiche più controverse e criticità più difficilmente arginabili, risultando in un’accuratezza diagnostica subottimale. Il maggiore problema risiede nella difficoltà di lettura e interpretazione dei tracciati, ad oggi affidata alla sola esperienza clinica, e alla mancanza di un accordo generale circa i criteri di valutazione dei tracciati stessi. A partire dell’introduzione della CTG nella pratica clinica, sono state proposte numerose linee guida per la valutazione visiva dei tracciati CTG, nessuno dei quali si è però nettamente imposto come standard clinico. Per giunta, diversi studi hanno evidenziato che, anche in centri che adottano gli stessi criteri di valutazione, raramente i singoli osservatori concordano nell’interpretazione dello stesso tracciato. Numerose prove hanno evidenziato come la semplice ispezione visiva del tracciato CTG non consenta l’estrazione di tutte le informazioni di variabilità cardiaca contenute nel segnale FHR. Caratteristiche quali l’entità delle componenti periodiche del segnale generato dal pacemaker cardiaco, la non linearità del sistema di controllo della frequenza cardiaca fetale (FCF) o anche la stessa variabilità a breve termine non possono essere colte semplicemente analizzando ad occhio nudo il tracciato CTG. I limiti della CTG convenzionale risiedono, pertanto, nella difficoltà di lettura e interpretazione dei tracciati e nell’impossibilità dell’ispezione visiva di permettere di estrarre informazioni quantitative dai tracciati.

Le sempre crescenti possibilità offerte del supporto informatico e la necessità di identificare un nuovo approccio più oggettivo e replicabile hanno determinato l’avvento della cardiotocografia computerizzata (cCTG).

A partire dall’introduzione della cCTG, la ricerca si è focalizzata sull’identificazione di un unico indice quantitativo che rappresentasse un gold standard capace di descrivere adeguatamente il comportamento dei segnali di variabilità cardiaca fetale. Pur avendo contribuito a conferire riproducibilità al metodo e quindi a ridurre l'elevata variabilità intra e inter-osservatore, l’utilizzo di un unico regressore non ha sortito i risultati sperati, non consentendo di ottenere una valutazione adeguata circa il benessere fetale. Ciò ha spinto i ricercatori a capire che il problema dell'interpretazione dei tracciati CTG non potesse essere risolto con l'identificazione di un singolo indice, ma che, al contrario, dovesse basarsi su un approccio multivariato, che integrasse informazioni eterogenee raccolte attraverso diversi approcci metodologici. Questo ha aperto la strada a un’analisi in grado di fornire una visione a più ampio spettro dei tracciati CTG, in grado, pertanto, di consentire una descrizione di dinamiche (lineari e non) non facilmente ravvisabili ad occhio nudo.

L'integrazione dell'intelligenza artificiale (AI) nell'analisi dei tracciati CTG, ha il potenziale di rafforzare il potere analitico di questa metodologia. I modelli di AI hanno, infatti, la capacità intrinseca di scoprire schemi e tendenze nei dati che potrebbero passare inosservati con la semplice analisi visiva.

Negli ultimi anni, in particolare, l'attenzione della ricerca si è spostata verso l'impiego del Deep Learning (DL), in quanto considerato come l’approccio più promettente in questo campo. L’efficacia degli algoritmi di DL dipende però strettamente dalla disponibilità di ingenti moli di dati annotati e strutturati.

Tuttavia, attualmente, dataset CTG di grandi dimensioni non sono facilmente accessibili e solo pochi di essi, pur con un numero limitato di casi, sono pubblicamente disponibili online.

Il presente elaborato di tesi si erge su queste premesse e si pone l’obbiettivo di proporre alcune soluzioni che consentano di affrontare e superare alcune delle limitazioni che ad oggi inficiano le capacità diagnostiche della metodica cardiotocografica.

In particolare, il corpus è articolato nel modo seguente:

Il **Capitolo 1** fornisce un’introduzione alle serie di variabilità cardiaca, motivando l’utilità della loro analisi sia nel caso adulto che fetale e dettagliando gli step necessari ai fini dell'ottenimento delle stesse.

Il **Capitolo 2** mostra un caso di studio che affronta il tema della qualità dei segnali ECG, a partire dai quali possono essere ottenute, per l'appunto, le serie di variabilità cardiaca. in particolare, nel corso del capitolo viene proposto un metodo per inferire la qualità di un gruppo di segnali ECG raccolti tramite un dispositivo indossabile. La metodologia qui proposta nel caso di segnali provenienti da soggetti adulti può anche essere estesa al caso fetale.

Il **Capitolo 3** si addentra nel tema dell'analisi delle serie di variabilità cardiaca fetale, fornendo un'introduzione alla CTG e presentandone pregi, limiti e possibilità di sviluppo.

Il **Capitolo 4** è dedicato alla presentazione di un nuovo dataset CTG (NAPAMI) di grandi dimensioni, specificamente pensato per consentire l'applicazione di tutte quelle tecniche di AI, (in particolare di DL) le cui performance sono strettamente vincolate alla quantità di dati a disposizione.

Il **Capitolo 5** presenta due nuovi approcci metodologici utili ai fini della comprensione e dell'interpretazione delle dinamiche dei tracciati CTG sia nel dominio del tempo che delle frequenze. In particolare, La **Sezione 5.1** introduce un metodo innovativo, basato su un modello Markoviano nascosto, progettato per l'identificazione automatica delle fasi comportamentali del feto, a partire dai tracciati FHR, mentre nella **Sezione 5.2**, viene invece presentato un nuovo approccio per l'analisi spettrale dei tracciati.

Infine, il **Capitolo 6** affronta il tema della classificazione dei tracciati FHR, proponendo una nuova architettura neurale ad input eterogenei, che possa consentire di distinguere, con un buon livello di confidenza, una situazione di sofferenza fetale da una situazione fisiologica.

Abstract.

This document summarizes the research activities carried out as part of my doctoral program, which took place at the Bioengineering Lab of the Electrical, Computer and Biomedical Engineering Department of the University of Pavia, under the guidance of Professor Giovanni Magenes.

The performed activities are part of the Italian PRIN project ICT4MOMs, which aims at realizing intelligent solutions to monitor the mother-fetus system during pregnancy. The project is leaded by Prof. Maria G. Signorini (Politechnic of Milano).

The activities carried out and collected in this paper are part of the European PRIN project ICT4MOMs, which aims to realize an intelligent monitoring system of the mother-fetus system during pregnancy.

Although pregnancy itself does not represent a life-threatening event for the mother-fetus system, antepartum is generally a critical period, which needs, as such, careful monitoring to avoid unfavorable outcomes.

It is in this context that the focal theme of this thesis is developed, which concerns the development and application of artificial intelligence (AI) solutions for the processing, analysis, and classification of heart rate variability series (HRV), with a particular focus on the fetal case.

Fetal heart rate variability signals (FHR) are the object of study in cardiotocography (CTG), a noninvasive clinical practice used to monitor fetal well-being throughout pregnancy.

Introduced since the early 1970s, this method has been widely used in clinical practice, so much so that it is now the gold standard for monitoring and assessing antenatal well-being.

Abstract.

Several studies have demonstrated the validity of CTG in intrapartum, leading to the reasonable conclusion that a "normal" CTG pattern should be considered a good indicator of fetal well-being, while an "abnormal" test results in poor predictivity toward the presence of fetal suffering.

In contrast, the effectiveness of CTG in identifying fetal distress in antepartum has more controversial features and critical issues that are more difficult to curb, resulting in suboptimal diagnostic accuracy. The greatest problem lies in the difficulty of reading and interpreting the tracings, to date left to clinical experience alone, and the lack of general agreement about the criteria for evaluating the tracings. Since the introduction of CTG into clinical practice, numerous guidelines for the visual evaluation of CTG tracings have been proposed, none of which, however, has clearly established itself as a clinical standard. What is more, several studies have shown that even in centers adopting the same assessment criteria, individual observers rarely agree in their interpretation of the same tracing. Numerous trials have shown that mere visual inspection of the CTG tracing does not allow extraction of all the cardiac variability information contained in the FHR signal. Features such as the magnitude of the periodic components of the signal generated by the cardiac pacemaker, the nonlinearity of the fetal heart rate (FCF) control system, or even the short-term variability itself cannot be captured by simply analyzing the CTG trace with the naked eye.

The limitations of conventional CTG lie, therefore, in the difficulty of reading and interpreting the tracings and the inability of visual inspection to allow quantitative information to be extracted.

The ever-increasing possibilities offered by computer support and the need to identify a new, more objective and replicable approach led to the advent of computerized cardiotocography (cCTG).

Since the introduction of cCTG, research has focused on identifying a single quantitative index that would represent a gold standard capable of adequately describing the behavior of fetal cardiac variability signals. Although it helped to give reproducibility to the method and thus reduce the high intra- and inter-observer variability, the use of a single regressor did not yield the hoped-for results, failing to provide an adequate assessment about fetal well-being.

This prompted the researchers to realize that the problem of interpreting CTG tracings could not be solved by the identification of a single index but, on the contrary, should be based on a multivariate approach, integrating heterogeneous information collected through different methodological approaches. This paved the way for an analysis that could provide a broader spectrum view of CTG tracks, capable, therefore, of enabling a description of dynamics (linear and nonlinear) not easily discernible to the naked eye. The integration of artificial intelligence (AI) into CTG trace analysis has the potential to strengthen the analytical power of this methodology. Indeed, AI models have the inherent ability to uncover patterns and trends in data that might go unnoticed by simple visual analysis.

In recent years, in particular, research attention has shifted toward the use of Deep Learning (DL), as it is seen as the most promising approach in this field. The effectiveness of DL algorithms, however, is strictly dependent on the availability of large masses of annotated and structured data. However, currently, large CTG datasets are not readily accessible and only a few of them, albeit limited in number, are publicly available online.

The present thesis paper stands on these premises and aims to propose AI solutions to address and overcome some of the limitations that to date mar the diagnostic capabilities of the cardiotocographic method.

Specifically, the corpus is organized as follows:

Chapter 1 introduces heart rate variability series, motivating the usefulness of their analysis in both adult and fetal cases and detailing the steps necessary for the purpose of obtaining them.

Chapter 2 shows a study case, concerning adults but eventually extendable to the fetal case, that addresses the issue of the quality of electrocardiograms (ECG), from which cardiac variability series can be obtained; in particular, a method for inferring the quality of a group of

ECG signals collected via a wearable device is proposed throughout the chapter.

Chapter 3 delves into the topic of fetal heart rate variability series analysis, introducing CTG and presenting its merits, limitations and possibilities for development.

Chapter 4 is dedicated to the presentation of a new large CTG dataset (NAPAMI), specifically designed to allow the application of all those AI techniques, (particularly DL) whose performance is strictly constrained by the amount of available data.

Chapter 5 presents two new methodological approaches useful for understanding and interpreting the dynamics of CTG traces in both the time and frequency domains. Specifically, **Section 5.1** introduces a novel method, based on a hidden Markovian model, designed for automatic identification of fetal behavioral stages from FHR tracings, while in **Section 5.2**, a new approach for spectral analysis of tracings is presented instead.

Finally, **Chapter 6** addresses the issue of classification of FHR tracings, proposing a new neural architecture with heterogeneous inputs, which can allow to distinguish, with a certain level of confidence, a situation of fetal distress from a physiological situation.

Contents.

Chapter 1

A gentle introduction to heart rate variability (HRV) in adults and fetuses.

1.1. Introduction to HR and HRV: on the importance of their analysis.

The term heart rate (HR) is used to indicate the number of heart beats in a specified time interval (generally 1 minute). Heart Rate Variability (HRV), on the other hand, represents the amount of fluctuations around the average HR **[1.1]**. In other words, HRV can be seen as the variation in the time intervals between adjacent heart beats (see **Figure 1.1**).

A healthy heart does not adhere to a rigid metronome-like beat; on the contrary, the oscillations of a healthy heart's rhythm exhibit intricate and nonlinear patterns.

HRV can be viewed as both a reflection of the cardio-respiratory control system activity and a valuable tool for studying the sympathetic and parasympathetic functions of the autonomic nervous system (ANS). It serves as a reflection of neuro-cardiac functions, arising from intricate heart-brain interactions and the dynamic, nonlinear processes of ANS.

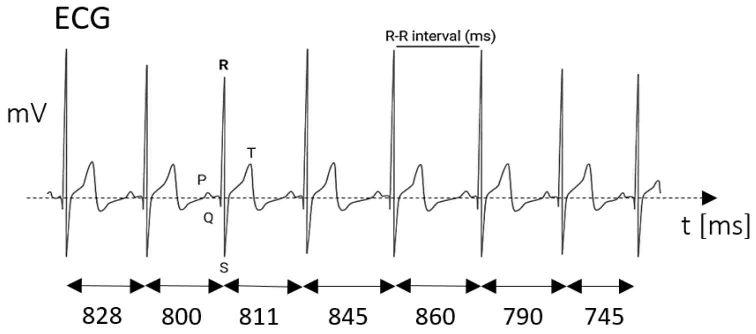


Figure 1.1: Typical healthy ECG trend with heart beats interspersed by time intervals with different durations.

The ANS communicates with the Cardiac System (CS) via the cardiac nerves, some of which originate from the cervical portion of the sympathetic nervous system, while the remaining come from the vagus nerve. Nerves from the sympathetic system have an excitatory role on the CS, leading to an increase in cardiovascular functions like heart rate (HR) and blood pressure (BP). Conversely, nerves from the vagus (or parasympathetic) system have inhibitory functions, working to decrease these cardiovascular activities (**Figure 1.2**). The interplay between the vagus and sympathetic systems is intricate and is known as the sympathovagal balance. Continuous changes in the sympathovagal balance result in fluctuations in HR around the average rate. These control mechanisms operate in opposition to one another, striving to rectify any disruptions introduced into the cardiovascular system which tend to alter heart rate. In adults, under normal conditions and at rest, this system responds to both external and internal stimuli, maintaining the heart rate at approximately 70 beats per minute **[1.2]**.

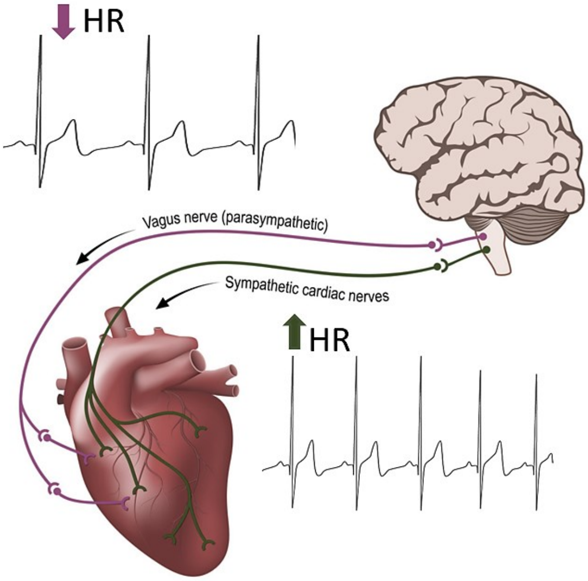


Figure 1.2: The autonomic nervous system (ANS), composed of the sympathetic and parasympathetic branches, plays a central role in regulating heart rate. The sympathetic nervous system tends to increase heart rate (a "fight or flight" response), while the parasympathetic nervous system decreases heart rate (a "rest and digest" response).

Furthermore, besides cardiac activity, HRV mirrors the regulation of blood pressure (BP), gas exchange, gastrointestinal function, vascular tone (which governs BP by adjusting blood vessel diameter), and potentially even facial muscle activity **[1.3].**

The fluctuations in a healthy heart's beat-to-beat intervals find their most accurate description in the realm of mathematical chaos **[1.4]**. This variability endows the heart with the capacity to swiftly respond to an uncertain and ever-changing environment. However, it's important to note that a higher HRV is not universally indicative of better health, as pathological conditions can also give rise to elevated HRV. On the other hand, optimal HRV levels are linked to overall health, self-regulatory capacity, and adaptability or resilience. Elevated levels of resting vagally-mediated HRV are associated with enhanced performance in executive functions such as attention and emotional processing, which are regulated by the prefrontal cortex **[1.5]**. Observing the behavior of the sympathovagal balance through the analysis of HRV signals allows for the monitoring of patients with heart disorders as well as with conditions involving disruptions in the functions of the autonomic nervous system. These illnesses, in fact, by altering the dynamic equilibrium of the sympathovagal balance, render the modulation of the heart's natural pacemaker less effective and modify the HRV signal spectrum compared to that of a healthy individual.

Hence, HRV is a source of important information regarding the general health status and its monitoring is of great interest both in the case of adult and fetus. As a matter of fact, as detailed in **Chapter 4**, many quantitative indexes have been presented in the literature, trying to provide a reliable description of autonomic functions. However, sympathovagal interplay is a complex phenomenon, that has not been proven to be accurately described by a unique index.

Measurements of heart rate variability are simple to perform, non-invasive, and therefore widely employed in clinical practise.

The next sections are dedicated to HRV series and will detail the necessary steps to obtain the latter both in the case of adult subjects (**Section 1.2**) and in the fetal case (**Section 1.3**).

1.2. Construction of HRV series in adults.

In adults, HRV series can be obtained through different approaches, each of which makes use of different sensors. Asides from the electrocardiogram (ECG), which represents the gold-standard, different methodologies have been developed to estimate HRV series and include wearable garments (as will be treated in **Chapter 2**) and devices, such as smartwatches and fitness trackers, which rely on diverse kinds of technologies, as optical sensors, bioelectrical impedance-based sensors and even textile ones **[1.6]**. Obviously the precision in estimating HRV series can vary substantially among these methods, accordingly to the embedded systems and the choice of the most suitable approach depends on specific needs and intended use.

Regardless from the employed technology, the goal is being able to spot heart beat events with a certain degree of accuracy, since the more precise is the beat event localization, the more reliable are the computed parameters. The most accurate method to estimate HRV series is electrocardiogram, which makes use of electrodes to record the heart's electrical activity. The obtained ECG signals present a quasi-periodic trend, characterized by the repetition of typical patterns (**Figure 1.3**), each of which comprehends a P wave (resulting from atrial depolarization), a QRS complex (associated with atrial repolarization and ventricular depolarization), and a T wave (attributed to ventricular repolarization).

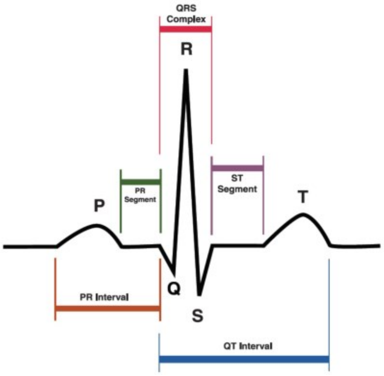


Figure 1.3: Typical P-QRS-T complex of an ECG signal. It comprehends a P wave (resulting from atrial depolarization), a QRS complex (associated with atrial repolarization and ventricular depolarization), and a T wave (attributed to ventricular repolarization).

The first step towards obtaining HRV series is the precise localization of R peaks, which represent the local maxima within each QRS complex.

The identification of R peaks can be achieved through various approaches, some of which rely, for example, on the computation of the Hilbert transform **[1.7]** or on some other digital filtering methods **[1.8]** or again on pattern recognition **[1.9]**.

The most classical and vastly employed method for the detection of R peaks is the Pan-Tompkins algorithm **[1.10]**. The latter comprehends a cascade of digital elaborations, the first step of which includes bandpass filtering (with a 5-50Hz bandpass) to remove unwanted noise and baseline wander. Then, the filtered ECG is differentiated, squared and integrated in moving windows to emphasize the high-frequency components and hence highlight the QRS complex peaks. A dynamic threshold is set to distinguish QRS complex peaks from other parts of the ECG signal. This threshold is often determined based on the local maximum of the integrated signal. Peaks above the threshold are considered as QRS complex detections. Last, the algorithm identifies the R-wave peak of each QRS complex, which represents the most prominent point in the ECG waveform.

Once R peaks are detected, the temporal distance between consecutive R peaks (RR interval) is computed, for each couple of consecutive R peaks.

This allows for obtaining the tachogram, which shows the trend of RR intervals (generally in milliseconds) over time.

The RR interval time series exhibits irregular sampling intervals due to variations in the durations of consecutive heartbeats. The representation of R-R intervals as a function of time poses challenges, particularly in frequency-domain analysis. To mitigate this issue, various strategies have been employed prior to spectrum analysis. One approach involves calculating the power spectrum directly from the RR interval time series, which is available as a function of the beat index. However, in this method, the spectrum is not a function of frequency; instead, it depends on cycles per beat **[1.11]**. Another approach is to resample the RR interval time series using different interpolation techniques, such as spline interpolation. The aim is to evenly distribute the non-uniformly sampled RR intervals, making them equally spaced. The third approach utilizes an integral pulse frequency modulation (IPFM) model. The IPFM method employs delta functions that represent a series of impulses occurring at the times of heartbeats **[1.12]**.

Regardless from the adopted strategy, HRV series can be obtained straightly from tachogram by applying **equation 1.1**, which converts the RR series in beats per minute (bpm).

$$\boldsymbol{HRV(bpm) = \frac{60}{RR(s)}} \qquad (1.1)$$

In ideal conditions, RR interval time series are solely consisting of pure sinus beats. However, these series are in most cases imperfect, due to several kinds of artefacts, which can represent a significant problem affecting the reliability of HRV series.

Being able to stand if a HRV series can accurately represent the real fluctuations of heart beats can’t help but evaluating the quality of collected ECG signals. This means assessing whether P-QRS-T complexes are well maintained or either if noise and artefacts cause their deterioration and the loss of their typical trends.

Chapter 2 addresses this issue and illustrates a novel method, specifically designed to automatically identify ECG signal artifacts, and hence to provide a numerical quantification of the global quality of collected ECGs.

1.3. Construction of fetal HRV series.

For what concerns, instead, the fetal case, the extraction of HRV series in fetuses (FHR) can be obtained through diverse methodologies.

In clinical practice, the golden standard for fetal heart rate monitoring is based on the Doppler ultrasound (US) **[1.13]**.

Continuous monitoring of the fetal heart rate (FHR) both before and during labor is achieved by securing a US transducer to the maternal abdomen. This continuous FHR monitoring, along with simultaneous monitoring of uterine activity, is referred to as cardiotocography (CTG), as will be detailed in Chapter 2. In contrast, when intermittent measurements of the FHR are needed, a handheld Doppler US transducer is typically employed **[1.14]**.

Doppler ultrasound imaging operates by emitting an ultrahigh-frequency sound wave beam (exceeding 20 kHz) that is aimed at the fetal heart. This beam is directed through an ultrasound probe positioned on the mother's abdomen. Ultrasound waves transmitted through the maternal skin and underlying subcutaneous tissue travel through the uterine muscle, the amniotic sac containing amniotic fluid, and reach the fetal heart. During each cardiac cycle, both atria and ventricles undergo a diastolic and systolic phase, which cause the rhythmic contraction and dilation of the heart walls. These movements alter the frequency components of US echoes coming back to the transducer. In fact, when a US wave encounters a moving target, the latter changes its frequency accordingly to **equation 1.2** (Doppler effect):

$$\boldsymbol{fs = \frac{2fv\cos(\theta)}{c}} \qquad (\mathbf{1.2})$$

where fs represents the frequency shift, *vcos(θ)* is the velocity of the moving target (the fetal heart walls contracting and dilating) along the direction of the US beam and c is the c the speed of sound in the tissue (m/s). In general, if the target approaches the transducer, the received frequency increases (i.e., *fs* > 0), while it decreases when the target moves away (**Figure 1.4**).

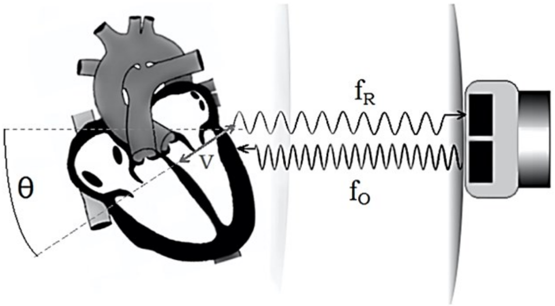


Figure 1.4: Visual representation of US Doppler effect. When US waves encounter a moving object, their frequency components are altered according to the ratio between the speed of the moving object and the one of the propagating US waves.

After detecting the envelope of the Doppler signal (commonly performed through the computation of Hilbert transform), the autocorrelation function is computed to estimate the number of beats per minute and hence obtain the FHR. What makes the US Doppler method the most employed approach in the antepartum is its robustness to extract FHR from signals with low signal to noise ratio (SNR) together with its total lack of invasiveness. On the other hand, the drawbacks of this method include the necessity for skilled clinicians or nurses to accurately position the ultrasound probe on the maternal abdomen to detect the fetal heart rate. Additionally, it is not suitable for prolonged FHR monitoring due to its high sensitivity to both fetal and maternal movements, which often necessitates frequent probe repositioning and can lead to ambiguous records with respect to accelerations and decelerations. Maternal movements can introduce Doppler-shifted reflected waves, which may be stronger than the actual cardiac signal. Moreover, the equipment is bulky, and monitoring must take place in a hospital setting. Lastly, monitoring sessions are infrequent during pregnancy, requiring pregnant women to travel to the hospital for these examinations **[1.15]**.

Besides US Doppler, another possible solution to estimate FHR is abdominal ECG (AECG). AECG involves the placement of electrodes directly on the maternal abdomen.

These electrodes capture faint electrical signals originating from the fetal heart, enabling noninvasive FHR assessment during pregnancy (**Figure 1.5**).

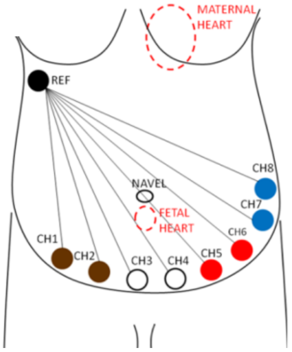


Figure 1.5: Scheme of AECG electrodes disposition onto the maternal abdomen. The latter lead to record electrical signals originating from the fetal heart, enabling noninvasive FHR assessment during pregnancy.

AECG can be employed to extract the Fetal Electrocardiogram (FECG), identify fetal QRS complexes, and generate HRV data. However, the recorded signal is a blend of various sources, including the desired FECG signal and numerous interfering noises, such as the maternal ECG, alternating current (AC) interference, motion artifacts, muscle activity, and others. To record these electrical signals, various electrode technologies and configurations can be employed. The primary advantage of AECG is its noninvasiveness and unobtrusiveness, making it suitable for extended monitoring periods, even up to 24 hours. AECG is also preferred over US CTG, in case of pregnant women with BMI > 30 kg/m^2. Additionally, AECG devices have modest power requirements and can be designed to be compact and portable. This method allows for the detection of both maternal ECG and heart rate. Signal processing algorithms enable the extraction of both fetal and maternal ECG signals. In comparison to US Doppler, AECG offers supplementary insights into fetal cardiac activity, as FECG contains pathophysiological details not present in HRV series obtained through Doppler. Furthermore, in AECG, it is still possible to derive HRV after identifying QRS complexes.

Nonetheless, a notable drawback of this technique is that acquiring a reliable FECG signal can be challenging and, in many instances, quite demanding. Furthermore, the signal-to-noise ratio (SNR) is often quite low due to the numerous sources of interference. Consequently, the use of noise reduction and FECG extraction algorithms is imperative to address these challenges.

When uncertain fetal heart rate patterns are observed during labor, Internal Fetal Monitoring (IFM) is employed. This method involves the insertion of an electrode at the tip of a catheter through the cervix, positioning it just beneath the fetal scalp's skin (**Figure 1.6**). The electrode directly records the Fetal Electrocardiogram (FECG) and transmits this signal to a recording device via a wire. Since the internal fetal monitor is directly connected to the baby, the Fetal Heart Rate (FHR) signal is notably clearer and more consistent than what is typically obtained with an external monitoring device. However, it's essential to acknowledge a minor risk of infection associated with internal monitoring. Additionally, the scalp electrode may leave a mark or a small cut on the baby's head, though these typically heal swiftly.

It's important to note that internal fetal monitoring can only be initiated after the rupture of the fetal membranes, meaning it is employed exclusively during labor **[1.15]**.

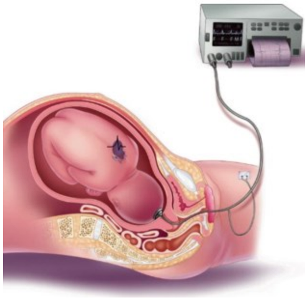


Figure 1.6: Visual representation of Internal fetal monitoring. This method involves the insertion of an electrode at the tip of a catheter through the cervix, positioning it just beneath the fetal scalp's skin and collects the FECG signal.

Other existing technologies include the INVU from Nuvo Cares **[1.16]**, a wireless, wearable device containing passive electrical and acoustic sensors which fuses ECG and audio signals to furnish a reliable estimation of fetal HRV; another important technology is Magnetocardiography (mECG) which leads to high quality FHR signals at the expense of costs and instrumentation size **[1.17]**.

1.4. References for Chapter 1.

[1.1] R.McCraty, F.Shaffer,. "Heart rate variability: new perspectives on physiological mechanisms, assessment of self-regulatory capacity, and health risk", Glob Adv Health Med, 2015 4:46–61. 10.7453/gahmj.2014.073.
[1.2] M.J.Capilupi, S.M.Kerath et al., "Vagus Nerve Stimulation and the Cardiovascular System", Cold Spring Harb Perspect Med. 10(2):a034173, 2020 doi: 10.1101.
[1.3] R.N.Gevirtz, P.M.Lehrer et al., "Cardiorespiratory biofeedback", 4th ed In: Schwartz MS, Andrasik F, editors. Biofeedback: A Practitioner's Guide. New York: The Guilford Press, p. 196–213, 2016.

[1.4] A.Goldberger, "Is the normal heartbeat chaotic or homeostatic?" News Physiol Sci 6:87–91, 1991.
[1.5] R.D.Lane, E.M.Reiman, "Activity in medial prefrontal cortex correlates with vagal component of heart rate variability during emotion". Brain Cognit 47-97, 2001.
[1.6] P.Castiglioni, P.Meriggi et al., "Heart Rate Variability from Wearable Photoplethysmography Systems: Implications in Sleep Studies at High Altitude". Sensors doi: 10.3390/s22082891, 2022.
[1.7] D.Benitez, P.A.Gaydecki et al., "The use of Hilbert transform in ECG signal analysis". Comput. Biol. Med. 31, 399–406, 2001.
[1.8] S.A.Israel, J.M.Irvine et al., "ECG to identify individuals", Pattern Recognit. 38, 133–142, 2005.
[1.9] S.Mehta, S.S. Saxena "Computer-aided interpretation of ECG for diagnostics", Int. J. Syst. Sci. 27, 43–58, 1996.
[1.10] J.Pan and W.J.Tompkins, "A real time QRS detection algorithm", IEEE Trans. Biomed. Eng. 32, 230–236, 1985.
[1.11] R.W.DeBoer, J.M.Karemaker, "Comparing spectra of a series of point events particularly for heart rate variability data", IEEE Trans. Biomed. Eng. 31, 384–387,1984.
[1.12] R.W.DeBoer, J.M.Karemaker, "Spectrum of a series of point events, generated by the integral pulse frequency modulation model", Med. Biol. Eng. Comput. 23, 138–142, 1985.
[1.13] P.Hamelmann, R.Vullings et al., "M. Doppler Ultrasound Technology for Fetal Heart Rate Monitoring: A Review", IEEE Trans Ultrason Ferroelectr Freq Control. 2020 Feb;67(2):226-238. doi: 10.1109/TUFFC.2019.2943626, 2019.
[1.14] R. Martis, O. Emilia, D. S. Nurdiati, and J. Brown, "Intermittent auscultation (IA) of fetal heart rate in labour for fetal well-being," Cochrane Database Syst. Rev., no. 2, pp. 1–68, 2017, Art. no. CD008680. doi: 10.1002/14651858.CD008680.pub2.
[1.15] M.Hasan, M. Reaz, M. Ibrahimy, M. Hussain, and J. Uddin, "Detection and processing techniques of FECG signal for fetal monitoring," Biological procedures online, vol. 11, no. 1, pp. 263–295, 2009.

[1.16] M.Mhajna, N.Schwartz et al., “Wireless, remote solution for home fetal and maternal heart rate monitoring”, American Journal of Obstetrics & Gynecology MFM, Volume 2, Issue 2, 2020.

[1.17] J.F. Strasburger, B. Cheular et al., “Magnetocardiography for fetal arrhythmias”, Rhythm. Jul;5(7):1073-6. doi: 10.1016, 2008.

Chapter 2

Case of study: automatic ECG signal quality assessment in Mountain Rescuers through the computation of Sample Entropy.

As anticipated in **Chapter 1**, the quality of the acquired ECG signal is pivotal in obtaining a precise and dependable estimation of HRV series. An ECG of high quality offers clean and interference-free data, which is indispensable for an accurate comprehension of HRV and its implications for cardiac health and well-being. As explained in the previous paragraph, HRV relies on a precise identification of R-R intervals to measure time intervals between consecutive heartbeats. If the ECG exhibits disturbances or artifacts, these discrepancies might be incorrectly interpreted as variations in heart rate, leading to an inaccurate estimation of HRV. For instance, a low-quality ECG may include electrical interferences or body movements that need to be rectified or removed before analysis. Artefacts can either

affect the whole ECG or be localized at some particular signal portions. However, the automatic identification of distorted portions of ECG signal is a nontrivial task.
In this chapter we present the developed method to automatically assess the quality of a set of ECG signals collected through a wearable device in typical mountain rescuers activities. ECGs signals have been obtained during sessions of programmed field tests at the Bormio Ski Resort (Valtellina, Lombardy, Italy) in the month of March 2022. Here, following the defined protocol, a group of 15 mountain rescuers has carried out daily rescuers' activities, while wearing a wearable textile system. The test protocol was designed to simulate the real physiological demands of mountain rescuers during their emergency deployments. Among the performed activities, rescuers had to walk up and down hill in snow-covered trails and carrying stretchers onto which simulated victims were located. To infer the quality of ECG signals recorded we developed an algorithm for the automatic evaluation of collected signal deterioration, which is based on the analysis of regularity of ECGs' P-QRS-T complexes pattern.

2.1. Background.

In recent years, wearable devices became a consolidated reality and their use has been extended to always wider fields of application. In particular wearables providing physiological signals are highly diffused, not only for medical application but, nowadays are becoming central, as source of data, in more complex systems that interact with human being in each area of his daily activity (work, home, sport.). Reliability of the signals acquired is fundamental since these signals are often used to trigger advice, alerts and feedbacks coming from such integrated system to the final user. Even if the overall quality of the signals offered by wearable devices increased considerably, there are intrinsic difficulties such as motion artifacts due to the nature of these devices. Moreover, signals measured by wearable devices in daily life activities result much more challenging due to uncontrolled environment and the consequent noise that this condition can bring. For this reason, a good signal quality assessment (SQA) results fundamental for a functional use and diffusion of wearable devices.

In this context, we present our work, describing the development of an automatic ECG SQA, capable to verify the quality and acceptability of the ECG signal acquired by a wearable textile system a sensing vest and a data logger. This study has been carried out in the frame of EU funded project SIXTHSENSE **[2.1]**.

Analysis have been conducted on a subset of the data collected during the field trials held in Bormio (Italy) to verify the functionalities of the first prototype delivered. In particular, it has been analyzed the ECG signal acquired by the sensing vest during operation designed to simulate the real physiological demands of mountain rescuers during their emergency deployments. The SQA developed aims to quantify the global quality of ECGs recorded, in terms of the percentage of conservation of P-QRS-T complexes. To achieve this result, the ECG signal has been decomposed using the *sym*4 mother wavelet, then a frequency-localized version of the ECG waveform has been reconstructed, only using the wavelet coefficients at scales 4 and 5, in order to maximize the QRS energy. The estimation of the regularity of ECG patterns is obtained evaluating the Sample Entropy (SampEn) (see the **Appendix** section for more details) of the WT signal obtained from the squared absolute value of each coefficient composing the sequence. Low values of SampEn relate to high regularity, while high values indicate irregular or noisy signal portions. Furthermore, SQA has been completed evaluating the correlation between the intensity of the activity with the quality of ECG signal.

2.2. The employed wearable system.

The data analyzed in this study have been acquired by a wearable textile system by Smartex Srl. The system, based on Wearable Wellness System (WWS) **[2.2]**, has been re-designed and customized to meet SIXTHSENSE project requirements in which several sensors have to be embedded in an integrated wearable prototype. The platform used for this study, part of the ALFA version of the SIXTHSENSE prototype, is composed by a garment, equipped with electrodes for ECG acquisition and a piezoresistive sensor for breath measurement, and a data logger (**Figure 2.1**).

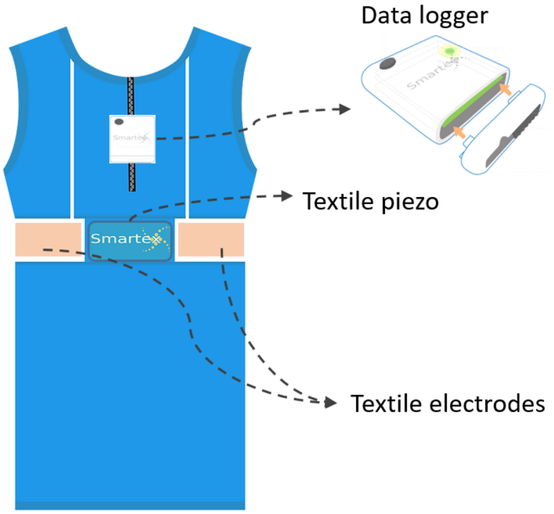


Figure 2.1: Illustration of the employed WWS. The vest is equipped with two textile electrodes to collect the ECG signal, one piezoresistive textile sensor to measure respiratory signals and a data logger (RUSA device).

The physiological signals provided by the system can be grouped in the following categories:

1. **ECG measurements**: One lead ECG (with a sampling frequency of 250 Hz), and derived Heart Rate, Heart Rate Variability, and R-R interval.

2. **IMU measurements**: The system is equipped with one IMU with 9 degrees of freedom (DOF) placed on the chest. Besides the IMU raw signals additional information are provided on activity intensity

3. **Respiration measurements**: The piezoresistive sensor placed on the thorax is used to measure the strain on the thorax caused by the participant's breathing.

2.3. Participants and experimental protocol.

Fifteen mountain rescuers from different countries (Serbia, Bosnia, and Italy) took part in the study (age, 31.7±8.6 yr.; body mass, 78.2±12.6 kg, and height, 178.1±7.7 cm). The study was performed according to the recommendations of the Helsinki Conference for research on humans and was approved by the Ethics Committee of the University of León, Spain. Written informed consent was obtained from all subjects before starting the tests.

All participants performed a field test at the Bormio Ski Resort (Valtellina, Lombardy, Italy) in the month of March 2022. The test was designed to simulate the real physiological demands of mountain rescuers during their emergency deployments. For this, the participants executed the most common activities during their deployments **[2.3]**. Specifically, the mountain rescuers had to walk uphill until they reached a point where a simulated victim was located (~75 kg). After that, the victim was evacuated on a stretcher

(~12 kg) to the base of the ski resort. The starting and ending test point was at an altitude of about 2,000 m. Likewise, the victim was located at 2,263 ± 61 m. The test was performed on snowy terrain at an environmental temperature and relative humidity of 6.9 ± 2.1 ºC and 17.1 ± 11.6%, respectively. The test was performed in crews of 3−4 members, each crew used the techniques of movement on snow most used by them (i.e., using crampons or mountain skis). The mean duration of the test was ~80 min (~40 min ascent, ~15 min rescue and ~25 min descent). Participants were encouraged to move at a speed consistent with an actual deployment and allowed rest stops if required. Rescuers wore their standard team clothing (i.e., cold protective trousers and jackets over a cotton T-shirt; helmet, gloves and boots) and each carried personal equipment during the whole field test. Finally, each subject was equipped with different devices to measure heart rate: a commercial heart rate monitor (RS800, Polar Electro Oy, Kempele, Finland) and the ALPHA version vest. The size of this garment was chosen before starting the test to fit the shape and size of the participants.

2.4. The proposed automatic ECG SQA method.

The proposed method is designed to numerically assess the quality of ECG signals, based on the regularity of the patterns composing the latter, regardless from the way signals are recorded.

As anticipated at the beginning of **Section 1.2**, in ideal conditions, ECG signals present a quasi-periodic trend, characterized by the recurrence of typical patterns (**Figure 2.2 a**). A normal ECG pattern contains a P wave (due to atrial depolarization), a QRS complex (due to atrial repolarization and ventricular depolarization) and a T wave (due to ventricular repolarization). However, due to many factors, ECG signals are often corrupted by different kinds of noise (e.g., baseline wander, muscle artefact, power line interference, instrumentation noise, etc.) **[2.4]**, that alter the regularity of its normal patterns, making it difficult to identify P-QRS-Ts complexes (**Figure 2.2 b**).

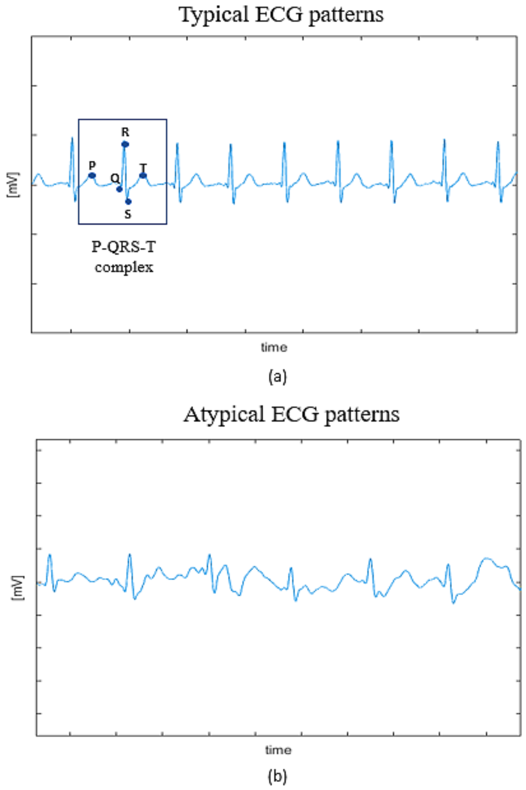


Figure 2.2: **(a)** Regular ECG patterns characterized by the repetition of identifiable P-QRS-T complexes. **(b)** Atypical ECG patterns resulting from different kinds of artefacts.

Since our goal was to quantify the global quality of ECGs recorded, in terms of the percentage of conservation of P-QRS-T complexes, we developed an algorithm for the automatic evaluation of ECG signals' regularity based on the computation of Sample Entropy (SampEn) **[2.5]** on sliding windows. In particular, the first step composing the aforementioned method is the computation of the maximal overlap discrete wavelet transform (MODWT) of the ECG tracing. The idea is to use wavelets as templates **[2.6]** for a process of pattern matching in order to enhance the typical ECGs patterns respect to noisy atypical ones. First, we proceeded to decompose the ECG using the sym4 mother wavelet, exploiting the similarity of the latter with the P-QRS-T complex (**Figure 2.3**).

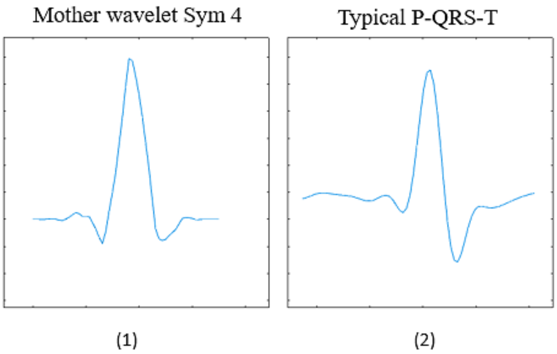


Figure 2.3: Mother wavelet Sym4 **(1)** with its typical shape resembling a typical ECG P-QRS-T complex **(2)**.

Then, to maximize the QRS energy and to improve signal to noise ratio, we reconstruct a frequency-localized version of the ECG waveform, only using the wavelet coefficients at scales 4 and 5. Finally, the computation of the squared absolute value of each coefficient composing the sequence, gives back the signal (WT), used to estimate the regularity of ECG patterns. For this aim we moved a sliding window of 5 seconds-length with a stride of 1 second along the WT signal and, at each shift, we computed the SampEn. The latter is a measure of the amount of regularity and predictability of fluctuations of time series data. Low values of SampEn relate to high regularity, while high values indicate unpredictable behaviors. In the specific case of study, low values of SampEn describe parts of ECG where P-QRS-T patterns are well maintained, while atypical patterns are related to high levels of this parameter. The resulting series of SampEn indices obtained is then resample at the same sample frequency of ECG (fs = 250 Hz).

By analyzing ECGs in relation to SampEn series, we identified two values of threshold through which we could define three levels of signal quality that we refer to as green, yellow, and red. Green quality is related to parts of the signal where P-QRS-T complexes are perfectly identifiable while yellow and red ones are respectively related to portions where complexes are still present but distort and to parts where signal is totally corrupted by noise, such that P-QRS-T are completely impossible to spot. The process discussed so far is shown in **Figure 2.4**, which illustrates all the steps that, starting from a row 70-minutes ECG, bring to the localization, on the ECG signal itself, of good, middle, and low quality points, respectively depicted as green, yellow, and red ones.

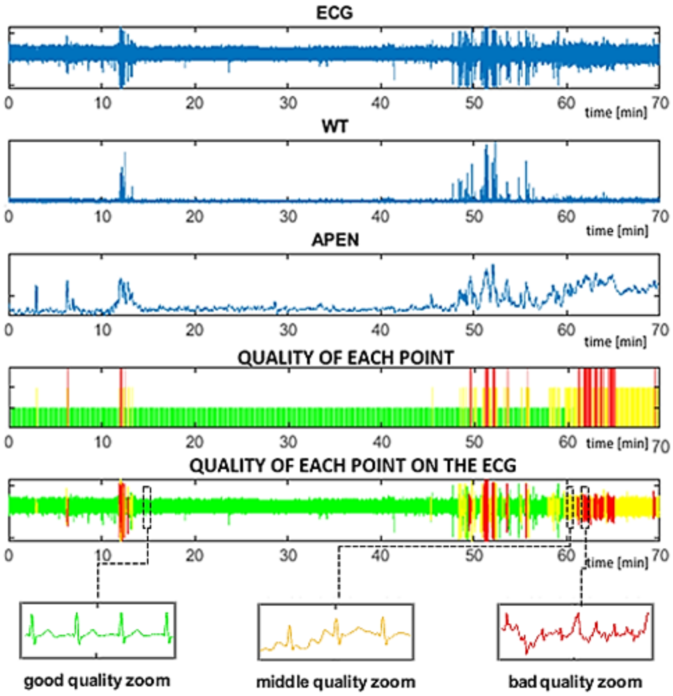


Figure 2.4: steps that, starting from a row 70-minutes ECG, bring to the localization, on the ECG signal itself, of good, middle, and

low quality points, respectively depicted as green, yellow, and red ones. From the top to the bottom, we have the ECG signal, the WT signal, the signal obtained after the SampEn computation, the classification of each point in good (green), middle (yellow) and bad quality (red), and the same classification on the ECG signal.

2.5. Achieved results.

After testing the overall ECG signal quality obtained, regardless from the activity carried out, we evaluated the quality maintenance while performing high intensity activities as compared to low intensity ones.
To evaluate the global quality of the ECGs recorded by the system, we computed the average value of the percentage of good, middle, and low-quality points on the 15 ECG signals recorded, regardless from the intensity of the activity performed.
The obtained results are summarized in the pie chart in **Figure 2.5**.

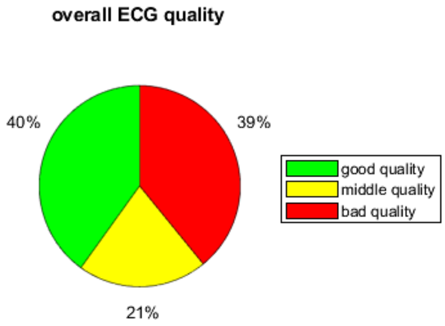


Figure 2.5: Pie chart showing the average value of the percentage of good, middle, and low-quality points on the 15 ECG signals recorded, regardless from the intensity of the activity performed.

From the analysis of the obtained results it appears that good and bad quality points are well balanced (40% green – 39 % red) while 21% of points are of middle quality. This means that globally, approximately 60% of points in the analyzed ECGs present a sufficient maintenance of P-QRST-complexes, leading to the estimation of heart rate.

After evaluating the global quality of the recorded signals, we proceeded to compare the average of the percent ECG signal quality while performing high and low intensity activities.

WWS system provides information on the intensity of the activity performed by the subject wearing it. This estimation is based on the evaluation of Signal Magnitude Area (SMA) **[2.7]**, the most extended feature used to measure the physical activity, obtained from the accelerometer signals coming from the IMU embedded in the WWS electronic. SMA is expressed through **equation 2.1** as:

$$\boldsymbol{SMA} = \sum_{\boldsymbol{i}=1}^{\boldsymbol{N}} (|\boldsymbol{x}(\boldsymbol{i})| + |\boldsymbol{y}(\boldsymbol{i})| + |\boldsymbol{z}(\boldsymbol{i})|) \qquad \textbf{(2.1)}$$

where $x(i)$, $y(i)$ and $z(i)$ indicate the values of respectively x-axis, y-axis, and z-axis acceleration signals and N is the length of the sliding window (N=5,000 in our case).

The wearable system provides several levels for the energy/intensity of activity but for the purpose of this work this feature has been binarized to distinguish low and high intensity activity and correlate this information with possible degradation of ECG signal quality.

Intensity Energy value signal was resampled at the same sample frequency of the ECG (fs = 250 Hz) and then, after defining a reasonable threshold value, was binarized to identify points relating to low and high intensity activities (**Figure 2.6**). After the energy related to each point of the tracing was determined, we computed the average value of the percentage of good, middle, and bad-quality points on the 15 ECG signals analyzed, stratified by intensity (high/low) of the activity carried out.

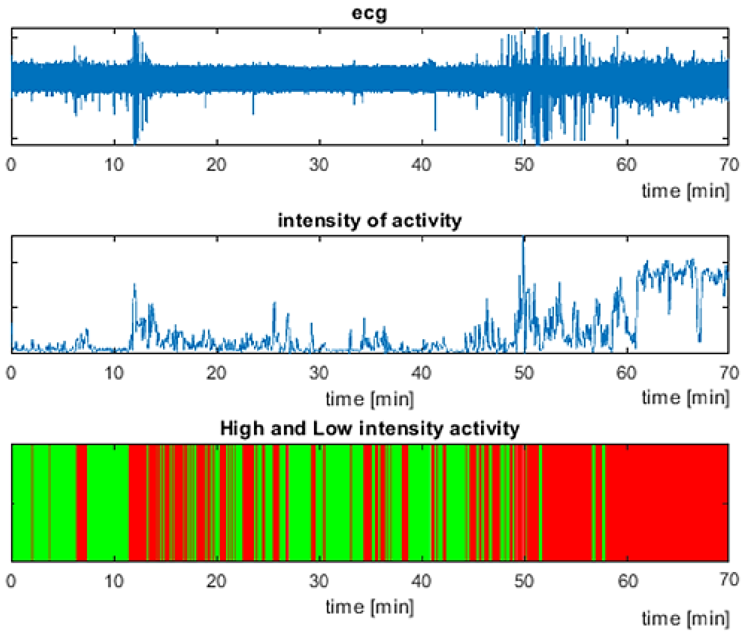


Figure 2.6: From the top to the bottom: ECG signal, Intensity Energy value signal, Classification of points in time in High and Low intensity of activity.

The obtained results are summarized in **Figure 2.7**. By comparing the pie charts relating to high and low intensity points, we can state a strong dependency of the ECG signal quality from the intensity of the activity performed. In fact, the percentage of good quality points, in low energy case, almost triples the ones in high energy condition (59% vs 21%). The mirror image situation is for bad quality points. In fact, as we can observe, the percentage of bad quality points in the low energy case is about one third of the one in the high energy case. The percentage of middle quality points is instead balanced between the two cases.

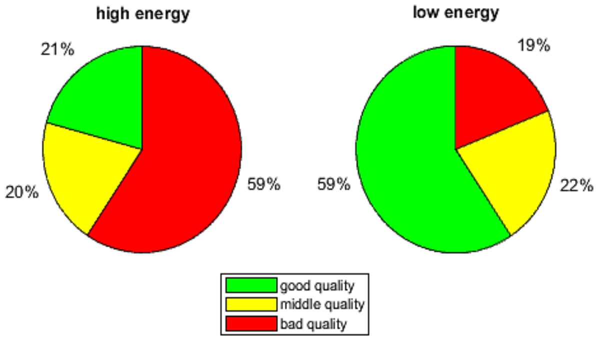


Figure 2.7: Pie charts quantifying the average of the percentage of good, middle, and bad-quality points on the 15 ECG signals analyzed, stratified by intensity (high/low).

2.6. Discussions and Conclusions for the proposed method.

In this chapter we have presented the development of a method for the automatic assessment of ECG signal regularity based on the computation of SampEn on fixed length sliding windows.

This SQA approach was employed to evaluate the average quality of 15 ECGs recorded during scheduled field tests, in order to infer the ability of wearable system to register ECGs of sufficient quality, in terms of maintenance of P-QRS-T complexes, both in stress and non-stress conditions.

From the analysis of the obtained results, it appears that the average percentage of acceptable portions of ECG signals recorded by WWS system is about 60% of signal length.

If we separately analyze the system behavior under both high and low intensity regimes, we can observe an important degradation of ECG signal quality, when a high intensity activity is performed. In fact, while the maintenance of P-QRS-T complexes acceptable integrity is about 80% in the case of a low energy activity, it approximately reduces of a half, when a high intensity activity is carried out.

The gathered results confirm the difficulties to acquire reliable data in the field. Harsh and uncontrolled condition and activity of a certain intensity can imply sudden degradation of the signal quality. In this particular application an added source of artifacts relies on the use of heavy clothes and equipment (i.e., bulky backpack) of the mountain rescuers that affects the performance of the wearable device.

Lastly, it's worth to observe that the presented automatic SQA approach is not uniquely applicable to the specific case illustrated in the chapter but can also be extended to any possible context where it's necessary to infer the acquired ECGs' quality (both local and global). For example, this SQA method could even be employed to integrate and better the estimation accuracy of systems computing HR.

2.7. References for Chapter 2.

[2.1] https://sixthsenseproject.eu

[2.2] R. Paradiso, L. Caldani. "Electronic Textile Platforms for Monitoring in a Natural Environment,"Research Journal of Textile and Apparel, vol. 14, no. 4, pp. 9-21, 2010.

[2.3] N.Callender, J.Ellerton et al., "Physiological demands of mountain rescue work". Emerg. Med. J. 29 (9), 753–757. https://doi.org/10.1136/emermed-2011-200485, 2012.

[2.4] R.Panda, U.C.Pati, "Removal of Artifacts from Electrocardiogram Using Digital Filter", IEEE Students' Conference on Electrical, Electronics and Computer Science, pp. 1-4., 2012, doi:10.1109/SCEECS.2012.6184767.

[2.5] S.M.Pincus, "Approximate entropy as a measure of system complexity", Proceedings of the National Academy of Sciences of the United States of America, 88(6), 2297–2301, 1991, https://doi.org/10.1073/PNAS.88.6.2297

[2.6] Z.Wang, Z.Junjiang et al., "A new modified wavelet-based ECG denoising", Computer Assisted Surgery, 24:sup1, 174-183, 2019, doi: 10.1080/24699322.2018.1560088

[2.7] W.Chung, A.Purwar et al. " Frequency domain approach for activity classification using accelerometer", 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 1120-1123, 2008. doi: 10.1109/IEMBS.2008.4649357

Chapter 3

Analysis of fetal heart rate variability series.

3.1. Introducing cardiotocography.

In developed countries, all mothers are submitted to medical examinations to monitor fetal wellbeing throughout pregnancy. Although most pregnancies proceed physiologically, complications affect approximately 8% of the total ones **[3.1]**. These might arise due to adverse mother's health conditions, thus leading to various medical issues, further impacting the health of both the mother and fetus. The negative impact on the fetus health is usually referred to as "fetal distress", which is strictly linked to alterations in the FHR signal.

The most employed diagnostic examination in the clinical practice to assess the fetal health during pregnancy is cardiotocography (CTG). In Italy, during pregnancy, each woman can undergo three or even more ambulatory CTG monitoring tests. In case mothers are diagnosed with a condition that makes the pregnancy high-risk, the screening occurrence can be scheduled on a weekly or even daily basis. Since the introduction of the first commercial cardiotocograph in Europe by Hammacher in 1968, there has been a significant scientific effort and tremendous enthusiasm directed toward a method that appeared to offer a tangible means of understanding when and why fetal health deterioration occurred, thereby enabling more effective intervention **[3.2]**.

CTG monitoring has gained widespread use in clinical practice since the 1970s. It remains the most widely employed method for monitoring the ongoing clinical condition of the fetus, despite the emergence of newer techniques such as Doppler velocimetry and ultrasound.

Conventional CTG is a non-invasive method which monitors the fetal wellbeing by analyzing two different tracings, i.e., the fetal heart rate (FHR) and uterine contractions (TOCO) signals. The latter are recorded simultaneously by means of two different instruments which are placed on the abdomen of the mother to be (see **Figure 3.1**). The first US transducer is positioned near the fetal back and is used to record the fetal heart rate. The second one is used to collect the uterine contractions and is hence posed not far from the uterine fundus, which is located at the upper part of the uterus.

Numerous studies have validated the reliability of CTG in predicting fetal hypoxia during the early stages of labor.

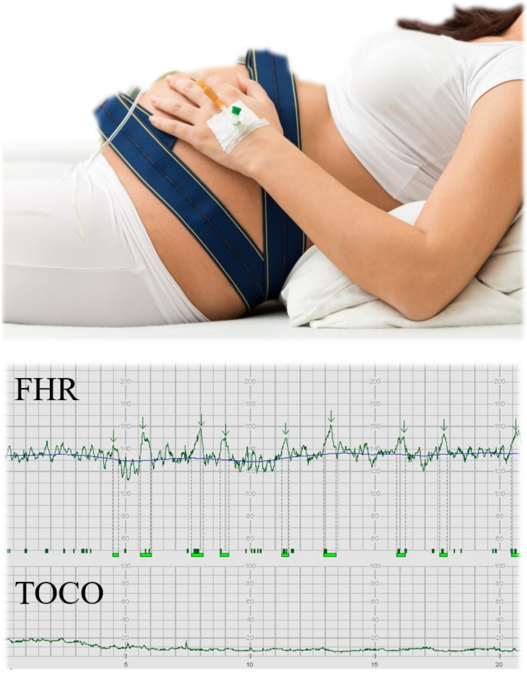


Figure 3.1: CTG consists of the simultaneous recording of two diverse tracings, i.e., FHR and TOCO, which are obtained with two different transducers, placed on the abdomen of the pregnant woman.

In contrast, an abnormal CTG test has limited predictive value for identifying fetal distress or hypoxia when compared to other indicators of fetal health, such as analyzing the acid-base balance of blood sampled from the umbilical artery or assessing the Apgar score. Indeed, a normal fetal heart rate (FHR) pattern predicts an Apgar score greater than 7 at 5 minutes with 99% accuracy, while a pattern indicative of fetal distress is associated with the birth of a healthy neonate in approximately 50% of cases **[3.3]**.
Conversely, the effectiveness of antepartum cardiotocography, or non-stress testing, in detecting fetal distress during pregnancy remains a subject of controversy. The primary challenge lies in the absence of a universally agreed-upon set of criteria for interpreting CTG tracings.

As a matter of fact, the complete non-invasiveness of CTG, which is the main characteristic that has determined the great diffusion of this technique, is however counterbalanced by the fact that in most clinical contexts CTG tracings are analyzed through eye inspection by obstetricians and/or gynecology specialists. Since the introduction of CTG in clinical practice, numerous visual reading systems have been developed, none of which has become the established standard. Furthermore, studies have highlighted that even in centers adhering to the same guidelines, individual observers rarely reach a consensus in evaluating the same tracing. Moreover, there is ample evidence that visual assessment of CTG tracings fails to capture all the intricacies of cardiac variability present in the FHR signal. Factors such as the extent of periodic components in the signal generated by the cardiac pacemaker, the non-linearity of the FHR control system, and short-term variability cannot be adequately discerned through visual inspection of the CTG tracing **[3.4-3.5]**.

The considerable inter and intra-observer variability and the inability of the human eye to extract quantitative information from the FHR signal played a key role representing the real weakness of the method **[3.6]**. Moreover, attempts made so far to interpret the tracings, according to various guidelines, did not provide the desired results **[3.7]**.

3.2. Computerized cardiotocography.

The analysis of CTG tracings received a boost since the early 1980s, with the introduction of computerized CTG (cCTG), which had the goal of overcoming the limitations of classical CTG.

The cCTG relies on quantitative measurements of both FHR and UCs and extracts from the two signals a set of features that could offer better insights and assist clinicians in their evaluations **[3.8-3.9]**. Since the introduction of cCTG, for at least two decades, the research has focused on finding a single methodological approach that could represent the gold standard in discriminating healthy from pathological fetuses. This allowed to quantitatively reproduce the standard analysis method based on eye inspection of CTG signals' time course **[3.10]**, conferring reproducibility to the method and playing a role in reducing the high intra and inter-observer variability. However, it was not enough to reach a satisfactory assessment of fetal wellbeing. The reliability of such approach has been limited for long time using basic time domain analysis, considering linear parameters only. As a matter of fact, a better understanding of physiological mechanisms impinging FHR control showed how the Autonomous Nervous System (ANS) acts in a very complex way, particularly when pathological conditions arise.

On the other hand, it has been observed that FHR changes anticipate and can predict fetal distress as well as adverse conditions before the insurgence of any other recognizable symptom **[3.11]**. In this context, more sophisticated FHR Variability (FHRV) investigations have been proposed, stressing the importance of considering multiple parameters to assess fetal state **[3.12]**. Moreover, even frequency analysis parameters started to be used for quantifying fetal cardiovascular control mechanisms as it happened for adults **[3.13]**.

A further development was introduced with the application of non-linear methods to biological time series, which can investigate the geometric and dynamic properties of the FHR signal. Entropy estimators **[3.14]**, complexity indices **[3.15]** as well as wavelets **[3.16]** and other nonlinear related parameters were applied with the aim to improve the information enhancement from the FHR **[3.17-3.19]**. Such techniques allowed to describe and understand complex physiological control mechanisms thanks to novel available tools. A review of the most used nonlinear indices applied to FHR was recently published by Ribeiro et al. **[3.20]**. These nonlinear indices were added to more traditional signal processing parameters developed in time domain such as the ones derived from classical analysis in time domain: Short-

and Long-Term Variability (STV and LTV), Delta and Interval Index (II) as proposed by Arduini et al **[3.21]**.
Thus, researchers started to understand that the problem of interpreting CTG tracings could not be solved by using a single index, but on the contrary it should make use of various features, extracted through different methodological approaches, to provide a reliable antenatal diagnosis.
Dawes & al. in 1991 were the first to publish a pioneer multivariate approach to CTG analysis, stressing the importance of how the integration of multiple parameters is effectively improving the assessment of the fetal wellbeing **[3.22]**.
Other examples of what above described are two papers **[3.23-3.24]** which focus on the early identification of Intra Uterine Growth Restricted (IUGR) fetuses. The multivariate model makes use of a small set of parameters and clearly shows how it outperforms accuracy, sensitivity, and specificity of each parameter in a univariate approach.
Moreover, it is clear that the mother-fetus system during pregnancy should be considered as a whole; thus, it can be viewed as a continuously evolving system, which must be monitored by means of time-varying approaches. These should take into account the development of FHR control mechanisms and, as a consequence, the changes in the relevant parameters/features. Thus, the search for a unique, robust and reliable set of parameters to accurately describe the FHR signal still remains an open challenge due to the evolution and complex nature of the fetal heart rate variability **[3.25]**.

3.3. On the limits of cCTG: new perspectives from AI.

The multiparameter approach aroused great interest, in particular, in the evaluation of the onset of states of fetal pathology. However, as the number of parameters increased, even a multifactorial statistical analysis became very difficult to be applied and researchers in the field of fetal monitoring started to consider AI techniques.

In the current era, there has been a notable surge in the abundance of data across various domains of knowledge, and the field of medicine stands to benefit significantly from this wealth of information. However, clinical data often face challenges in terms of lack of structured records and standardization **[3.26]**. In fact, clinical data often consist of not organized and not structured collections of measurements, signals and information which limit their analysis and interpretability. Addressing these issues becomes crucial to extract meaningful insights from data.

A medical exam that can benefit from a large amount of standardized and well-organized data is CTG.

The integration of Artificial Intelligence (AI) into CTG data analysis holds the potential for improving the discriminative power of this methodology **[3.27]**. AI models have the intrinsic ability to discover patterns and trends in data that may go unnoticed by visual analysis, providing valuable insights to manage risky pregnancies and to take appropriate decisions **[3.28]**.

Indeed, several attempts have been made to address the problem of classifying CTG tracings through Machine Learning (ML) techniques **[3.29-3.30]**.

In recent years, the focus has shifted towards employing Deep Learning (DL) algorithms, which can implicitly understand signal characteristics without relying on predefined features. DL algorithms are considered the most promising approach in this field **[3.31-3.33]**. However, their effectiveness heavily relies on the availability of large amounts of labeled training data. Insufficient data quantity and/or the lack of accurate labels may hinder practical implementations, leading to overestimation of model generalization capabilities and an increased risk of misclassification. Moreover, to achieve high performance, these methods must deal with appropriately structured data.

Currently, large CTG datasets specifically designed for deep learning purposes are not readily accessible, and only a few CTG datasets are publicly available online.

To the best of our knowledge the largest of them is the one presented by Ayres de Campos et al. **[3.34]**. This dataset does not include row CTG tracings, but only a series of 23 features extracted from the FHR tracings. These describe statistical aspects of the latter such as variability and central tendency. Each record also contains a tag added by an expert clinician, indicating one among three possible fetal conditions, i.e., pathological (P), normal (N) or suspect (S). The dataset comprises a total of 2,126 instances split in 1,655 N, 295 S and 176 P.

Another freely accessible CTG dataset is the one described by Chudaceck et al. **[3.35]**, which can be found in PhysioNet **[3.36]**. This contains 552 intrapartum recordings collected at the obstetrics ward of the University Hospital in Brno, Czech Republic. It includes the FHR time series and UC signals together with a set of maternal, delivery, and fetal clinical details.

These two datasets are the most frequently used in the literature. For example, Fergus **[3.37]** used the dataset from **[3.35]**to prove that the employment of AI techniques has the potential to outperform visual inspection to determine whether a cesarean section was necessary or not. This dataset was also employed used to test various DL architectures such as those presented in the work from Zhao et al. **[3.38]** and in the study of Ogasawara et al. **[3.39]**.

Also the dataset from **[3.34]** was used in different works, including the one described in **[3.40]** which analyses the performance of ten ML classification models in predicting the fetal risk. The same dataset was also used in the work from **[3.41]** which explores the capacity of diverse ML algorithms such as Support Vector Machine, Random Forest and K-Nearest Neighbors to assess the fetal health state.

Another available dataset is the one described in Signorini et al. **[3.42]** which includes a set of 12 linear and nonlinear indices referred to a little cohort of 120 fetuses, equally split in 60 healthy and 60 Intra Uterine Growth Restricted

(IUGR). This dataset has been used in several works, such as **[3.43]** which analyses the employment of the most common ML algorithms to predict IUGR.

Among these datasets, only the one by Chudaceck et al. **[3.35]** contains the CTG signals and is therefore suitable for DL studies, despite its relatively low sample size.

In general, publicly accessible CTG datasets suffer from some important limitations, particularly related to both limited numerosity and the lack of a fair balance in the representation of the different classes contained in the dataset.

There are also several papers in the literature that refer to datasets that cannot be publicly accessed, such the one reported in the work from **[3.44]**, which is said to contain more than 35,000 records, and the one mentioned in **[3. 45]**, which is said to contain 73,802 nonstress tests.

An extensive collection of CTG data could empower clinicians and researchers with a more accurate and comprehensive understanding of fetal well-being. In fact, a rich collection of organized clinical data could lead to an enhancement of CTG's diagnostic capabilities, helping to detect signs of fetal distress or abnormalities.

By analyzing a vast collection of CTG data, clinicians could identify patterns and correlations that might indicate potential risks or complications earlier, leading to more accurate diagnoses and timely interventions and hence improving patient outcomes. A large CTG dataset could even lead clinicians to have access to a broader spectrum of cases, including maternal conditions (such as diabetes) that could compromise the fetal outcomes. This could enable them to develop a deeper understanding of these situations, recognize subtle patterns associated with specific conditions, and make more informed decisions in challenging scenarios.

In the next chapter we will illustrate the steps we took to face the aforementioned limits and that could lead to the setup of a large, balanced and structured dataset of labelled CTG recordings. The latter is suitable for the application of those

AI techniques, the application of which requires can't do without large volumes of data to achieve acceptable results.

3.4. References for Chapter 3.

[3.1] John Hopkins Medicine. (2018). 4 common pregnancy complications. https://www.hopkinsmedicine.org/health/conditions-and-diseases/staying-healthy-during-pregnancy/4-common-pregnancy-complications.

[3.2] K.Hammacher, P.H.Werners, "Über die auswertung und dokumentation von ctg-ergebnissen", Gynecologic and Obstetric Investigation 166 410-423,1968.

[3.3] S.B. Thacker, "The efficacy of intrapartum electronic fetal monitoring", Am J Obstet Gynecol. Jan;156(1):24-30, 1987.

[3.4] T.Todros, C.U.Preve et al., "Fetal heart rate tracings: observers versus computer assessment", European Journal of Obstetrics & Gynecology and Reproductive Biology., 68:83–6,1996.

[3.5] S.Schiermeier, G.Westhof et al., "Intra- and Interobserver Variability of Intrapartum Cardiotocography: A Multicenter Study Comparing the FIGO Classification with Computer Analysis Software", Gynecol Obstet Invest.;72(3):169–73, 2011.

[3.6] D. Ayres-de Campos, J. Bernardes et al., "Inconsistencies in classification by experts of cardiotocograms and subsequent clinical decision". British journal of obstetrics and gynaecology 106 1307-1310, 1999.

[3.7] D. Ayres-de-Campos, J.Bernardes, "Twenty-five years after the FIGO guidelines for the use of fetal monitoring: time for a simplified approach?", International journal of gynaecology and obstetrics: the official organ of the International Federation of Gynaecology and Obstetrics 110 1-6, 2010.

[3.8] G.S.Dawes, m.Lobb et al., “Antenatal cardiotocogram quality and interpretation using computers”, BJOG. ;99(10):791–7, 1992.
[3.9] S.L.Sholapurkar, “The present and future of intrapartum computerized cardiotocography: role of pattern recognition incorporating single vs. multiple parameters”, The Journal of Maternal-Fetal & Neonatal Medicine. ;35(25):7452–8, 2022.
[3.10] G.Visser, G.S.Dawes et al., “Numerical analysis of the normal human antenatal fetal heart rate”, BJOG: An International Journal of Obstetrics & Gynecology 88 792-802, 1981.
[3.11] D.Hoyer et al. “Monitoring fetal maturation - Objectives, techniques and indices of autonomic function”, Physiological Measurement 38 R61, 2017.
[3.12] H.P.Van Geijn “Developments in CTG analysis”, Bailliere's Clinical Obstetrics and Gynecology 10, 185-209, 1996
[3.13] M.Malik et al. “Heart rate variability” Standards of measurement, physiological interpretation, and clinical use, European Heart Journal 17 354-381, 1996.
[3.14] S.M.Pincus “Approximate entropy as a measure of system complexity”, Proc. Natl. Acad. Sci. USA 88 2297-2301, 1991.
[3.15] A.Lempel, J.Ziv, “On the complexity of finite sequences”, IEEE Trans. Inf. Theory. 22 75-81, 1976.
[3.16] I.Daubechies “The wavelet transform, time-frequency localization and signal analysis”, IEEE transactions on information theory 36 , 961–1005,1990.
[3.17] S.M.Pincus, R.R.Viscarello, “Approximate entropy: a regularity measure for fetal heart rate analysis”, Obstet. Gynecol. 79 249–55, 1992.
[3.18] M.Ferrario, M.G.Signorini, G.Magenes, “Comparison between fetal heart rate standard parameters and complexity indexes for the identification of severe intrauterine growth restriction. Methods Inf. Med. 46 186-190, 2007
[3.19] J.Spilka et al. “Using nonlinear features for fetal heart rate classification” Biomed. Signal Process. Control. 7 350-357, 2012.

[3.20] M.Ribeiro, J.Monteiro-Santos et al. “Non-linear Methods Predominant in Fetal Heart Rate Analysis: A Systematic Review”, Frontiers in Medicine 8 , 2021.
[3.21] D.Arduini , A.Rizzo A et al., “Computerized analysis of fetal heart rate”, Journal of Perinatal Medicine 22 22-27, 1994.
[3.22] Dawes GS, Moulden M, Redman CWG. System 8000: Computerized antenatal FHR analysis. jpme.;19(1–2):47–51,1991.
[3.23] G.Magenes, R.Bellazzi et al., “Comparison of data mining techniques applied to fetal heart rate parameters for the early identification of IUGR fetuses”, Annu Int Conf IEEE Eng Med Biol Soc.:916–9, 2016.
[3.24] M.G.Signorini, N.Pini , “Integrating machine learning techniques and physiology based heart rate features for antepartum fetal monitoring”, Comput Methods Programs Biomed. 2020 Mar;185:105015.
[3.25] D.Hoyer, J.Zebrowski et al. “Monitoring fetal maturation - Objectives, techniques and indices of autonomic function”, Vol. 38, Physiological Measurement. IOP Publishing Ltd, 2017.
[3.26] J.L.Aeberhard, A.P.Radan et al., “Artificial intelligence and machine learning in cardiotocography: A scoping review”, Vol. 281, European Journal of Obstetrics and Gynecology and Reproductive Biology, Elsevier Ireland Ltd, 2023.
[3.27] M.E.O’Sullivan, E.C.Considine et al., “Challenges of Developing Robust AI for Intrapartum Fetal Heart Rate Monitoring”, Front Artif Intell., 2021.
[3.28] G.Georgoulas, P.Karvelis et al., “Investigating pH based evaluation of fetal heart rate (FHR) recordings”, Health Technol (Berl), 2017.
[3.29] R.C.Deo, “Machine Learning in Medicine” Circulation, 2015.
[3.30] C.Ricciardi, G.Improta et al., “Classifying the type of delivery from cardiotocographic signals: A machine learning approach”, Comput Methods Programs Biomed, 2020.
[3.31] A.Petrozziello, C.W.G.Redman et al., “Multimodal Convolutional Neural Networks to Detect Fetal Compromise

During Labor and Delivery", IEEE Access. 2019; 7:112026–36.

[3.32] E.Spairani, B.Daniele et al., "A deep learning mixed-data type approach for the classification of FHR signals", Front Bioeng Biotechnol, 2022.

[3.33] Z.Cao, G.Wang et al., "Intelligent antepartum fetal monitoring via deep learning and fusion of cardiotocographic signals and clinical data", Health Inf Sci Syst., 2023.

[3.34] D.Ayres-de-Campos, J.Bernardeset al., "Sisporto 2.0: A program for automated analysis of cardiotocograms", J Matern Fetal Med., 2000.

[3.35] V.Chudáček, J.Spilka et al., "Open access intrapartum CTG database", BMC Pregnancy Childbirth., 2014.

[3.36] A.L.Goldberger, L.A.N.Amaral et al., "PhysioBank, PhysioToolkit, and PhysioNet Components of a New Research Resource for Complex Physiologic Signals, available from: http://www.physionet.org, 2000.

[3.37] P.Fergus , A.Hussain et al., "Classification of caesarean section and normal vaginal deliveries using foetal heart rate signals and advanced machine learning algorithms", Biomed Eng Online, 2017.

[3.38] Z.Zhao, Y.Deng et al., "DeepFHR: intelligent prediction of fetal Acidemia using fetal heart rate signals based on convolutional neural network", BMC Med Inform Decis Mak. 2019.

[3.39] J.Ogasawara, S.Ikenoue et al. "Deep neural network-based classification of cardiotocograms outperformed conventional algorithms", Sci Rep., 2021.

[3.40] Z.Hoodbhoy, M.Noman et al., "Use of Machine Learning Algorithms for Prediction of Fetal Risk using Cardiotocographic Data", Int J Appl Basic Med Res. 2019.

[3.41] A.Mehbodniya, A.J.P.Lazar et al., "Fetal health classification from cardiotocographic data using machine learning", Expert Syst., 2022.

[3.42] M.G.Signorini, N.Pini, et al., "Dataset on linear and non-linear indices for discriminating healthy and IUGR fetuses", 2020.

[3.43] V.P.C.Magboo, “Prediction of Late Intrauterine Growth Restriction using Machine Learning Models”, Procedia Comput Sci.;207:1427–36, 2022.
[3.44] A.Petrozziello, I.Jordanov et al., “A. Deep Learning for Continuous Electronic Fetal Monitoring in Labor” Annu Int Conf IEEE Eng Med Biol Soc., 2018:5866–9, 2018.
[3.45] J.Pardey, M.Moulden et al., “A computer system for the numerical analysis of nonstress tests”, Am J Obstet Gynecol., 186(5):1095–103, 2022.

Chapter 4

Data collection: the creation of NAPAMI, a novel dataset for cCTG.

4.1. Background.

As anticipated at the end of **Chapter 3**, the potential application of artificial intelligence (AI) techniques in the analysis of CTG data seems to be more than a hope to stem the limitations deriving from the visual inspection of CTG tracings. However, nowadays, the poor quantity of reliable, structured and labelled real clinical datasets represents a non-trivial obstacle, since the most promising AI methods (e.g., deep learning algorithms-DL) inevitably require a large amount of data to obtain high-quality results.

This section illustrates the steps that, starting from an unorganized set of CTG recordings, collected at the Federico II Hospital in Napoli, Italy, have taken to the setup of a large, balanced, structured dataset of labelled CTG recordings, which we have baptized with the name NAPAMI.

In particular, in the course of the current section, it will be detailed how, starting from an unorganized collection of 37,095 CTG recordings, stored in a MS-Access database during ambulatory non-stress tests, we managed to achieve the final version of NAPAMI. A four step procedure was adopted and allowed to a) extract the labels identifying pregnancy at risk b) preprocess the FHR signals by filtering out artifacts and loss of signal, c) extract time, frequency and non-linear domain indices and compute new parameters on the FHR signal, d) fix the remaining bugs.

The obtained dataset has the necessary features allowing to explore the classification performances of all those algorithms which, by their nature, require many examples.

In fact, NAPAMI is a large, structured and annotated dataset of CTG recordings (NAPAMI), that could be used for developing and testing AI supervised solutions to improve the care and outcomes of pregnant women and their offspring.

This result opens application perspectives toward obtaining a system for the automatic classification of CTG recordings representing a reliable diagnostic tool to support clinical decisions.

In fact, NAPAMI can be a valuable resource for testing and validating new approaches for the CTG classification, improving the management and outcomes of pregnant women in the antenatal period and during labor. In addition, the availability of a wide range of case histories could also grant clinicians access to a wider range of information and this could result in a deeper understanding of diverse antenatal scenarios.

4.2. Data Collection and Access Database Completing.

The CTG recordings were collected at the Gynecology and Obstetrics Unit, Department of Neuroscience, Reproductive Sciences and Dentistry, School of Medicine, Federico II

University of Naples, Italy, as part of routine antepartum fetal monitoring exams.

The CTG monitoring was conducted in a controlled clinical environment, with the patients lying on an armchair, using a Philips Avalon FM30 device (**Figure 4.1**) equipped with an ultrasound transducer and a transabdominal tocodynamometer. These fetal monitors utilize an autocorrelation technique to compare the demodulated Doppler signal of a given heartbeat with the subsequent one, in order to obtain the heart period (equivalent to the RR period). The heart period is then translated into a heart frequency in beats per minute (bpm) upon the detection and acceptance of a new heart event.

Figure 4.1: The upper part shows the Philips Avalon FM30 device, which is equipped with an ultrasound transducer and a transabdominal tocodynamometer for the registration of FHR and TOCO tracings. The latter are analyzed with the 2CTG2 software for the cCTG, which is illustrated below.

Each CTG recording has a duration ranging from about 20 to 60 minutes.

The raw signals acquired were processed through the 2CTG2 program **[4.1]**, a software designed for the computerized CTG (**Figure 4.1**). This system reads the FHR and UC signals with a sampling frequency of 2Hz, which represents a good compromise to minimize the repetition of FHR values and to provide at the same time an appropriate level of detail for advanced analysis, including frequency and non-linear parameters. The 2CTG2 software allows users to inspect the FHR and UC signals on a computer screen.

The 2CTG2 also includes algorithms, that perform the computation of several quantitative features, starting from the raw FHR tracings **[4.2]**, as will be detailed in the following sections.

At each exam, the CTG tracings, together with the computed parameter set, are automatically stored in a Microsoft Access Database (.mdb).

Each recording is labeled with a class assigned by the members of the medical team at the occurrence of the first examination. The labeling process was carried out following the main indications for antepartum CTG monitoring with documented pathologies at risk of chronic hypoxia according to ACOG guidelines **[4.3]**.

Code 01 refers to physiological pregnancies and includes the largest group of records. CTG monitoring was performed in physiological pregnancies starting from 39 weeks of gestation. Prior to this gestational week, the CTG in this group was only offered for study purposes.

Code 02 is used to identify physiological twin pregnancies.

Code 03 includes pregnant women with fetuses affected by IUGR or fluximetric alterations (fetal pathology).

Code 04 includes all maternal pathologies that can cause chronic fetal hypoxia, such as type I, type II or gestational diabetes, essential or gestational hypertension, thyroid dysfunctions, maternal chronic kidney disease, systemic lupus erythematosus, connective tissue disease with autoantibodies to SS-A and SS-B, antiphospholipid antibody syndrome, maternal hypoxia, hemoglobinopathies and other maternal anemias with hemoglobin < 8 g/dl, cyanogenic maternal heart disease. It also includes pregnant subjects with severe obesity (Body Mass Index ≥40) and dyspnea or drug addicted.

Code 05 is used to label cases of fetal malformation with potential cardiovascular impact.

Code 06 mainly includes cases of maternal pathologies that have caused chronic fetal hypoxia. Therefore, this group of fetuses already experienced complications related to the maternal pathology (e.g., pregestational or gestational diabetes and macrosomia or hypertension/preeclampsia and IUGR, respectively), differently from fetuses with code 04 where the risk is only potential.

Code 07 includes fetuses affected by a malformation diagnosed or suspected during the 2nd trimester of pregnancy (e.g. heart disease, suspected esophageal or duodenal atresia, clubfoot, spina bifida) associated with at least one obstetric complication developed during the 3rd trimester of pregnancy (e.g. IUGR, macrosomia, polyhydramnios, oligohydramnios). These are fetuses at increased risk for genetic and/or chromosomal syndromes, which are often diagnosed after birth.

Code 08 includes pregnant women with an obstetric disease potentially causing chronic fetal hypoxia associated with a fetal malformation diagnosed or suspected during 2nd trimester screening ultrasound (e.g. chronic/gestational hypertension and fetal heart disease or pregestational diabetes and fetal heart disease or spina bifida, respectively). It is not necessary that a correlation between the maternal pathology and the fetal malformation has been demonstrated.

Code 09 includes twin pregnancies in which one or both fetuses developed a pregnancy related complication, almost always IUGR.

Code 10 includes twin pregnancies in which the pregnant woman is affected by pathologies with a potential risk of chronic fetal hypoxia.

Code 11 includes twin pregnancies in which one or both fetuses have a malformation diagnosed or suspected during 2nd trimester screening ultrasound. These are fetuses at increased risk for genetic and/or chromosomal syndromes, often diagnosed after birth.

The 12th group (“xx”) includes records without a code. They are associated with recordings belonging to pregnancies classified as high-risk but in which fetuses were never associated to a risk of fetal chronic hypoxia. These were collected exclusively for study purposes.

A code for late-term or post-term pregnancy has not been included, because pregnant women over 40 weeks + 6 days of gestation were managed in the emergency room.

Oligohydramnios has not been considered as an independent risk category, because it was frequently associated with other pregnancy complications, such as IUGR or premature rupture of membranes. The term or preterm premature rupture of membranes with suspected chorioamnionitis was managed in the emergency room and then admitted in the Obstetrics Unit.

Decreased fetal movements is an occasional indication to CTG. However, since if the trace is reactive and the amniotic fluid is regular (maximum pocket >2 cm) there is no indication to repeat the exam, a specific code was not assigned. Maternal-fetal isoimmunization and previous fetal death (unexplained or recurrent risk) are also infrequent events, therefore no specific code has been assigned.

A summary reporting codes and their related description can be found in **Table 4.1**.

Table 4.1: Outpatient code with associated description.

CODE	DESCRIPTION
01	physiological pregnancy
02	twin pregnancy
03	fetal pathology (Intra Uterine Growth Restriction - IUGR, fluximetric alterations)
04	maternal pathologies that can cause chronic fetal hypoxia
05	fetal malformation (with possible cardiovascular impact)
06	maternal pathology + fetal pathology
07	fetal pathology + fetal malformation
08	maternal pathology + fetal malformation
09	twin pregnancy + fetal pathology
10	twin pregnancy + maternal pathology
11	twin pregnancy + fetal malformation
xx	Unknown (records that do not fall into any of the above categories)

The Access Database (.mdb) includes a set of records relating to 10,918 pregnant women, yielding a total corpus of 37,095 CTG registrations. Before it was provided to us, the database was completely anonymized by the hospital itself to safeguard the privacy of pregnant women who underwent the examination. Avoiding compromising the confidentiality of the individuals represented in the data, each personal information, including name, surname, address etc., was removed.

The aforementioned Access database has been organized into 4 tables:

- *PCC_INFO*: includes a "notes" column containing the outpatient code together with

details such as the Apgar score, and any other comment, manually added by the clinician.
- *CTG_EXAM*: contains the reference ID associated to the patient and to the exam and the gestational week.

- *CTG_PARAMETERS*: contains the ID associated to the exam and the parameters computed by the 2CTG2 software.

- *CTG_TRACES*: contains the ID of the exam, the length of the FHR trace and the row CTG signals, i.e., FHR, UC and fetal movement profile (FMP), compressed as Binary Long Objects (BLOB).

Figure 4.2 shows the schematic representation of the Microsoft Access Database (.mdb). The dashed lines in represent the linkages between the four tables composing the structure.

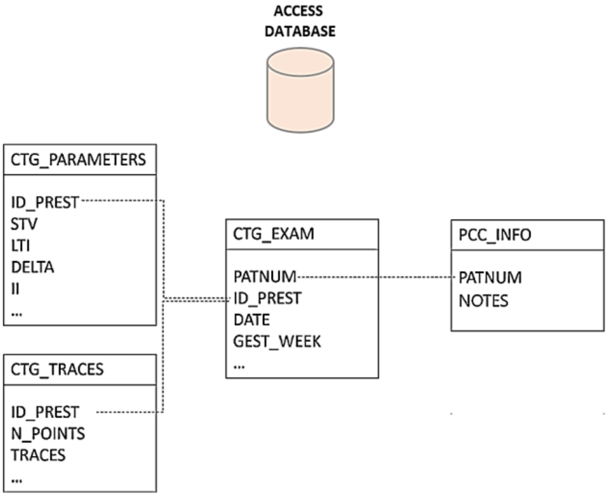


Figure 4.2: Structure of the Microsoft Access Database (.mdb file). CTG_EXAM and PCC_INFO are linked through the key PATNUM. While CTG_PARAMETERS, CTG_EXAM and CTG_TRACES are linked through the key ID_PREST.

4.3. Extracting data from the Access database.

The first step towards building up the final structured dataset is the retrieval, from the Access (.mdb) database, of all the information relative to each exam, including the row tracings and the parameters computed by the 2CTG2. This was obtained by establishing a connection with the Access file and retrieving the fields of interest through an SQL query. To this end MATLAB 2021a was employed.

After each record was retrieved from the Access database, the BLOB objects codifying the row CTG tracings were decrypted to obtain FHR signals, together with perceived fetal movements (FMP) and UCs. A more detailed description of FMP will be furnished in **Chapter 5**.

Once the raw FHR traces were extracted, a preprocessing step was applied to remove artifacts. As seen in **Chapter 1**, in clinical practice, the FHR signal is recorded using an ultrasound (US) probe placed externally on the abdomen of the pregnant woman. Movements produced by the opening and closing of the fetal cardiac valves represent the information content of the US signal. By an algorithm based on the autocorrelation function, the firmware of the CTG monitors reconstructs with good accuracy the occurrence of the fetal beats providing the FHR. However, there are several factors that can affect the measurement of FHR, such as the movement of the mother and fetus, the displacement of the transducer, and also events in the external clinical environment.

This can result in artifacts and signal losses which are the major sources of noise in the fetal signal. Therefore, the main goal of the pre-processing phase is to reduce these disturbances that worsen the successive phases of the analysis of the signals. Signal intervals with losses having a duration of less than 15 s, were linearly interpolated, while losses of longer duration, were instead completely removed from the signal. **Figure 4.3** shows an example of a noisy row FHR signal before **(a)** and after **(b)** pre-processing.

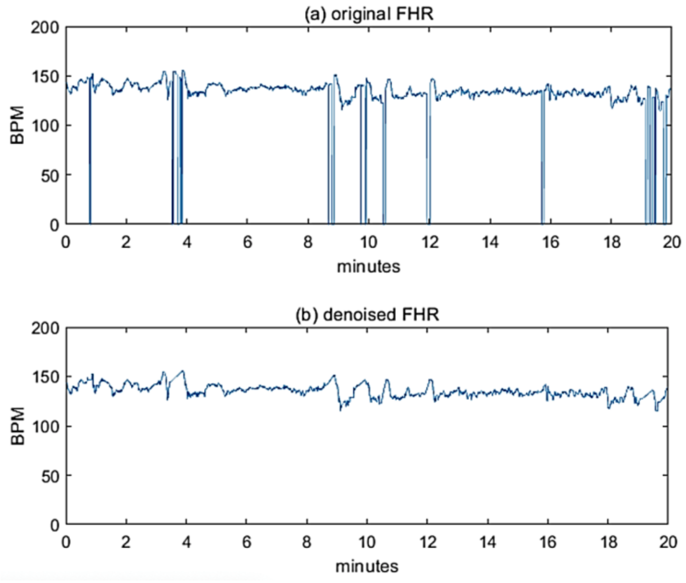


Figure 4.3: example of row FHR signal before (a) and after (b) artifacts removal.

For what concerns instead parameters, the group of regressors automatically computed by the 2CTG2 includes a heterogeneous set of indices, pertaining to different domains, which are yet known in literature and allow for a description of FHR signals from different viewpoints. The 2CTG2 performs the calculation of parameters on non-overlapping 3-minutes windows and then averages the obtained values. The classical set of features automatically calculated by the 2CTG2 comprises:

1) *Time domain parameters*: mean FHR, Delta **[4.4]**, Interval Index (II) **[4.5]**, Short Term Variability (STV) **[4.4-4.5]**, Long-Term Irregularity (LTI) **[4.4]** and number of large and small accelerations per hour **[4.6]**.

2) *Frequency domain parameters*: Low Frequency (LF), Movement Frequency (MF), High Frequency (HF) **[4.7]**.

3)*Non-linear parameters*: Approximate Entropy (ApEn) **[4.8-4.9]**.

4.4. Computation of additional parameters at different time scales.

We decided to integrate the set of parameters computed by the 2CTG2 with a new set of indices already known in the literature, although some come from different research domains. The new feature set includes the binary and ternary Lempel-Ziv Complexity indexes (LZ2, LZ3) **[4.10-4.11]**, Acceleration and Deceleration Phase Rectified Slope (APRS, DPRS) indices extracted from the Phase Rectified Signal Averaging (PRSA) **[4.12-4.13]**, acceleration and deceleration capacities (AC, DC) **[4.12]**, deceleration reserve (DR) **[4.14]**), the normalized power in the LF, MF and HF bands computed from the PRSA signal **[4.12]**, Sample Asymmetry **[4.15]** and Sample Entropy (SampEn) **[4.16-4.17]**.

We decided to perform the computation of the different parameters introduced so far at diverse time scales.

In particular, a group of 16 global parameters (PG) was computed on a single segment of 20 minutes extracted from the whole FHR signal, which corresponds to the longest segment available for all recordings, and includes: mean value, median value, standard deviation, inter-quantile range, acceleration and deceleration capacity (AC, DC), DR, APRS, DPRS, normalized power in the LF, MF and HF bands computed from the PRSA signal, Sample Asymmetry, SampEn , ApEn, LZ2, LZ3.

In addition to the features calculated by considering the signal in its entirety, other parameters were computed by shifting a fixed size moving window (1 min or 3 min) along the FHR signal, and are stored as series. More specifically, the parameters computed at each slide of the 3-minute length window (P3) include 19 indices, i.e., mean frequency, total signal power, power in the Very Low Frequency (VLF), LF, MF and HF bands, percentage of power in LF, MF and HF bands, LZ2, LZ3, ApEn and SampEn (both computed with 3 different parameter choices), LTI both with and without excluding large accelerations.

Moreover, 6 parameters were computed at each shift of the 1-minute length sliding window (P1): DELTA, STV and II, all computed with and without excluding accelerations and decelerations.

An exhaustive description of each of the presented parameters can be found in the **Appendix** section.

4.5. Final adjustments.

Once the row tracings had been extracted and processed, the outpatient code retrieved, and the 2CTG2 parameters integrated with the new set of features computed at different time scales, a processing step was necessary to fix inconsistencies within the dataset assembled so far.

The process involved several steps, such as eliminating duplicated records. Additionally, records with a duration of less than 20 minutes were removed, along with those where the gestational age or maternal age was missing or lacked plausible values (e.g., mother age over 70). Furthermore, excessively poor-quality signals, characterized by significant noise distortion, were directly discarded. Finally, we proceeded to correct typos from the “notes” field. The latter contains several information about the health status of both mother and fetus as well as the related outpatient code. Since this field is manually filled out by the physician performing the analysis, it is possible that the entered text includes spelling and typographical errors that needed to be fixed to allow a correct retrieval of meaningful information. **Figure 4.4** summarizes all the aforementioned steps.

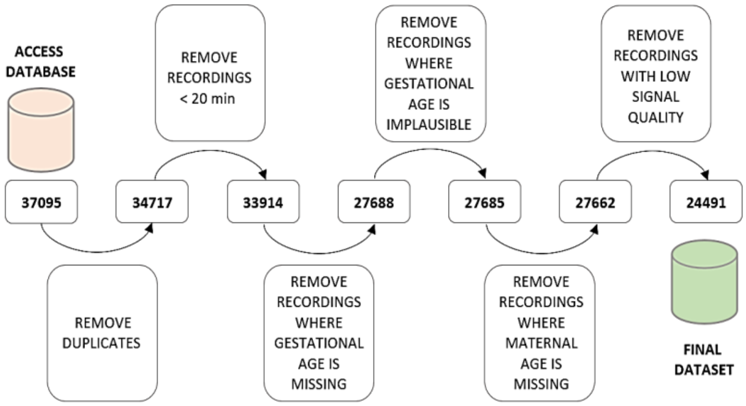


Figure 4.4: Visual illustration of the necessary steps taken to clean the records retrieved from the Access Database and obtain the final Dataset.

4.6. NAPAMI: the obtained dataset.

After all the steps discussed so far, the obtained dataset includes a total of 24,491 CTG records, organized in the form of a MATLAB table (. mat).

The NAPAMI Dataset comprises 16 fields which can be organized into 3 groups by their semantic meaning, as schemed in **Figure 4.5**.

The first group includes information related to pregnancy. In particular, it comprises the patient number (PATNUM), the age of the pregnant woman, the gestational week, the outpatient code and the note field. This contains several information items about the maternal and fetal health status, such as the presence of pre-existing pathologies, the assumption of drugs, the maternal weight etc. It is important to note that the PATNUM does not match the identifier originally assigned in the Access database. In fact, the latter has been replaced by a new numeric code whose sole purpose is to identify examinations related to the same patient, within the same pregnancy.

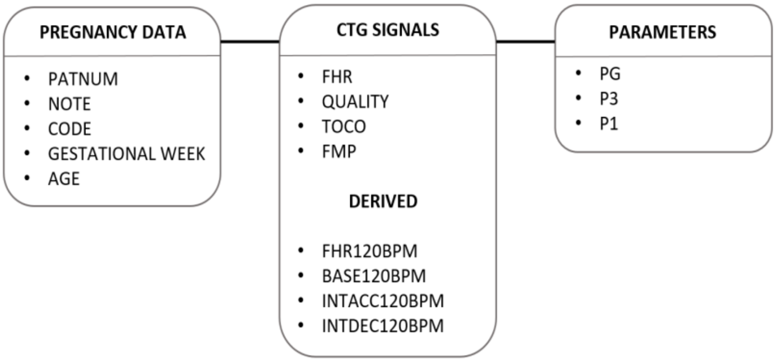


Figure 4.5: Schematization of the final dataset. The obtained dataset, consisting of 16 fields, can be organized into 3 groups. The first group includes data related to pregnancy while the second and the third ones respectively comprise the CTG signals and the computed parameters' sets.

The second group comprises the CTG tracings, including the row FHR signal together with QUALITY, TOCO and FMP. Each QUALITY trace has the same length as the corresponding FHR trace. For each FHR sample it indicates the related quality through a number ranging from 32 (good quality) to 224 (bad quality). The FMP trace is also as long as the corresponding FHR signal and is an array composed of zeros and ones, respectively indicating the presence/absence of perceived fetal movements.

In addition to the raw FHR signals, we enriched the final dataset by adding both the corresponding denoised FHR traces in beats per minute (FHR120bpm) and their associated baselines (Base120bpm). From these, we identified the time points where accelerations and decelerations occur and their duration in seconds (Intacc120bpm, Intdec120bpm).

The third group includes all the parameters computed at different time scales, which have been detailed in the previous session.

The number of records stratified by gestational code is illustrated in the bar plot in **Figure 4.6 (a)**, while **Figure 4.6(b)** depicts the number of occurrences per gestational week.

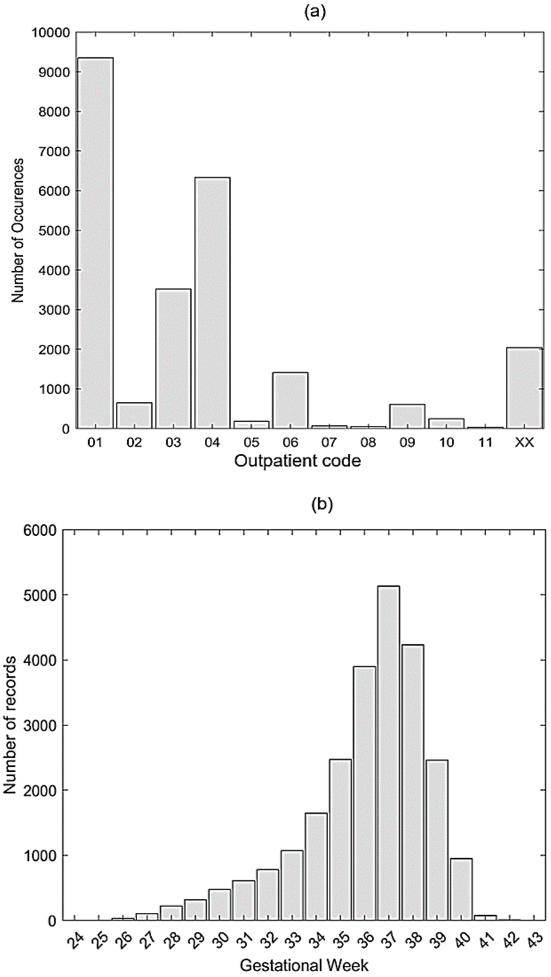


Figure 4.6: (a) Number of occurrences per outpatient code; (b) number of records per gestational week.

As regards the numerosity per outpatient code, **Figure 4.6 (a)** clearly shows that the most represented class in the dataset coincide with physiological pregnancies (01). Moreover, it is possible to observe that a good portion of records are related to codes 03 and 04, which refer to the fetal and maternal conditions at greatest risk of chronic fetal hypoxia, for which CTG monitoring is required.

Code 06 includes maternal conditions resulting in chronic fetal hypoxia. The numerosity of this group is likely underestimated because the code was assigned during the first outpatient access, and unfortunately was rarely updated according to the evolution of obstetric complications. Foer example, many patients had a code 04 for hypertension and the fetus later developed IUGR but remained with a code 04 until delivery.

Codes 05, 07, 08 and 11 are poorly represented because the incidence of fetal malformations in live fetuses is low during the 3rd trimester of pregnancy. Indeed, in many cases the patients voluntarily terminated the pregnancy during the 1st or 2nd trimester. When the patient decided not to terminate the pregnancy despite severe cardiac, neurological or renal malformations with a low probability of neonatal survival, CTG monitoring was not performed.

The group “xx” included recordings performed in high-risk pregnancies not directly associated with fetal chronic hypoxia, for which there is no indication for antepartum CTG monitoring. This is a very heterogeneous group that includes pregnancies obtained with medically assisted procreation techniques, pregnancies complicated by cholestasis or severe hypertransaminasemia, autoimmune hypothyroidism, high-risk thrombophilia, severe obesity, maternal seroconversion for toxoplasmosis, cytomegalovirus or HIV. These are maternal conditions that could affect fetal well-being. Therefore, it was decided to monitor their pregnancies only for clinical research purposes.

Since gestational weeks from 34 to 39 individuate the typical time interval during which antepartum CTG is performed, the majority of records fall within this period (see **Figure 4.6 (b)**). Generally, monitoring starts from the 32nd week and has a weekly frequency, but the decision to start the antepartum CTG, as well as the frequency of controls, depends on the severity of the risk of hypoxia. In many cases, monitoring was started before the 32nd week and the frequency was biweekly, triweekly or even daily. Except for specific situations, it was not performed multiple times a day, because CTG is not predictive of acute adverse events (49). Most of the monitoring sessions were performed near term or at the end of the pregnancy, to identify the best timing for delivery.

Figure 4.7 shows another important factor to consider: the resulting well-balanced composition of NAPAMI for records related to physiological pregnancies and those with potential risks. By aggregating records labeled with codes 01 and 02 into one group representing physiological pregnancies, and grouping the remaining data labeled with codes 03 to 11 as potentially risky pregnancies, we can observe that the dataset is nearly evenly distributed between these two groups.

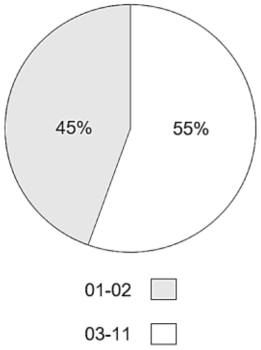


Figure 4.7: portion of records with codes 01-02 (physiological) vs 03-11 (risky).

4.7. Final considerations on NAPAMI.

In this chapter we have provided a detailed description of the steps taken to build up NAPAMI, which is, at date, one of the largest datasets on antepartum monitoring of healthy and high-risk pregnancies. It is a systematic and continuous collection of information on the clinical characteristics of patients and their fetuses during pregnancy up to delivery. In many cases of high-risk pregnancies, fetal monitoring data are available as early as the 26th week, and for some of these patients, data on delivery and neonatal outcomes are also available. The presence of an extensive collection of clinical data as the one provided by NAPAMI, coupled with the pursuit of fresh clinical evidence and the potential for ongoing assessment, would enable healthcare professionals in the fetal medicine to gain a more precise understanding of fetal well-being. This, in turn, can aid in making informed decisions regarding i.e., the optimal timing of delivery, particularly in cases where delivery must be conducted before the 34-week and carries the combined risks of prematurity and fetal pathology.

The dataset resulting from the various processing stages outlined in the manuscript enables the integration of existing clinical knowledge with findings from experimental research, verifying its practical applicability in clinical settings, and generating new evidence where it is currently lacking.

NAPAMI represents a reliable picture of the population of pregnant women over a quite large time window in a standard clinical context. Moreover, based on the evidence that the mother-fetus system should be considered as a whole, pregnancy can be viewed as a continuously evolving system which can be analyzed through time-varying approaches and our dataset contains longitudinal data coming from both healthy and risky pregnancies.

In fact, another significant benefit provided by the computed set of parameters within NAPAMI, is the possibility to evaluate the progression of fetal CTG features over time, without any restrictions, across all monitored patients. This presents a valuable opportunity for continuous auditing and analysis. The added value of these features will also be to facilitate a reliable analysis of FHR signal both for clinicians and researchers, independently from the methodological approach they decide to adopt. The dataset should be considered a tool to empower the diagnostic reliability of CTG.

In addition, the availability of a wide range of case histories could also grant clinicians access to a wider range of information, including those involving maternal conditions such as diabetes that could impact fetal outcomes.

Consequently, this could result in a deeper understanding of these scenarios, recognizing subtle patterns associated with specific conditions, and making more informed decisions even in challenging circumstances. Through the analysis of a vast set of CTG data, clinicians would be able to identify patterns and correlations that may indicate potential risks or complications at an earlier stage of pregnancy, enabling more accurate diagnoses and timely interventions.

4.8. References for Chapter 4

[4.1] G.Magenes, M.G. Signorini et al., “2CTG2: A new system for the antepartum analysis of fetal heart rate”, In: IFMBE Proceedings. Springer Verlag, p. 781–4, 2007.

[4.2] M.G.Signorini, G.Magenes, “Advanced Signal Processing Techniques for CTG Analysis”, In: IFMBE Proceedings, vol 57.. p. 1205–10, 2016.

[4.3] Practice Bulletin No. 145: antepartum fetal surveillance. Obstetrics & Gynecology. 2014 Jul;124(1):182–92.

[4.4] D. Arduini, G.Rizzo, C.Romanini, “Computerized Analysis of Fetal Heart rate”. jpme.;22(s1):22–7, 1994.

[4.5] J. de Haan, J.H.van Bemmel et al., “Quantitative evaluation of fetal heart rate patterns”, Eur J Obstet Gynecol. Jan;1(3):95–102, 1971.

[4.6] R. Mantel, H.P.Van Geijn et al., “Computer Analysis of Antepartum fetal heart rate”, Int J Biomed Comput.;25(4):261–86, 1990.

[4.7] M.G.Signorini, G.Magenes et al., “Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings”, IEEE Trans Biomed Eng.;50(3):365–74, 2003.

[4.8] S.Pincus “Approximate entropy (ApEn) as a complexity measure”, Chaos: An Interdisciplinary Journal of Nonlinear Science. Mar;5(1):110–7, 1995.

[4.9] X.Li, D.Zheng et al., “Approximate entropy of fetal heart rate variability as a predictor of fetal distress in women at term pregnancy", Acta Obstet Gynecol Scand. Sep;84(9):837–43, 2005.

[4.10] M.Ferrario, M.G.Signorini, G.Magenes, "Comparison between fetal heart rate standard parameters and complexity indexes for the identification of severe intrauterine growth restriction", Methods Inf Med.;46(2):186–90, 2007.

[4.11] A.Lempel, J.Ziv, "On the Complexity of Finite Sequences", IEEE Trans Inf Theory.;22(1):75–81, 1976.

[4.12] A.Bauer, J.W.Kantelhardt et al., "Phase-rectified signal averaging detects quasi-periodicities in non-stationary data", Physica A: Statistical Mechanics and its Applications.;364:423–34, 2006.

[4.13] A.Fanelli, G.Magenes et al., "Quantitative assessment of fetal well-being through ctg recordings: A new parameter based on phase-rectified signal average", IEEE J Biomed Health Inform.;17(5):959–66, 2013.

[4.14] M.W. Rivolta, T.Stampalija, M.G.Frasch et al., "Theoretical Value of Deceleration Capacity Points to Deceleration Reserve of Fetal Heart Rate", IEEE Trans Biomed Eng.;67(4):1176–85, 2020.

[4.15] B.P.Kovatchev, L.S.Farhyet al., "Sample Asymmetry Analysis of Heart Rate Characteristics with Application to Neonatal Sepsis and Systemic Inflammatory Response Syndrome", Pediatr Res.;54(6):892–8, 2003.

[4.16] D.E.Lake, J.S.Richman et al., "Sample entropy analysis of neonatal heart rate variability", American Journal of Physiology-Regulatory, Integrative and Comparative Physiology. Sep 1;283(3):R789–97, 2002.

[4.17] M.Ferrario, M.G.Signorini, G.Magenes, "Complexity analysis of the fetal heart rate variability: early identification of severe intrauterine growth-restricted fetuses", Med Biol Eng Comput.;47:911–9, 2009.

Chapter 5

Time and spectral analysis of CTG tracings.

In the previous chapter, we provided a comprehensive account of how we successfully constructed a rich dataset of CTG tracings (NAPAMI). This expansive collection paves the way for groundbreaking possibilities in the analysis of CTG signals.

In the following chapter, we will harness the wealth of this dataset to delve deeper into the comprehension and interpretation of FHR traces, exploring them both in temporal and spectral domains.

Section 5.1 introduces a novel methodology centered around a Hidden Markov Model, designed for the automatic identification of fetal behavioral stages within FHR tracings. This innovative approach promises to enhance our ability to discern and classify different stages of fetal activity, adding an invaluable layer of insight to our analysis.

Section 5.2 presents an innovative approach to the spectral analysis of FHR tracings. We aim to provide a fresh perspective on spectral analysis, comparing the performance of our approach with that of more commonly employed traditional methods.

Through this chapter we will shed light on the potential advantages and benefits of our proposed methodologies in elucidating the intricacies of FHR patterns.

5.1. Identification of fetal behavioral stages in FHR tracings.

For a proper understanding of CTG tracings, it is essential to know the pathophysiological mechanisms underlying changes in FHR both ante- and intra-partum. In particular, a correct interpretation cannot help but consider how fetal behavioral states influence heart rate and fetal responsiveness **[5.1]**.

The transition from one to the other behavioral state is a sign of integrity and good maturation of the fetal central nervous system (CNS) and autonomic nervous system (ANS). For example, a trace with no accelerations in 20 minutes of recording ("nonreactive") could be due to a quiet behavioral state. The longer the period of time in which there are no accelerations, the greater the risk that the fetus may be in a hypoxic condition. Conversely, a highly responsive well-oxygenated fetus might have incoming accelerations for more than an hour and simulate fetal tachycardia, a condition associated with hypoxia **[5.2]**.

For these reasons, the automatic identification of fetal behavioral states is a challenging but fundamental task in cCTG. Being able to spot the distinct fetal stages within the CTG tracings has the potential to increase the interpretability and reliability of this diagnostic methodology and is of fundamental clinical relevance. In fact, from the duration of the different phases, important information about the state of fetal well-being is derived. It has been observed that prolonged phases of inactivity are important indicators of pathological conditions, and alterations in the physiological alternation of fetal states have been associated with several conditions in pregnancy **[5.3-5.7]**. Moreover, CTG parameters have been shown to vary substantially according to the behavioral state, which suggests that they should be more correctly interpreted knowing the fetal state in which they were computed **[5.8-5.11]**.

Indeed, as discussed in **[5.12]**, the fetal behavioral state has been shown to affect the responsiveness of CNS.

Nijhuis **[5.1]** reported that for fetuses at term (i.e., from 36-38 weeks of gestation), 4 distinct behavioral states can be identified, which closely follow the behavioral states observed in neonates. These states are referred to as 1F (quiet sleep), 2F (active sleep), 3F (quite awake) and 4F (active awake). Each is characterized by a typical FHR pattern and is associated with a different motility and eye movements profile.

In fetuses not yet at term this distinction is less clear and is often difficult to identify these 4 behavioral states by definition.

In the context of the non-stress test, it is more common to consider only two states (i.e., “active” and “quiet”) which can be more consistently identified also at earlier gestational ages **[5.13-5.14]**.

The identification of different fetal states within CTG tracings is, to date, still left to the clinician's experience and can consequently lead to discordance in interpretations. In the wider context of automatic identification of fetal behavioral states in the FHR, only a few simple algorithms for their identification have been presented in **[5.15-5.17]**.

In this section, we present an unsupervised method based on Hidden Markov Models (HMMs) for clustering FHR signals’ points as belonging to “active” and “quiet” states. We believe that HMMs could be particularly well-suited for this task for several reasons. Unlike other clustering techniques, they correctly capture the inherent temporality of the analyzed signal and they naturally exploit the clinical observation that FHR signals change their characteristics in time, alternating among different phases that present consistent similarities and are generated by variations in the fetal state, which cannot be directly observed. Moreover, they are fully data-driven and present the great advantage to be very well interpretable.

The proposed approach is based on shifting a fixed-length sliding window on the FHR tracing and extracting a small set of features at each slide. The parameters are considered to be the set of emissions/observations of the HMM while the “active” and “quiet” phases are the hidden states. After the unsupervised training phase, the model can automatically assign each point of the shifting window to the state under which it is more likely to have observed the emitted set of observations **[5.18]**.
An illustration of the procedure is shown in **Figure 5.1**.

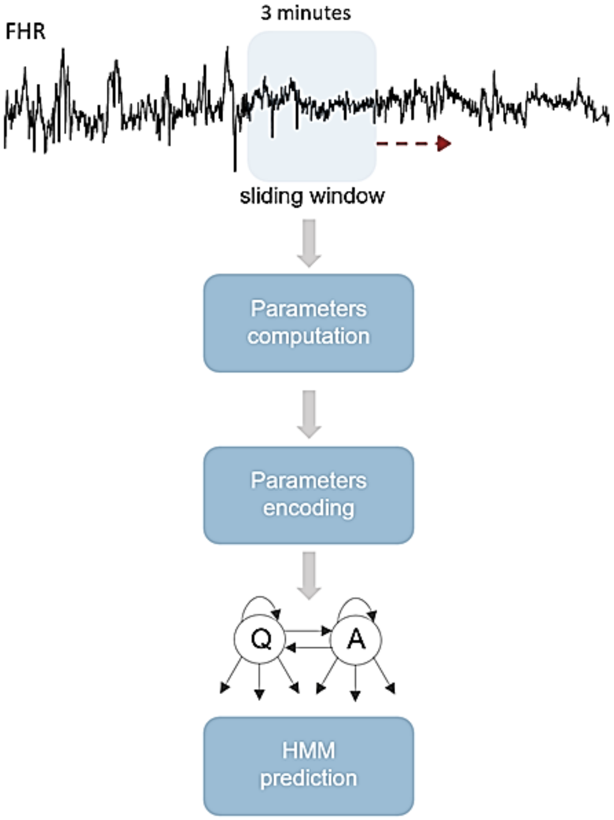


Figure 5.1: Illustration of the whole described procedure. At each shift of a 3-min sliding window, a set of selected parameters, describing different aspects of the FHR chunk within the 3-min window, is obtained. The computed set is then properly encoded and passed in input to the HMM. The

latter predicts the state which is more likely to have generated the observed sequence of parameters.

5.1.1. Overview of the proposed method.

The core idea behind this approach is to exploit the a priori knowledge of the existence of two distinct fetal stages (A and Q) to develop an unsupervised classification model for their identification within an FHR signal. The need for adopting an unsupervised learning approach, which aims at discovering patterns and relationships in data without any predefined guidance (no labelled data as in the supervised case) derives from the absence of an objective external source of information that can be used to identify the labels (i.e., the ground truth).

In particular, in our work, we set up a discrete emissions multivariate HMM with two possible states (A and Q), one for each possible fetal phase. The HMM states are not directly visible, but they are observable through a discrete set of emissions; the latter consists of the discretized values of a set of parameters computed on a moving window running on the FHR signal. Hence, at each shift of the sliding window, the computation of a set of quantitative features, describing different aspects of the FHR signal, is performed. Each variable in the set is then codified so that the parameters' set is in the form of a binary array, that is passed to the trained HMM. The latter gives back the state (A or Q) most likely to have given rise to the observed set of emissions.

The chapter has the following structure: a brief introduction on HMMs is provided in **Section 5.1.2**; **Section 5.1.3** furnishes details about parameters computation and encoding; **Section 5.1.4** illustrates the proposed HMM for the fetal states assessment; **Section 5.1.5** shows details about training and testing of the developed model, while the proposed HMM's results are illustrated in **Section 5.1.6** while **Section 5.1.7** provides a statistical comparison between the identified fetal phases. Lastly, final discussions about the work are presented in **Section 5.1.8**.

5.1.2. A brief introduction to Hidden Markov Models.

Classical Markov models are random stochastic processes in which the probability that rules the transition from one state of the system to another one only depends on the immediately preceding one and not on the whole trajectory of states that have brought to the actual one (memory-lessness property). In standard Markov models, the states of the system are directly observable.
HHMs, on the other hand, are Markov models in which the states are not directly observable, but are inferable through a set of random variables, called observations or emissions, that are probabilistically related to the unobservable states **[5.18]**.

An HMM is defined through:

- A set $\boldsymbol{S}$ of $\boldsymbol{N}$ possible states $\{S_1, \ldots, S_N\}$.
- A set $\boldsymbol{A}$ of $\boldsymbol{M}$ possible emissions/observations $\{a_1, \ldots, a_M\}$.
- A $N \times N$ transition matrix called $\boldsymbol{P}$ such that $P(i, j) = p_{ij} = p(S_j|S_i)$.
- A $N \times M$ emission matrix called $\boldsymbol{E}$ such that $E(i, j) = e_{ij} = p(a_j|S_i)$.
- A $1 \times N$ vector $\pi 0$ of prior probabilities.

The P matrix regulates state changes, while the E matrix rules the probability of a given state to emit a certain symbol (observation). HMMs can present univariate or multivariate emissions. In the first case, a single observation is emitted at each time, while in the second one, a set of observations is issued.

The *P* and *E* matrices are learnt during the training phase through the Baum-Welch algorithm, a special case of the expectation-maximization (EM) algorithm **[5.19]**; the algorithm requires an initial estimate of *π0*, *E* and *P*, which are then updated to the current values during the training.

Once an estimate of matrices *E* and *P* has been obtained, the model can be exploited to make predictions. Specifically, given a particular sequence of observations, the model will use the learnt knowledge to predict the trajectory of states most likely to have given rise to the observed sequence. This task is addressed through the Viterbi algorithm **[5.20]**. The prototype of an HMM is depicted in **Figure 5.2**.

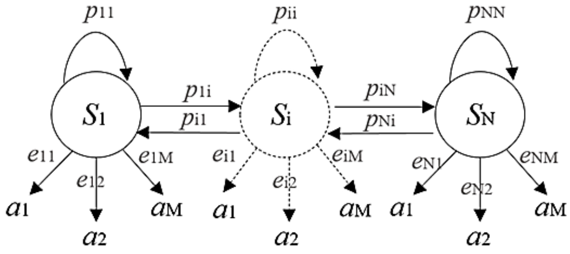


Figure 5.2: Scheme of a generic Hidden Markov Model with N states and M possible emissions for each state. The probability to observe the jth emission in the ith state is ruled by the emission probability eij. On the other hand, the probability of shifting from state i to state j is governed by pij.

5.1.3. Parameters computation and encoding.

The tracings employed to conduct the presented work are from the NAPAMI dataset (see **Chapter 3**). From this huge data collection, we only considered denoised signals for gestational weeks between 30 and 40.

We decided to perform the computation of a set of parameters on 3-minute moving windows (i.e., 360 points of the tracing), shifting along the FHR signal with a 5-second stride (10 signal points).

The parameters we decided to include in our study are those which, based on our prior knowledge, we expect to vary the most between Active and Quiet states.

In particular, the regressor set we considered comprehends two groups of features, the first of which includes parameters that are directly obtained from the FHR signal, and the second of which is formed by a single signal-independent feature, hence not straightly derived from the tracing.

The first group is formed by 5 parameters and includes the variance of the signal or total power (PWT), DELTA, Sample Entropy (SampEn), the power in the very-low frequency range (VLF) and the number of accelerations within the window (Accel.). All features are computed onto the whole three-minute windows except from DELTA, which is obtained only using the points belonging to the central minute of each window, for consistency with its definition **[5.21]**.

PWT and DELTA are two linear measures of variability in the time domain, which are known to increase during fetal activity **[5.8] [5.22]**, DELTA is a commonly used parameter in clinical practice and consists of the difference between the maximum and minimum value of the signal after the application of a low-pass and down-sampling procedure excluding accelerations and decelerations **[5.21]**. Wide variations in terms of this parameter among distinct behavioral states have been reported in the literature **[5.8]**. SampEn is a family of statistical indices that measure regularity, or predictability, by counting the presence of repetitive patterns. SampEn has been shown to increase during quiet states **[5.8]**, **[5.22]**. VLF was computed as the percentage of the power of the detrended signal at frequencies below 0.03Hz. This feature has been shown to vary substantially between active and quiet states in **[5.8]**.

The accelerations are identified according to the definition provided in **[5.23]** as periods longer than 15 seconds during which the FHR remains at least 5 bpm above the baseline and that have a maximum amplitude of at least 10 bpm.

The second group of parameters coincides with the percentage of perceived fetal movements (FMP) inside the window. The FMP signal is the result of the pregnant woman pressing a button, integrated within the system, to indicate the perception of fetal movements during the CTG examination. Technically, the FMP is a signal that has the same duration as the FHR trace, where the perceived movements are indicated by the value 1 and the remaining points have a value of 0. In our study, we refer to FMP as the percentage of perceived movements within the 3-min window.

Accelerations and fetal movements are perhaps the most typical characteristic of activity phases in CTG traces. However, their presence does not automatically indicate that the fetus is in an active state. FMPs, for example, may be present also in the quiet phase 1F, although more sporadically, and could be the result of the mother's misperception **[5.24]**.

The choice of integrating a set of indices derived from the FHR tracing (PWT, DELTA, SampEn, VLF, Accel.) together with a signal-independent feature (FMP) has the intent to allow objective and subjective information to co-participate and work in tandem to the process of identifying fetal stages.

A summary reporting the computed parameters is shown in Table 5.1. Parameters were computed using MATLAB R2022b (The Math Works, Inc.).

Table 5.1: Summary of computed parameters

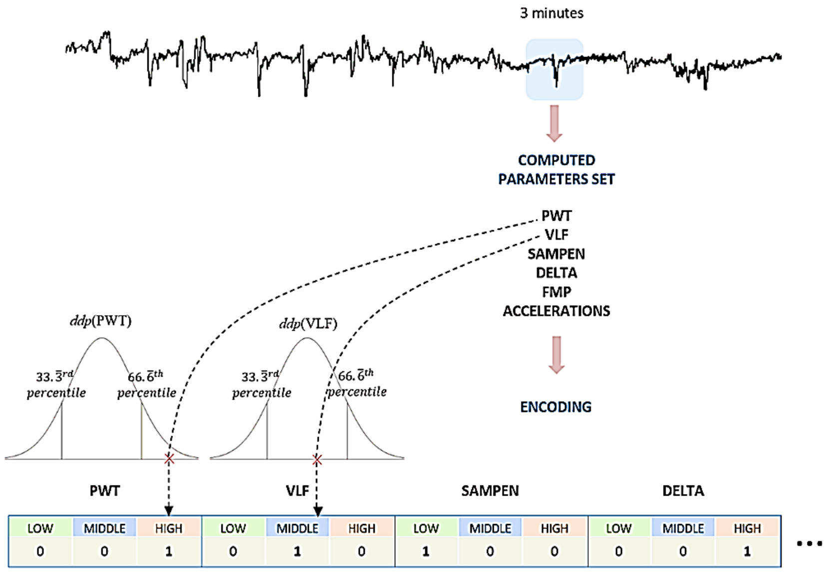
Method	Parameter	Sequence length	Hypothesis
Time domain	DELTA	1 min	Variability of FHR signal in the time domain
	PWT	3 min	
	Accel	3 min	
Frequency domain analysis	VLF	3 min	Quantification of the activity of the autonomic nervous system
Signal regularity and predictability	SampEn	3 min	Presence of recurrent patterns in a single scale
Signal Independent	FMP	3 min	Mother's perception of fetal movement

Once the parameters' set is obtained at each shift of the 3 minutes sliding window, the latter needs to be properly encoded to be passed as input to the HMM.

As previously stated, the idea is to interpret the group of regressors computed at each sliding of the moving window as the set of observable random variables emitted by the actual state of the HMM (i.e., the emissions). These observations are probabilistically related to the system's state through the emission matrix E, which is estimated during the training phase.

Since each parameter is continuous, we decide to put ourselves in a simplified case, by converting each variable to categorical. PWT, VLF, SampEn, DELTA and FMP are categorized according to the 33.3rd and 66.6th percentile values of their respective PDFs, which have been computed considering all the signals in the dataset. Accel., on the other hand, is binarized with respect to the presence or absence of accelerations within the 3-minute window. Thus, each 3-minute excerpt of the FHR signal is converted into a binary vector of 17 symbols, formed by 5 triplets and a final tuple. The first 5 groups of 3 digits respectively indicate low, middle and high values of PWT, VLF, SampEn, DELTA and FMP, and the last group of 2 symbols stand for the presence or absence of accelerations.

For example, let's assume we consider a 3 minutes' chunk of FHR signal, described by high PWT, middle VLF, low SampEn, high DELTA, high FMP and presence of accelerations; the latter will be coded as "0 0 1 0 1 0 1 0 0 0 0 1 0 0 1 1 0" (see **Figure 5.3**).



PWT			VLF			SAMPEN			DELTA			
LOW	MIDDLE	HIGH	LOW	MIDDLE	HIGH	LOW	MIDDLE	HIGH	LOW	MIDDLE	HIGH	...
0	0	1	0	1	0	1	0	0	0	0	1	

Figure 5.3: Schematization of the adopted encoding procedure. Each 3-min excerpt of FHR signal is summarized by a set of 6 parameters. Each of the latter is categorized accordingly to the values of the 33rd and 66th percentiles of their relating PDFs. For example, a 3 minutes' chunk of FHR signal, described by high PWT, middle VLF, low SampEn, high DELTA, high FMP and presence of accelerations; the latter will be coded as "0 0 1 0 1 0 1 0 0 0 0 1 0 0 1 1 0".

5.1.4. The proposed HMM for the fetal state assessment.

The intent of the present work is to demonstrate how HMMs can be used as a tool capable of enabling the unsupervised identification of fetal states within an FHR signal.

In our attempt to the problem, we make the modelling assumption that the system can be described by an HMM with two possible states ($N = 2$), i.e., activity (A) and quiet (Q). The evolutional dynamics of this system are governed by its emission and transition probabilities, which are estimated during the training phase, as will be explained in **Section 5.1.5**. The emission set A is hence composed of 17 possible observations: A= {PWT low, PWT middle, PWT high, VLF low, VLF middle, VLF high, SampEn low, SampEn middle, SampEn high, DELTA low, DELTA middle, DELTA high, FMP low, FMP middle, FMP high, Accel. yes, Accel.no}. This modelling choice hence makes the E matrix of size $N \times M$, and the P matrix of size $N \times N$ where $N = 2$ and $M = 17$. A representation of the developed HMM is depicted in **Figure 5.4**. The HMM was implemented in Python, version 3.7.

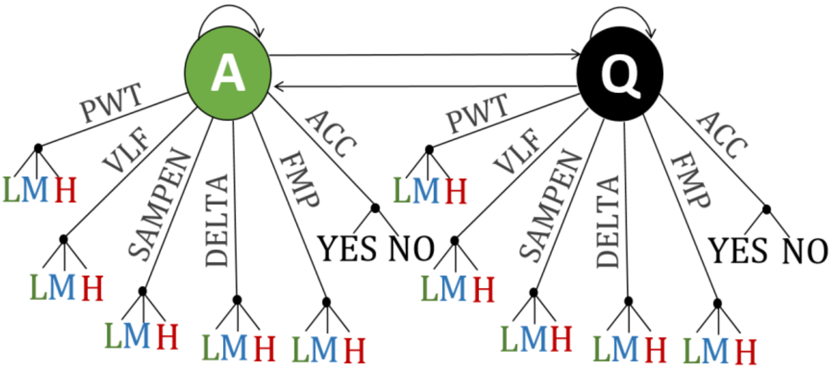


Figure 5.4: Illustration of the HMM structure in the complete case. The developed HMM consists of 2 possible states (Quiet - Q and Activity - A), each of which can generate a discrete set of observations. L, M and H respectively denote low, middle and high values for each discretized parameter.

5.1.5. HMM training and testing.

After defining the structure of the model and coding each 3 minutes' chunk of FHR, as explained in **Sections 5.1.3 and 5.1.4**, we proceed to the estimation of matrices E and P through the Baum-Welch's algorithm. This requires an initial estimate of $\pi0$, E and P, which are then updated to the current values during the training. To avoid any kind of bias due to particular a priori modelling choices, we chose the equiprobability condition for the initialization of $\pi0$, E and P.

For model training, we selected a subset of 9 signals, each containing at least one quiet and one activity stage. The training set thus results in 3'273 3-min excerpts.

To identify the top-performing observation set, we trained several HMMs, one for each possible combination of the parameters set. Excluding the case where no feature is included, since each parameter may or may not be included in the emission set, the number of possible combinations amounts to $2n$ -1 = 63, where n = 6 equals to the number of features.

Once an estimate of matrices E and P has been obtained through the Baum-Welch's algorithm, the model can be exploited to make predictions. Specifically, given a particular sequence of observations, the model will use the learnt knowledge to predict the trajectory of states most likely to have given rise to the observed sequence, through the Viterbi algorithm.

Once each FHR point is classified as belonging to the Active or Quiet phase, a post processing step is carried out. Since Quiet and Active stages are known to have a duration of the order of the tens of minutes **[5.25]**, the label associated with short sub-sequences, totally included in longer excerpts of the opposite state, is reversed.

5.1.6. Fetal states clustering results.

As discussed so far, the goal of the present study is to describe a method based on HMMs for the unsupervised clustering of fetal behavioral states within FHR signals. Since the developed method is completely data-driven, to evaluate the model's capacity to spot activity and quiet stages, we compared the predictions provided by our models with the annotations of an expert clinician. To avoid any kind of bias on both sides, we provided the doctor with a set of 35 unlabeled signals and asked him to annotate them. Meantime we used our HMMs to classify each data point within the same signals supplied to the clinician, which represent our testing set. Then, to assess the degree of concordance, we proceeded to compare the predictions of the HMMs with the annotations of the clinician, which we consider to be the ground truth. **Table 5.2** shows the scores for the 5 best performing HMMs and the ones for the best-performing HMM with the exclusion of FMP, ordered by decreasing accuracy. The green and red dots in **Table 5.2** respectively indicate included and excluded features. The scores reported in **Table 5.2** include the overall accuracy (ACC), the True Active Rate (TAR), the True Quiet Rate (TQR), the False Active Rate (FAR) and the False Quiet Rate (FQR) as defined in **equations** from **5.1 to 5.5**.

$$\boldsymbol{Acc = \frac{TA + TQ}{TA + TQ + FA + FQ}} \qquad \textbf{(5.1)}$$

$$\boldsymbol{TAR = \frac{TA}{TA + FQ}} \qquad \textbf{(5.2)}$$

$$\boldsymbol{TQR = \frac{TQ}{TQ + FA}} \qquad \textbf{(5.3)}$$

$$FAR = \frac{FA}{TA + FQ} \tag{5.4}$$

$$FQR = \frac{FQ}{TQ + FA} \tag{5.5}$$

Table 5.2: Top 5 performing HMMs + #1 without FMP

		Models					
		HMM#1	**HMM#2**	**HMM#3**	**HMM#4**	**HMM#5**	**HMM#1 without FMP**
FEATURES	**PWT**	🔴	🟢	🔴	🔴	🔴	🔴
	VLF	🟢	🟢	🔴	🟢	🟢	🟢
	SampEn	🟢	🟢	🟢	🔴	🟢	🟢
	DELTA	🟢	🟢	🟢	🟢	🔴	🟢
	FMP	🟢	🟢	🟢	🟢	🟢	🔴
	Accel.	🟢	🟢	🟢	🟢	🟢	🟢
SCORES	**ACC**	**0.9**	0.88	0.88	0.87	0.85	0.84
	TAR	**0.9**	0.86	0.89	0.85	0.82	0.80
	TQR	**0.87**	0.93	0.87	0.92	0.93	0.86
	FAR	**0.09**	0.14	0.11	0.14	0.17	0.20
	FQR	**0.13**	0.06	0.12	0.08	0.07	0.14

The obtained results indicate that HMM#1, represented in the first column of **Table 5.2**, stands out with the highest overall accuracy of 0.90. It includes VLF, SampEn, DELTA, FMP, and Accel. Its outperforming ACC indicates a strong concordance between the predictions of HMM#1 and the annotations made by the clinician. HMM#1 also demonstrates a high TAR of 0.90, suggesting its effectiveness in correctly identifying active states.

Comparing HMM#1 to HMM#2, which includes the whole set of computed parameters, we observe that HMM#1 has a higher accuracy (0.90 vs. 0.88) and a higher TAR (0.90 vs. 0.86). This indicates that HMM#1 is more successful in accurately identifying active states. However, HMM#1 has a lower TQR compared to HMM#2, implying that it may struggle somewhat in accurately identifying quiet phases.

HMM#3, which excludes PWT and VLF, on the other hand, achieves a TAR of 0.89 and a TQR of 0.87. These values are comparable to those of HMM#1, indicating that HMM#3 is also capable of correctly identifying both active and quiet states with a high degree of accuracy, at the expense of FAR.

HMM#4, which doesn’t comprise PWT and SampEn, exhibits slightly lower performance than HMM#1 and HMM#3, with an accuracy of 0.87 and a TAR of 0.85. However, it shows a higher TQR of 0.92, suggesting its ability to accurately identify quiet phases.

HMM#5, which doesn’t include PWT and DELTA, achieves the lowest overall accuracy of 0.85. However, it still presents a TQR of 0.93 showing off a high discriminative power in identifying quiet stages.

The choice of the most suitable HMM depends on the specific requirements and priorities of the analysis, considering factors such as the desired balance between accurately identifying both active and quiet states.

In our specific case we are mostly interested in maximizing the classification accuracy, since our goal is assessing the HMMs' ability to correctly classify both active and quiet states, providing a reliable measure of the model's overall performance. ACC, in fact, provides an overall assessment of the concordance between the HMM predictions and the clinician's annotations, giving us a comprehensive understanding of the model's effectiveness.

The transition matrix *P* for the most accurate HMM (i.e., HMM#1) is reported in **Table 5.3**.

Table 5.3: Transition Matrix P for HMM#1

State	***Quiet***	***Activity***
Quiet	**0.985**	0.015
Activity	0.018	**0.982**

The analysis of *P* provides important insights into the system's dynamics since its values indicate the probability of the system to transit from one state to another.

The high values on the main diagonal of *P*, hence, indicate the tendency of the system to remain in the actual state for long periods rather than rapidly shifting to the other one. In fact, the probabilities associated with remaining in a quiet or active state are respectively 0.985 and 0.982; on the other hand, the probability of shifting to an active state when the system is in a quiet phase is very low (0.015) and so is the probability of passing from an active stage to a quiet one (0.018).

This kind of behavior is consistent with what is generally observed in clinics, since both active and quiet phases are known to have a duration in the order of tens of minutes **[5.13]**. This means that once the fetus enters either the quiet or active phase, it tends to persist in that state for a considerable amount of time before transitioning to the other state.

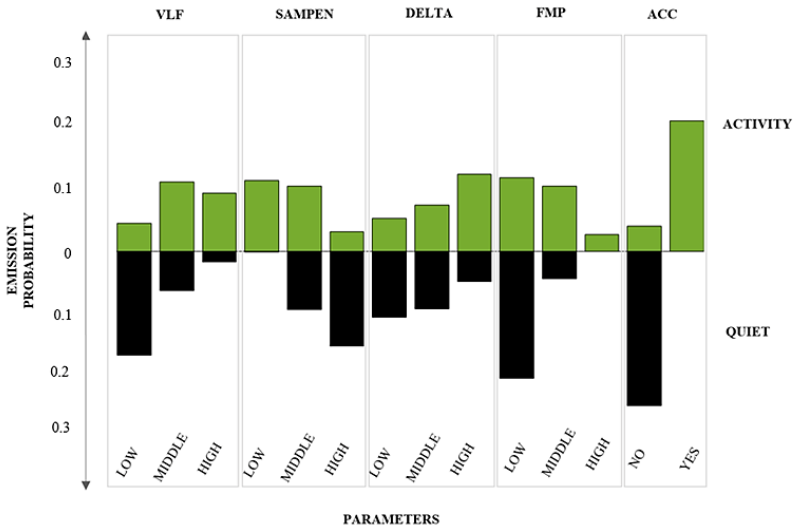


Figure 5.5: Illustration of the HMM structure in the complete case. The developed HMM consists of 2 possible states (Quiet and Activity), each of which can generate the discrete set of observations.

A visualization of the observation probabilities within the emission matrix E of HMM#1 is shown in **Figure 5.5**, which provides insights into the relationship between the probabilities associated with different feature values and the corresponding fetal states. Green and black bars in **Figure 5.5** stand to indicate the probabilities associated to each emission of the HMM, respectively in activity and quiet.

From the observation of **Figure 5.5** we can ascertain how high probabilities associated with high values of VLF, DELTA, FMP, Accel., together with low values of SampEn are more likely to reflect a state of fetal activity, rather than a quiet one. On the other hand, high probabilities associated with low values of VLF, DELTA, FMP, Accel. and high values of SampEn are very likely to be related to a quiet phase.

The analysis of **Table 5.3** combined with the findings depicted in **Figure 5.5**, sheds light on the heavy impact of Accel and FMP, which appear to be the most important features in the clustering process.

To evaluate the contribution of FMP, which is the only signal-independent feature, we compared the scores obtained by HMM#1 with the ones obtained with HMM#1 with the exclusion of FMP.

The inclusion of FMP in HMM#1 resulted in an improvement in the classification accuracy by approximately 6%. This suggests that FMP contributes positively to the overall performance of the model in accurately classifying fetal states.

However, it is noteworthy that even when FMP was excluded from the feature set, the obtained scores remained acceptable, particularly in terms of True Positive Rate (TPR). This indicates that the other features included in the regressor set are still capable of capturing relevant information and effectively distinguishing between fetal activity and quiet phases.

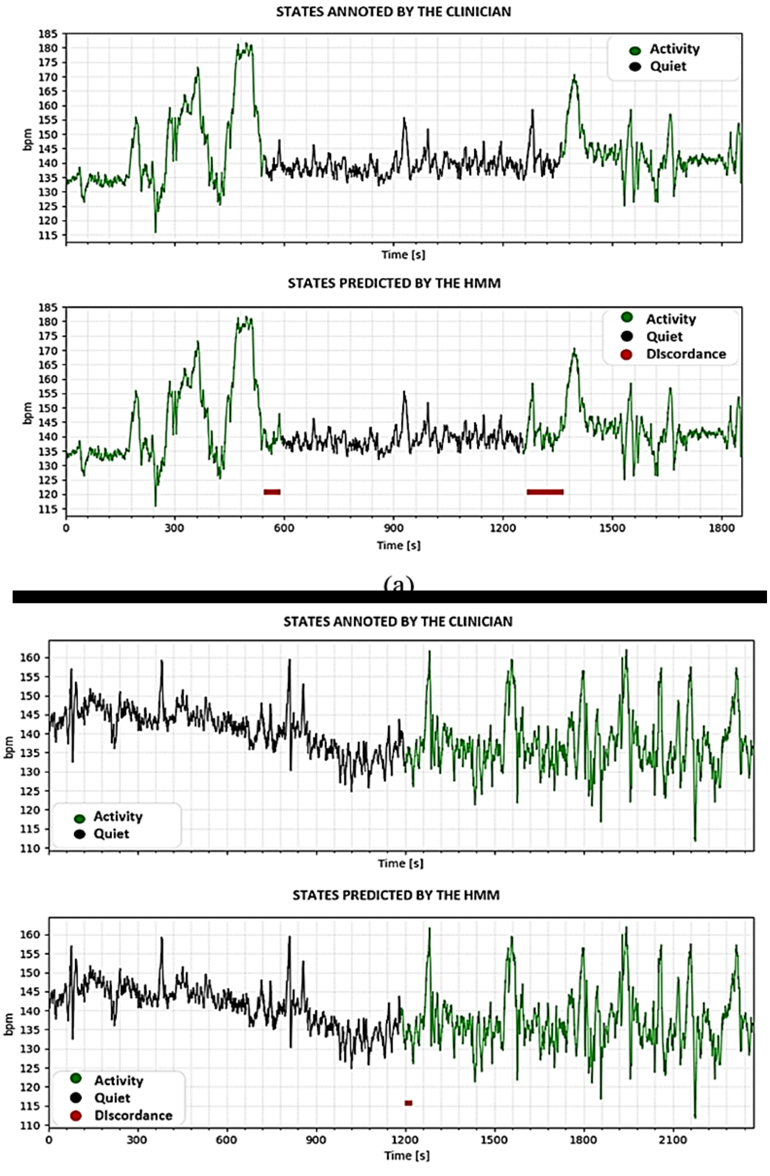
Figure 5. 6 illustrates a visual comparison between the annotations made by the clinician and the predictions generated by HMM#1. Two exemplary signals are shown, where the green and black points represent Active and Quiet

phases, respectively. Additionally, short red horizontal lines are used to indicate the segments in which the clinician's annotations differ from the model's predictions.

It is necessary to remark again that the labeling performed by the clinician was done on signals not annotated by our model, so that his judgment was not influenced by the predictions obtained from the HMM. **Figure 5.6 (a)** illustrates a signal presenting a prolonged phase of fetal quiet within two active stages. By comparing the HMM predictions with the labels provided by the clinician we can assess a high degree of concordance. The only equivocal portions are concentrated in areas that straddle two distinct phases.

Figure 5.6 (b), shows instead a signal which exhibits an initial quiet phase, lasting about 20 minutes, followed by an active stage approximately of the same duration. Even in this case, the model's predictions are aligned with the clinician's annotations, and the degree of concordance is still higher than in the preceding example.

From the observation of **Figure 5.6 (a) and (b)** it's possible to appreciate that portions labelled as quiet tend to exhibit lower variability and a lower tendency to accelerate as compared to activity phases. This tendency is consistent with what observed in **Figure 5.5**, since parameters associated with signal's variability (i.e., SampEn and DELTA) tend to assume higher values when the actual state is active. On the other hand, the more likely presence of accelerations within active phases determines a more predictable dynamic of the system. Again, this is consistent with what was observed from the inspection of E since low values of SampEn are probable to reflect a condition of fetal activity.



(a)

(b)

Figure 5.6: comparison between the predictions of the best performing HMM (HMM #1) and the clinician’s annotations for two example signals (a) and (b). Green and black points respectively denote Active and Quiet phases. The red horizontal lines in predictions plots underline the portions in which the physician's notes deviate from the model's predictions.

5.1.7. Characterizing statistical differences between Active and Quiet phases.

In the previous section, we presented the developed method for clustering Active and Quiet phases within FHR signals, which is based on a categorical HMM. Once validated, according to clinician's interpretations, the aforementioned algorithm was employed to explore statistical differences between the Quiet and Active stages. For that aim, the developed HMM was used to cluster fetal phases within the whole set of tracings contained in NAPAMI (i.e., 24'491 records). From the wider set of parameters explained in **Chapter 4**, we selected a subgroup of 14 features, which we retain to be the most suitable to grasp the characteristics of Quiet and Active stages and to provide an appropriate description of the latter, in terms of statistical properties. In particular, we decided to include PWT, LF, MF, HF, LZ2, LZ3, SAMPEN, LTI, DELTA, STV, II, FMP%, VLF and percentage of points in which the signal is accelerating while being in Active and Quiet phase (%pts ACC). Please refer to the Appendix section for a more in-depth description of each index.

To assess significant differences in terms of computed parameters' central tendency, within Active and Quiet phases, hypothesis testing was performed with a significance level $\alpha=0.05$. First, Kolmogorov-Smirnov test (KS) was applied to verify the null hypothesis (H0) of normally distributed PDFs. Since H0 was refused for each parameter, the non-parametric Mann Whitney U-test (MW) was employed to infer differences in terms of central tendency between features, respectively describing Active and Quiet phases.

A summary reporting median values for features in Active and Quiet phases is shown in **Table 5.4** while boxplots are illustrated in **Figure 5.7.**

Table 5.4: median values for parameters in A and Q. Bold numbers indicate higher values for each row.

	Active	Quiet
PWT	**242.6**	69.5
LF	40.0	**48.2**
MF	5.3	**8.7**
HF	1.4	**2.9**
LZ2	1.02	**1.05**
LZ3	0.9	**0.95**
SAMPEN	0.72	**1.06**
LTI	**29.6**	15.5
DELTA	**44.6**	28.2
STV	**6.8**	4.4
II	**0.85**	0.83
FMP %	**0.17**	0.07
VLF	**52.7**	38.4
% pts Acc	**0.1172**	0.001

From the analysis of **Figure 5.7** and **Table 5.4** we can observe very different behaviors between the two stages in comparison. In particular, parameters associated to signal's variability (i.e., PWT, DELTA, II, LTI, STV) assume significantly higher values in the Active stage as compared to those in Quiet stage. This suggests that the signal in the Active phase is more variable, tending to present more fluctuations compared to the Quiet stage. This is also confirmed by the percentage of points in which the signal is accelerating while being in Active and Quiet phase (%pts ACC); in fact, %pts ACC is significantly higher during the Activity (A: 11.72% vs Q: 1%), reflecting the signal's tendency to accelerate while being in phase A. On the other hand, features associated to signal’s regularity (i.e., LZ2, LZ3, SAMPEN) tend to assume slightly higher values in Q, revealing that the system presents a bit more predictable behavior in Quiet phase, as compared to Active one, where the trend is less regular.

For what concerns instead the percentage of HF, MF, HF, it’s possible to observe lower values in Activity, rather than Quiet, as the slow oscillations resulting from accelerations fall in the VLF band.

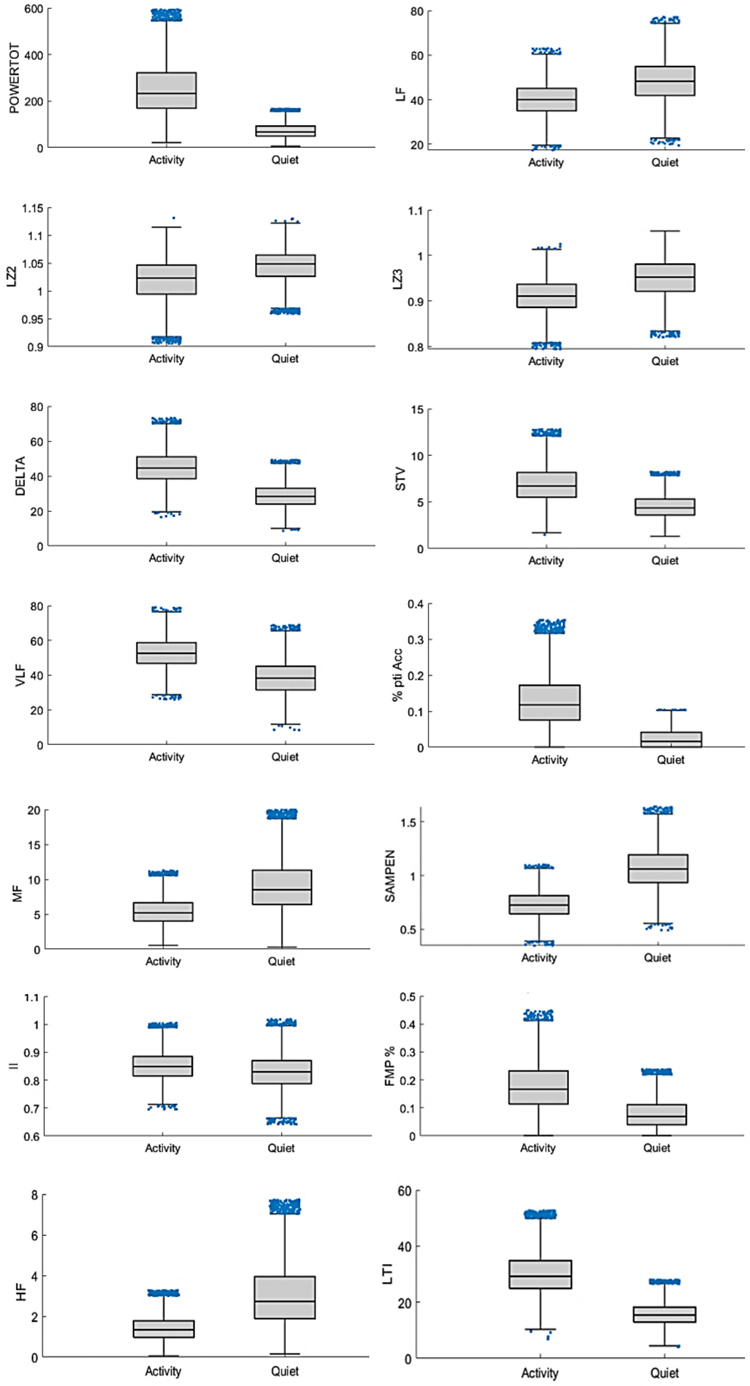


Figure 5.7: Boxplot for each computed parameter in Active and Quiet stages. Blue dots represent outliers.

To visually explore the spatial distribution of Active and Quiet phases, Principal component Analysis (PCA) was employed as a dimensionality reduction technique. To this aim, we started considering the set of 14 parameters, described above, for each of the labelled (through the aforementioned HMM) 3-mins non-overlapping chunk composing each signal of NAPAMI. PCA was hence applied and the first three principal components where considered, since they explain more than the 80% of variance. So doing, each FHR subsequence of 3 minutes is converted into a single point in a three-dimensional space, whose axes represent the three first PCA components (PC1, PC2, PC3). A 3-D Scatterplot representing the spatial distribution of points summarizing each 3-minutes chunk is depicted in **Figure 5.8**. Green and black color in **Figure 5.8** respectively represent active and quiet points, accordingly to the classifications provided by the HMM above described.

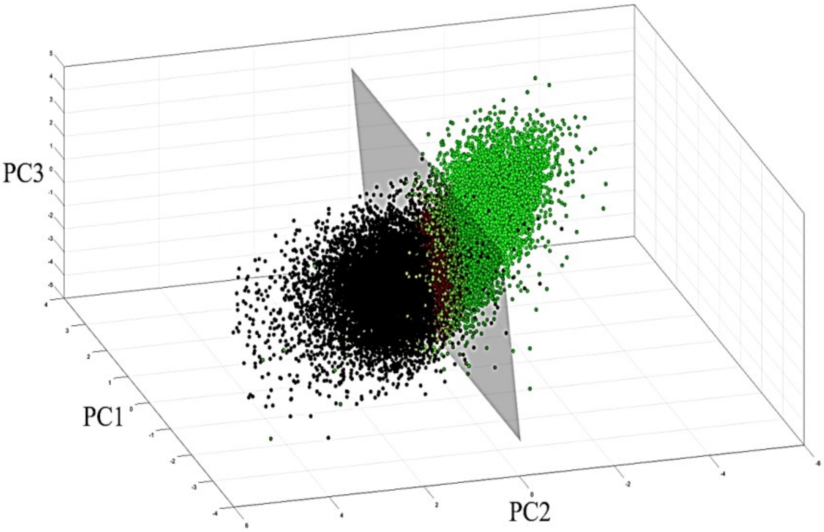


Figure 5.8: 3-D Visualization of first three principal components. Each 3-D point summarizes the properties of a 3-minute chunk of FHR tracing. Green points denote signal excerpts classified as Active by the HMM, while black ones are instead relating to those labelled as Quiet. Black and green dots are linearly separated by the gray hyperplane.

Observing the spatial distribution of points relating to active and quiet phases we can observe two slightly overlapping clusters, which are quite well linearly separable. In fact, it's possible to notice how they can be split into two separate groups through a plane boundary (see gray Hyperplane in **Figure 5.8**).

5.1.8. Discussions and conclusions on the proposed method.

In this section, we illustrated our proposed method for unsupervised FHR signal clustering, based on Hidden Markov Models (HMM), to automatically identify fetal behavioral states of quietness and activity within CTG tracings. More specifically, the developed model is a Multivariate HMM with categorical emissions.

The obtained results seem to suggest that the developed model can identify quiet and activity states with a good degree of reliability. In fact, the predictions of the model have shown, for the best model, a degree of agreement of 90% with the interpretations of an expert clinician.

A desirable feature of our method is the ease of interpreting the obtained results. Indeed, the choice of a small set of parameters, already used in computerized Cardiotocography, with a clear intrinsic meaning, provides explainability to this approach.

Another appealing feature of our approach is that the learning process is fully unsupervised. In fact, the states identified by the model using the Baum-Welch algorithm are associated with the “Active” and “Quiet” states only a-posteriori. The choice of using a totally data-driven approach was suggested by the need to find a classification method able to go beyond the clinician's interpretations. Letting the model autonomously identifying the clusters, rather than relying on expert annotations, removes the dependence on the annotator's choice, thus hopefully reaching a more objective evaluation that is only based on the intrinsic characteristics of the signal. Indeed, in the absence of a more objective external source of information that can be used to identify the labels, e.g., ultrasound, the advantages of a supervised approach are limited.

Despite the achieved results seem to suggest that the approach can be effectively used as a tool to cluster active and quiet fetal stages, it should be considered that just a limited amount of data was available to test the performance of the proposed HMM. The limited quantity of testing data can be primarily attributed to the time constraints faced by the participating clinician. The time required to increase the testing set of one order of magnitude would have caused an unacceptable delay in the publication of our work. Looking ahead, we are firmly committed to expanding our research efforts with the aim of gathering more extensive testing data. This could even benefit from the enrollment of a larger cohort of expert clinicians, boosting the truthful of testing ground truth. By doing so, we seek to enhance the robustness and generalizability of our results.

Future developments of the present work include defining different models tuned for gestational age and, for the last weeks of gestation, moving to a four-state model. Indeed, a model capable of distinguishing among all four behavioral states would be of great interest, although it has been shown in [4] that these states emerge with reasonable reliability only at the very end of the pregnancy.

We believe that the proposed method could represent a noticeable enhancement for the computerized analysis of the non-stress test. Indeed, the alternation of behavioral states is by itself of clinical interest and is an important pre-processing stage for the interpretation of CTG parameters that is often overlooked. Indeed, it has been clearly shown that CTG parameters widely vary between behavioral states. Reporting only their mean value without taking into account the states on which they were averaged could at least explain the large variability observed in CTG parameters, even within physiological pregnancies.

It is worth noting that including the FMP signal in the analysis only marginally increases the performance of the model, which can be successfully adopted by using the FHR signal alone.

Our method is applicable also when the FHR is extracted employing other methods rather than CTG, such as non-invasive electrophysiology. A compelling use case could be the analysis of behavioral states in long-term recordings, which could be very interesting for monitoring several pregnancy complications **[5.3]**, from Intra Uterine Growth Restriction **[5.4]**, diabetes **[5.7]**, **[5.26]**, or hypertension **[5.5]**, since all these conditions have been shown to have an impact on behavioral states.

5.2. Fetal heart rate spectral analysis.

The fetal heart rate (FHR) is a readily available source of physiological information **[5.27]**.

As anticipated in **Chapter 3**, an improvement to the FHR analysis came when several biomarkers were proposed in the literature together with systems for the computerized analysis of the signals. This allowed to overcome some of the limitations linked to the simple eye inspection of the traces **[5.28-5.33]**.

In fact, quantitative analysis of the FHR signal allows the identification of important characteristics that can go missed by visual inspection and ensures the reproducibility of the analysis.

Despite the advancements made in the last years, however, biomarkers of the most common fetal pathologies, e.g., intra-uterine growth restriction (IUGR), still have limited reliability **[5.34-5.35]** and most have been tested on a very limited number of subjects.

The rationale behind the study of the FHR antepartum is that disturbances in the normal intrauterine development lead to changes in the autonomic functions that are observable from the cardiovascular regulation **[5.36]**. Monitoring these changes is important both in perinatal medicine for their prognostic and diagnostic value and in the framework of the "developmental origin of health and disease" **[5.37]** for the possibility to predict disturbances later in life. Among the parameters presented in the literature, those that allow quantifying the frequency distribution of the oscillations in the heart rate variability (HRV) have the desirable property of being physiologically interpretable, since it has been shown that the sympathetic and parasympathetic branches of the autonomic nervous system (ANS) influence the HRV at different frequencies **[5.38]**.

When analyzing the power spectral density (PSD) of the FHR signal three main bands are typically defined in analogy with those employed for adults **[5.39]** and specifically adapted to the fetal case. These include the Very Low Frequency (VLF), Low Frequency (LF) and High Frequency (HF) bands. In addition, some authors also consider an additional frequency band, called movement frequency (MF), between LF and HF **[5.30]**. The exact frequency ranges vary between different authors and have been summarized in **[5.40]**. In general, the power in VLF is related to long period and non-linear contributions and gross body movements, LF with mainly sympathetic activity and HF with parasympathetic activity and fetal breathing **[5.30]**. MF has been hypothesized to be related to fetal movements and maternal breathing but overlaps with what other authors consider HF **[5.40]**.

The FHR signal presents characteristics that complicate its analysis in the frequency domain. The variability of the FHR and its frequency distribution change over time. While in adults the experimental conditions can be easily controlled (for example asking the patient not to move and breathe at a controlled rate) this is impossible to apply in the fetal case, due to frequent changes in behavioral states **[5.41]** (**Chapter 5.1**). It results that the FHR signal is inherently non-stationary, which must be addressed in frequency analysis. Moreover, it is expected that some oscillations in the FHR, like the ones induced by respiratory movements, are transient and possibly not phase-synchronized, which means they may not be captured by standard spectral analysis.

Lastly, FHR traces are often very noisy and subject to signal loss.

Since the task of analyzing the frequency content of the FHR signal is not trivial, several methodologies have been applied **[5.42-5.43]**. In particular, we expect that some of these may be more effective in detecting oscillations even when the assumptions of classical spectral analysis do not hold (more noticeably, stationarity and linearity). These techniques can be broadly classified as parametric (typically autoregressive - AR) and non-parametric (based on the Discrete Fourier Transform -DFT or the Hilbert–Huang transform) and may use explicit windowing (such as the Short Time Fourier Transform -STFT) or not (such as the Continuous Wavelet Transform-CWT).

More recently, the Phase-Rectified signal averaging (PRSA) technique has been proposed to detect quasi-periodicities in non-stationary signals **[5.44]**. The PRSA is not a method for spectral analysis itself, but rather a technique that produces a compressed version of the original signal (i.e., the PRSA curve) in which the noise is smoothed-out and (quasi) oscillations are highlighted, even in the presence of phase-resetting.

Several works employing measures extracted using PRSA have been applied to FHR analysis. In particular, acceleration and deceleration capacity **[5.45-5.47]**, acceleration and deceleration phase rectified slope **[5.48]** and deceleration reserve **[5.49]**. Some authors postulate that the deceleration capacity is a measure of the vagal control of the heart rate and the acceleration capacity is a measure of sympathetic activity **[5.50]**. However, this assumption has been challenged in **[5.51]** and **[5.49]**.

In this chapter, we illustrate our proposed approach for the analysis of the PRSA curve, derived from FHR signals. Similarly, to what proposed by Bauer et al. **[5.44]**, we perform the CWT of the PRSA curve but evaluate the relative distribution of the oscillations in the frequency domain instead of evaluating it only at specific scales, like it is done when calculating acceleration and deceleration capacity.

The proposed approach is compared to four more traditional methods used to perform spectral analysis of FHR signals from CTG recordings **[5.43]**, i.e., DFT, AR modelling, CWT and Empirical Mode Decomposition (EMD).
This section describes the performance of the proposed method and of the other existing in the detection of changes in the FHR due to two very common complications of pregnancy, which are expected to produce changes in the FHR signal characteristics: Intrauterine growth restriction (IUGR) and gestational diabetes (GDM). For the analyses, we used a very large database of antepartum CTG recordings collected at different gestational ages.

5.2.1. Description of the used data.

The present study has been conducted by exploiting NAPAMI (**Chapter 3**).
In this study, we selected the first 20 minutes of each recording with less than 10% of interpolated points. Only tracings relating to fetuses between the 32nd and 38th gestational weeks have been considered. Then, three distinct populations were considered: Controls, i.e., physiological pregnancies without known maternal or fetal pathologies, IUGRs (diagnosed when the fetal weight is lower than the 10th percentile for gestational age and are present alterations in the umbilical artery flow, in agreement with **[5.52]**), and GDMs (diagnosed following a positive 1-step glucose tolerance test **[5.53]**). For each week, we selected a subset of equal size from each population, randomly down sampling from the largest groups. The final dataset thus contains the same numerosity for the three analyzed groups in each week and includes a total of 2,178 recordings.

5.2.2. Classical methods for the time-frequency analysis.

Among the traditional methods to perform spectral analysis we selected: DFT, AR modelling, CWT and EMD. All these methods aim to estimate the PSD, albeit in different ways. In this paragraph, we briefly outline some details about their implementation and their differences. DFT and AR are based on an explicit windowing, while CWT and EMD are time-varying algorithms.

Explicit windowing methods: DFT & AR

The DFT and AR are the most common methodologies to estimate the PSD. Since both assume stationarity, the FHR signal was divided into windows of 2 minutes overlapped by 1 minute, which is a compromise between spectral resolution and the fulfilment of the stationarity condition. Inside each window, we removed the linear trend from the signal. The DFT was estimated by applying the Fast Fourier Transform and directly used to estimate the PSD. Parametric spectral analysis with AR models was performed as reported in **[5.30]**.

Time varying algorithms: CWT & EMD

Time-varying algorithms, i.e., the CWT and the EMD, are expected to provide better time-frequency resolution as compared to methods that employ explicit windowing. Moreover, since the EMD assumes neither stationarity nor linearity of the signal **[5.54]**, it appears to be particularly suitable for the analysis of FHR. Prior to their application, the signal was detrended by removing its moving average, computed over shifting windows of 1 minute and padded with a periodized extension of 240 samples to reduce distortions at the borders.

For the computation of the CWT, we employed the Morlet Wavelet with non-dimensional central frequency equal to 6. The scales "*s*" were defined according to **equation 5.6**.

$$\begin{cases} \delta_j = J^{-1} \cdot \log_2\left(\frac{N \cdot \delta_t}{s_0}\right) \\ s_j = s_0 \cdot 2^{j\delta_j},\ j = 0, 1, \ldots, J \end{cases} \quad (5.6)$$

where N is the length of the signal and δ_t the sampling period (i.e., 0.5s). s_0 was set to 1s ($2 \cdot \delta_t$) and J to 179. Wavelet software was provided by C. Torrence and G. Compo **[5.55]** and is available at the URL: http://paos.colorado.edu/research/wavelets/.

Spectral features, i.e., LF%, MF% and HF% are computed by integrating the spectra (or spectrograms) over the frequency and considering the mean over time. Segments with more than 5% of interpolated points were excluded from averaging. We consider the following frequency bands: LF (0.03 – 0.15 Hz), MF (0.15 – 0.5 Hz) and HF (0.5 – 1 Hz), following the definition provided in **[5.30]**. Spectral features are expressed as percentages of the total power. We do not report the values of VLF, which can be trivially obtained from the others. Indeed, what we are interested in quantifying in this study is the frequency distribution of the signal variability, rather than the variability itself.

5.2.3. PRSA spectrum evaluation: classical computation and the proposed method

In this section, we briefly describe the PRSA technique, which was introduced by Bauer et al. **[5.44]**; than follows the description of the method that has led to the computation of the PRSA spectrum **[5.57]** and the proposed approach for its evaluation and extraction of spectral features.

The first step to obtain the PRSA curve is finding the deceleration anchor points (x_{dec}), which are defined as samples satisfying the condition (**equation 5.7**):

$$\frac{1}{T}\sum_{i=0}^{T-1} x[t+i] > \frac{1}{T}\sum_{i=1}^{T} x[t-i] \quad (5.7)$$

Where x is the FHR series expressed in milliseconds (**Figure 5.9 a**). For each x_{dec}, a window of length $2L$ is defined taking the values of the original signal that go from x_{dec-L} to $x_{dec+L-1}$. The PRSA curve (x_k^{PRSA}) is constructed by averaging all these windows (**Figure 5.9 b**). This procedure highlights components that are phase-synchronized with the anchor points and cancels out the others. In this study, T was set to 1 sample (i.e., 0.5 seconds). We acknowledge that the choice of such a small value of T can make the procedure more sensitive to noise, but selecting a bigger value would filter the power in the HF band, which is of interest in this study. Indeed, the larger the value of T, the lower the frequency of the components that are highlighted by the PRSA **[5.58]**. To reduce the influence of noise, signal samples with poor quality were prevented from being anchor points and were not included in the averaging procedure.

Once the PRSA curve was obtained, to compute the PRSA spectrum (PRSA_Spt) we computed the scalogram using the CWT according to **equation 3** in a similar fashion to **[5.57]**:

$$\boldsymbol{X_w^{PRSA}(s,p) = \sum_{k=-L}^{L-1} x_k^{PRSA} \cdot \frac{w[(k-p)/s]}{s}} \qquad (5.7)$$

Where s is the scale, p is the position and w is the mother wavelet. We employed the analytic Morse wavelet with γ equal to 3 and time-bandwidth product equal to 60 and applied L1 normalization. The spectrogram is obtained as the square of the wavelet coefficients (**Figure 5.9 c**) and is evaluated at $k=0$ (**Figure 5.9 d**), thus obtaining a single spectrum.

As discussed in **[5.44]**, the PRSA_Spt presents relevant differences compared with conventional spectral analysis. The signal-to-noise ratio is improved by two effects:

Short patches of periodicities with a particular frequency cancel out in conventional spectral analysis while, due to the way it is defined, they all contribute to the PRSA curve and, therefore, to its spectrum.

A sinusoidal component of amplitude A_f produces an oscillation proportional to $A_f^2 \cdot f$ in the PRSA curve. It derives that while its contribution in the conventional power spectrum is proportional to A_f^2 in the PRSA_Spt is proportional to $A_f^4 \cdot f^2$. Therefore, a $1/f$ noise has a flat PRSA spectrum.

Both these properties are useful in the analysis of FHRV. Indeed, oscillatory patterns in the FHR are usually transient and not phase synchronized. Moreover, most of the variability is contained at very-low frequencies, consistently with a long-term correlated series, rendering it difficult to identify high-frequency oscillations that, despite their smaller amplitude, may offer important information on fetal physiology.

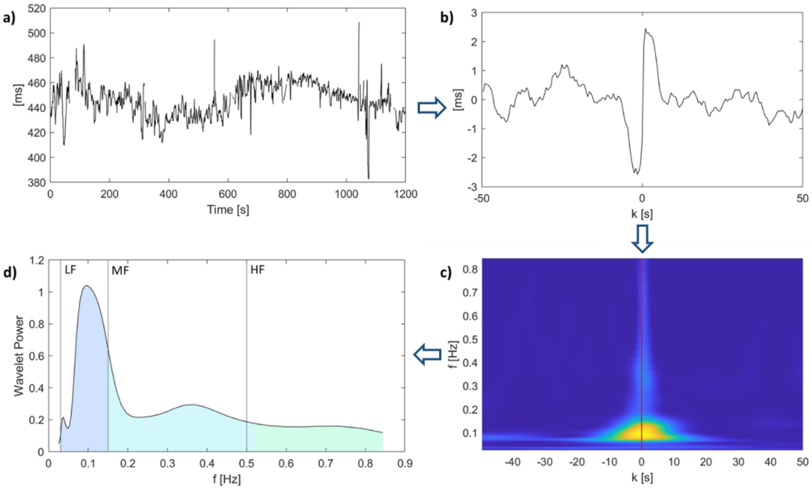


Figure 5.9: Illustration of the technique employed to compute the PRSA spectrum. a: FHR signal; b: PRSA curve; c: CWT spectrogram of the PRSA curve obtained by squaring wavelet coefficients; d: CWT spectrum at k=0 (section of the spectrogram).

The method we present in this study is a variation of the one presented in **[5.57]** to evaluate the PRSA_Spt. Here, instead of considering the Wavelet coefficients at a single scale, we propose to quantify the distribution of the oscillations integrating the PRSA_Spt (**Fig. 5.9 d**) in the frequency bands used in traditional spectral analysis (i.e., LF MF and HF). The features considered are ultimately the percentages of power obtained integrating the PRSA_Spt in the frequency bands described previously.

It should be noted that spectral features computed after the application of PRSA can no longer be regarded as an estimate of the distribution of the signal variance in the frequency domain. Rather, it is a measure of the localization in frequency of the oscillations that survive the PRSA procedure evaluated in descending signal segments.

This approach allows using the distribution of the oscillations in the frequency domain to estimate the fetal autonomic activity, rather than the fact that they are aligned around portions in which the signal increases or decreases. Therefore, unlike the other features that can be extracted from the PRSA, the ones proposed should not suffer from the limitations in the interpretation of their physiological meaning which have been pointed out in **[5.51]** and **[5.49]**.

All analyses were conducted using MATLAB R2022a.

5.2.4. Results.

Recordings have been divided into two groups: from week 32 of gestation to week 36 (pre-term) and from 37 to 38 (early-term), since considerable differences in fetal maturation are expected between the two periods. The first includes a total of 1,161 recordings and the second 1,017.

Table 5.5 reports values in relative spectral power obtained in the first and in the second group of weeks respectively in the first and second column. It can be noticed that the values achieved from classical methods of spectral analysis did not show substantial differences among each other. On the other hand, the PRSA_Spt accentuates higher frequency components, a result consistent with the scaling behavior of the PRSA_Spt. As a result, the PRSA_Spt reports higher values for MF% and HF% compared to the other methods. Moreover, employing classical spectral analysis a considerable portion of the total power is contained in the VLF band (i.e., f<0.03 Hz) while the power in this band is almost entirely filtered-out by the PRSA.

Table 5.5: Medians and quartiles of the spectral features at weeks 32-36 (first columns) and 37-38 (second column). All values are reported in percentages of the total power.

	Weeks 32-36			**Weeks 37-38**		
		LF%			**LF%**	
	Control	GDM	IUGR	Control	GDM	IUGR
FFT	39 (33-45)	39 (33-46)	37 (32-44)	37 (32 -43)	38 (33 -44)	37 (32 -44)
AR	35 (25-48)	35 (25-48)	32 (25-43)	32 (22 -41)	33 (25 -44)	33 (24 -43)
CTW	40 (34 -46)	41 (35 -46)	40 (35 -46)	38 (33 -44)	39 (35 -45)	39 (34 -46)
EMD	33 (28 -39)	34 (29 -40)	33 (28 -39)	32 (27 -38)	33 (28 -39)	33 (28 -28)
PRSA _Spt	**38 (25 -52)**	**36 (22 -50)**	**37 (23 -50)**	**38 (24 -53)**	**37 (24 -53)**	**36 (26 -51)**
		MF%			**MF%**	
	Control	GDM	IUGR	Control	GDM	IUGR
FFT	5.6 (4.0-7.0)	6.4 (4.5-8.7)	6.0 (4.5-7.9)	6.0 (4.5 -7.8)	6.3 (4.7 -8.1)	5.9 (4.7 -8.2)
AR	4.3 (3.0-5.8)	5.2 (3.5-7.1)	5.0 (3.6-6.7)	4.7 (3.4 -6.4)	5.1 (3.6-7.0)	4.9 (3.5-4.8)
CTW	5.3 (3.9-7.2)	6.5 (4.5-8.7)	6.3 (4.6-8.7)	5.9 (4.4-7.9)	6.5 (4.5-8.5)	6.1 (4.7-8.5)
EMD	5.5 (4.2-7.1)	6.6 (5.0-9.1)	6.4 (4.9-8.3)	6.4 (4.9-8.5)	6.6 (4.9-9.0)	6.6 (4.9-9.0)
PRSA _Spt	**20 (16-25)**	**30 (22-37)**	**28 (20-35)**	**25 (19 -33)**	**29 (22 -37)**	**28 (21 -36)**
		HF%			**HF%**	
	Control	GDM	IUGR	Control	GDM	IUGR
FFT	1.8 (1.2-2.7)	1.7 (1.1-2.5)	1.5 (1.1-2.4)	1.7 (1.1-2.5)	1.7 (1.2-2.5)	1.6 (1.1-2.3)
AR	1.8 (1.2-2.8)	1.9 (1.2-2.8)	1.5 (1.1-2.4)	1.7 (1.1-2.5)	1.6 (1.2-2.4)	1.6 (1.1-2.3)
CTW	1.7 (1.1-2.4)	1.5 (1.1-2.4)	1.4 (1-2.3)	1.4 (0.9-2.3)	1.5 (0.9-2.3)	1.5 (1.0-2.1)
EMD	3.2 (2.1-4.7)	3.0 (1.9-4.5)	2.7 (1.9-4.6)	2.8 (1.9-4.6)	2.9 (1.9-4.5)	2.8 (1.9-4.1)
PRSA _Spt	**33 (19-46)**	**25 (17-37)**	**25 (14-38)**	**26 (14 -41)**	**25 (14 -36)**	**23 (14 -36)**

Since the discussed features are not normally distributed, we employed the Mann-Whitney U-test to evaluate differences between the two high-risk groups and controls and computed the Cohen's r size effect to quantify them **[5.59]**. Results are considered significant when $p<0.0167$, following Bonferroni correction for multiple comparisons. The results related to the comparison between GDMs and Controls are reported in **Figure 5.10**, while comparisons between IUGRs and Controls in **Figure 5.11**. Confidence intervals were computed using empirical bootstrap with 1000 repetitions.

[redacted]

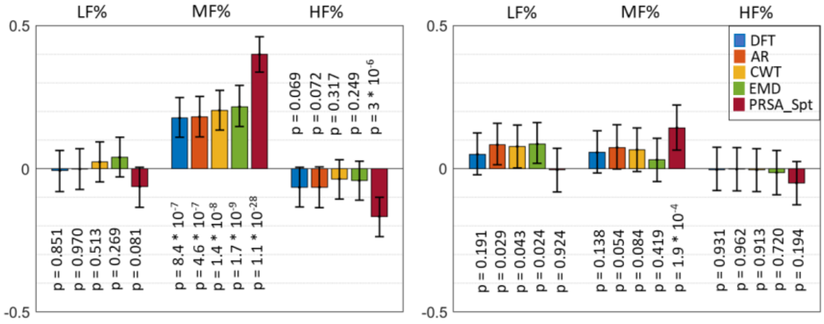


Figure 5.10: Cohen's effects size for the Mann-Whitney U-test in comparison between Controls and GDMs. Positive values indicate an increase in the high-risk population. Confidence intervals are computed using empirical bootstrap with 1000 repetitions.

[redacted]

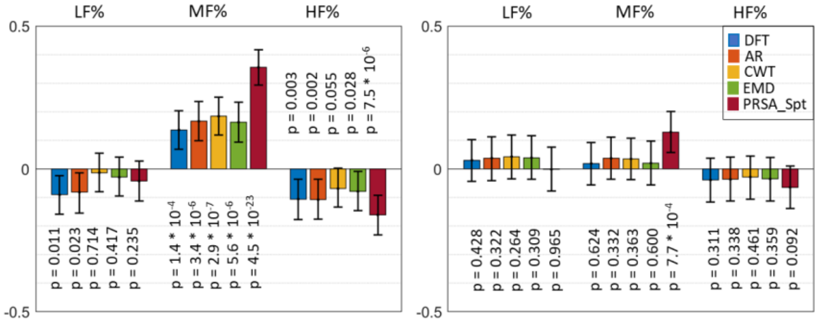


Figure 5.11: Cohen's effects size for the Mann-Whitney U-test in comparison between Controls and IUGR.

In general, more differences were identified in the first group of weeks. This result is in agreement with other studies **[5.31] [5.60]**, which also found more differences in the pre-term period compared to the early term, even though the FHR features analyzed were different.

In the first groups of weeks, all methods identified a significant increase in MF% both in the comparison between GDMs against Controls and IUGRs against Controls. It can be noticed that the PRSA_Spt method resulted in a significantly bigger size effect. Indeed, traditional methods resulted in both cases in small size effects (lower than 0.2), while the ones obtained with PRSA_Spt were moderate (0.40 for GDMs and 0.36 for IUGRs).

Classical methods did not detect any significant difference in the HF band for GDMs, while a small reduction was identified by AR and DFT in the IUGR population. The PRSA_Spt instead evidenced significantly lower values in the HF band for both pathological groups compared to controls. The effect sizes however are small (-0.17 for GDMs and -0.16 for IUGRs), even though the p-values are well below the 5% significance level ($3x10^{-6}$ and $7.5x10^{-6}$, respectively).

No differences were identified in the LF band for GDMs, while a small reduction was identified by AR in the GDMs. The PRSA_Spt did not evidence a significant difference in this band.

In the second group of weeks some traditional methods, but not PRSA_Spt, suggest a slight increase in LF% for GDMs. An increase in MF% in both high-risk groups was identified only by PRSA_Spt. No differences were identified in HF%.

The differences between groups identified with the PRSA_Spt method can also be clearly seen by looking at the spectra, which are reported in **Figure 5.12**. It can be noticed that at weeks 32 to 36 high-risk pregnancies and controls show different behaviors in MF and HF, while they substantially overlap in the LF band. Moreover, it is clearly visible a reversal of the trends of physiological and high-risk pregnancies around 0.55Hz. The differences among the average PRSA_Spt for the three analyzed groups at weeks 37 and 38 are much less evident.

[redacted]

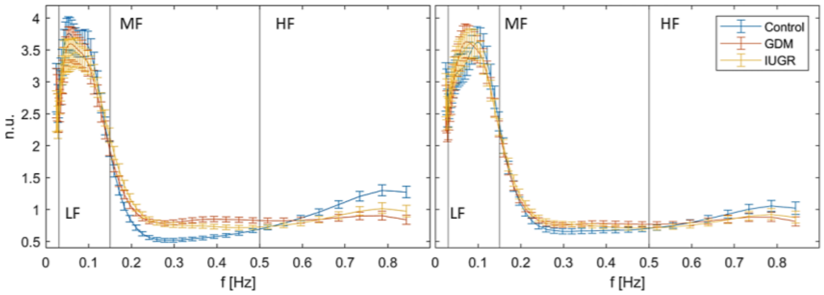


Figure 5.12: Average normalized deceleration related PRSA_Spt with 95% confidence intervals computed using empirical bootstrap with 1000 repetitions.

Notice that high-risk pregnancies show similar behaviors (increase in MF and decrease in HF) and more relevant differences are observed in the first group of weeks. Spectra have been normalized by dividing by the total power of the PRSA_Spt and are presented in normalized units.

As reported in **Table 5.6** the FHR variance, or total power, shows a trend which aligns with expectations, i.e., higher variability in Controls. However, the differences are not statistically significant.

Table 5.6: Medians and quartiles of the total power of FHR.

32-36

	Control	GDM	IUGR
Tot. Pow. [ms^2]	259 (158 - 404)	220 (143 – 391)	243 (137 - 391)

37-38

	Control	GDM	IUGR
Tot. Pow. [ms^2]	297 (178 - 452)	265 (166 - 418)	266 (172 - 451)

5.2.5. Discussions and conclusions.

The purpose of the work presented in this chapter is two-fold. Firstly, we elaborated on the applicability and robustness of spectral analysis in FHR signals using a very large dataset of healthy and complicated pregnancies and compared several existing methods of spectral analysis. Secondly, we proposed a different approach for evaluating the spectrum of the PRSA and compared the results with those obtained with more classical spectral analysis methods.

We found that the application of the PRSA method before computing the spectra improves substantially the capability to distinguish between uncomplicated and high-risk pregnancies compared to classical spectral analysis while maintaining the physiological interpretability of the results. On the other hand, we did not observe substantial differences among the classical methods of spectral analysis.

Time-frequency analysis has the advantage that allows localizing in time the results, allowing for example to investigate how the fetus reacts to stimuli (e.g., uterine contractions **[5.61]**) and differentiating among different fetal behavioral states **[5.62]**. However, the reported results suggest that when an aggregate measure is pursued, the PRSA_Spt presents a clear advantage over the other methods, which we attribute to its ability to reduce noise, capture oscillations that are not phase synchronized, and its advantageous scaling behavior.

The most relevant difference observed between controls and high-risk groups is a relative increase in the power in MF for the latter. Previous studies have attributed oscillations in this band to the presence of fetal movements **[5.30]**. However, we did not observe a significant increase in their occurrence as perceived by the mother in the high-risk groups. This may be indicative that other pathophysiological mechanisms are at the source of this pattern. Indeed, this frequency band is arguably the one with the least clear physiological interpretation, which requires further study. Quite interestingly, in **[5.63]** the authors found very little power in this band in uncomplicated low-risk pregnancies.

The reduction in HF that was observed is consistent with a reduction of respiratory movements in the pathological groups. Concerning IUGR foetuses, this is backed up by other studies which found that respiratory movements are lower in speed, power and intensity and in general have lower "quality" **[5.64]**.

We did not observe significant differences between the groups in LF% with PRSA_Spt, while minimal differences were identified using other methods. This is probably due to the scaling behavior of the PRSA and the choice of T=1. Indeed, quasi-oscillations that are enhanced the most lay around 1⁄2.5T **[5.44]** which in our case corresponds to the HF band. We did identify a significant reduction in LF% at weeks 32-36, but not at weeks 37-38 using T=2 (size effect:-0.19 for GDM and -0.15 for IUGRs). This however came at the expense of substantially reduced differences in MF% and HF%. At T=4 the relative power in HF was virtually null.

Interestingly, we did not find substantial differences when considering the acceleration-related PRSA curve instead of the deceleration-related one.

One limitation of the present study is that we only analysed data collected using CTG with a sampling frequency of 2Hz. We acknowledge that the results may change using FHR acquired with different techniques, such as fetal electrocardiography. Another limitation is that we did not differentiate between behavioral states, which is something we aim to do as future development.

The analysis in the frequency domain of the FHR provides useful insights into the fetal physiology, since it allows to assess non-invasively the functioning of the ANS. In this work, we compared four relatively traditional methods to perform spectral analysis and a novel approach based on the CWT of the PRSA curve and conclude that the latter allows to identify more clearly the differences in the frequency content of the FHR induced by GDM and IUGR. We believe that this approach may have relevant applications, for example for improving multi-parameters classification.

5.2.6. Reference for chapter 5.

[5.1] J. G. Nijhuis, H. F. R. Prechtl, C. B. Martin, and R. S. G. M.Bots, “Are there behavioural states in the human fetus? *”, 1982.

[5.2] FIGO Intrapartum Fetal Monitoring Expert Consensus Panel, “FIGO consensus guidelines on intrapartum fetal monitoring”, International Journal of Gynecology & Obstetrics, vol. 131, no. 1, 2015, doi: https://doi.org/10.1016/j.ijgo.2015.06.020.

[5.3] C. Romanini and G. Rizzo, “Fetal behaviour in normal and compromised fetuses. An overview”, Early Hum Dev. 1995. doi: 10.1016/0378-3782(95)01667-8.

[5.4] M. A. T. van Vliet, C. B. Jr Martin, J. G. Nijhuis and H. F. Prechtl, “The relationship between fetal activity and behavioral states and fetal breathing movements in normal and growth-retarded fetuses”, Am J Obstet Gynecol. 1985 Nov 1;153(5):582-8. doi:10.1016/0002-9378(85)90483-1.

[5.5] H. Valensise, G. Ciotti, D. Giobbi, A. L. Tranquilli, D. Arduini, and C. Romanini, “Behaviour of fetuses from hypertensive mothers”, Hypertens Pregnancy, vol. B7, no. 1–2, pp. 227–240, 1988, doi: 10.3109/10641958809023519.

[5.6] K. Sumiyoshi, Y. Kawagoe, M. Ohhashi, S. Furukawa, H. Sameshima, and T. Ikenoue, “Delayed rhythm formation of normal structured, growth-restricted fetuses using fetal heart rate monitoring patterns", Journal of Obstetrics and Gynaecology Research, vol. 46, no. 8, pp. 1342–1348, Aug. 2020, doi: 10.1111/jog.14316.

[5.7] A. Kurjak, S. Panchal, and S. Porovic, "Fetal behavior in normal pregnancy and diabetic pregnancy", Donald School Journal of Ultrasound in Obstetrics and Gynecology, vol. 12, no. 2. Jaypee Brothers Medical Publishers (P) Ltd, pp. 124–136, Apr. 01, 2018.
doi: 10.5005/jp-journals-10009-1562.

[5.8] H. Gonçalves, J. Bernardes, A. Paula Rocha, and D. Ayres-de Campos, "Linear and nonlinear analysis of heart rate patterns associated with fetal behavioral states in the antepartum period", Early Hum Dev, vol. 83, no. 9, pp. 585–591, Sep. 2007, doi: 10.1016/j.earlhumdev.2006.12.006.

[5.9] L. Semeia, K. Sippel, J. Moser, and H. Preissl, "Evaluation of parameters for fetal behavioural state classification", Sci Rep, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-07476-x.

[5.10] J. Brändle, H. Preissl, R. Draganova, E. Ortiz, K. O. Kagan, H. Abele, S. Y. Brucker, and I. Kiefer-Schmidt "Heart rate variability parameters and fetal movement complement fetal behavioral states detection via magnetography to monitor neurovegetative development", Front Hum Neurosci, vol. 9, no. APRIL, Apr. 2015, doi: 10.3389/fnhum.2015.00147.

[5.11] S. Lange, P. Van Leeuwen, U. Schneider, B. Frank, D. Hoyer,
D. Geue, D. Grönemeyer, "Heart rate features in fetal behavioural states", Early Hum Dev, vol. 85, no. 2, pp. 131–135, Feb. 2009, doi: 10.1016/j.earlhumdev.2008.07.004.

[5.12] I .Kiefer-Schmidt, J. Raufer, J. Brändle, J.Münßinger, H.Abele, D.Wallwiener, H.Eswaran and H.Preissl, "Is there a relationship between fetal brain function and the fetal behavioral state? A fetal MEG-study", J Perinat Med. 2013 Sep 1;41(5):605-12. doi: 10.1515/jpm-2013-0022.

[5.13] D. Arduini, G. Rizzo, A. Vizzone, H. Valensise, and C. Romaninit, "The Fetal Behavioural States : an ultrasonic study", Prenat Diagn, vol. 5, pp. 269–276, 1985.

[5.14] I. E. Timor-Tritsch, L. R. J. Dierker, R. H. Hertz, N. C. Deagan, and M. G. Rosen, “Studies of antepartum behavioral state in the human fetus at term”, Am J Obstet Gynecol, vol. 132, no. 5, pp 524–528, Nov. 1978, doi: 10.1016/0002-9378(78)90747-0.

[5.15] S. Vairavan, U. D. Ulusar, H. Eswaran, H. Preissl, J. D. Wilson, S. S Mckelvey, C. L. Lowery, R. B. Govindan, “A computer aided approach to detect the fetal behavioral states using multi-sensor Magnetocardiographic recordings”, Comput Biol Med, vol. 69, pp. 44–51, Feb. 2016, doi: 10.1016/j.compbiomed.2015.11.017.

[5.16] A. Samjeed, M. Wahbah, A. H. Khandoker, and L. Hadjileontiadis, “Unsupervised Fetal Behavioral State Classification Using Non-Invasive Electrocardiographic Recordings”, doi: 10.22489/CinC.2021.240.

[5.17] N. Pini, M. Lucchini, W. P. Fifer, and R. Barbieri, “Point Process Framework for the Characterization of Fetal Sleep States”, Annu Int Conf IEEE Eng Med Biol Soc. 2020. doi: 10.1109/EMBC44109.2020.9176169.

[5.18] L. R. Rabiner and B. H. Juang, “An Introduction to Hidden Markov Models”, IEEE ASSP Magazine, vol. 3, no. 1, pp. 4–16, 1986, doi: 10.1109/MASSP.1986.1165342.

[5.19] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, “A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains”, Source: The Annals of Mathematical Statistics, vol. 41, no. 1, pp. 164–171, 1970.

[5.20] A. J. Viterbi, “Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm”, IEEE Trans Inf Theory, vol. 13, no. 2, pp. 260–269, 1967, doi: 10.1109/TIT.1967.1054010.

[5.21] D. Arduini, G. Rizzo, and C. Romanini, “Computerized Analysis of Fetal Heart Rate”, J Perinat Med, vol. 22, no. Suppl.1, pp. 22–27, 1994, doi: 10.1515/jpme.1994.22.s1.22.

[5.22] M. G. Signorini, G. Magenes, S. Cerutti, and D. Arduini, “Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings”, IEEE Trans Biomed Eng, vol. 50, no. 3, pp. 365–374, 2003, doi: 10.1109/TBME.2003.808824.

[5.23] R. Mantel', H. P. van Gelin, F. J. M. Caron, J. M. Swartjes, E. E. van Woerden, and H. W. Jongsmab, "Computer Analysis of antepartum fetal heart rate: 2. Detection of Accelerations and Decelerations", 1990.

[5.24] A. C. de Wit and J. G. Nijhuis, "Validity of the Hewlett Packard actograph in detecting fetal movements", Ultrasound in Obstetrics and Gynecology, vol. 22, no. 2, pp. 152–156, Aug. 2003, doi: 10.1002/uog.155.

[5.25] I. J. M. Nijhuis, J. ten Hof, J. G. Nijhuis, E. J. Mulder, H.Narayan, D. J. Taylor and G. H. Visser, "Temporal organization of fetal behavior from 24-weeks gestation onwards in normal and complicated pregnancies", Dev Psychobiol, vol. 34, no. 4, pp. 257 268, 1999, doi: 10.1002/(SICI)1098-2302(199905)34:2<257::AID DEV2>3.0.CO;2-V.

[5.26] E. J. H. Mulder, G. H. A. Vissera, D. J. Bekedam, and H. F. R. Prechtl, "Emergence of behavioural states in fetuses of type-l diabetic women", Early Hum Dev, vol. 15, p. 231, 1987.

[5.27] R. M. Grivell, Z. Alfirevic, G. M. L. Gyte, and D. Devane, "Antenatal cardiotocography for fetal assessment," Cochrane Database of Systematic Reviews, vol. 2015, no. 9, pp. 1–39, 2015, doi: 10.1002/14651858.CD007863.pub4.

[5.28] A. Galli, E. Peri, C. Rabotti, S. Ouzounov, and M. Mischi, "Automatic optimization of multichannel electrode configurations for robust fetal heart rate detection by Blind Source Separation," IEEE Trans Biomed Eng, pp. 1–12, Oct. 2022, doi: 10.1109/tbme.2022.3212587.

[5.29] J. Pardey, M. Moulden, and C. W. G. Redman, "A computer system for the numerical analysis of nonstress tests," Am J Obstet Gynecol, vol. 186, no. 5, pp. 1095–1103, 2002, doi: 10.1067/mob.2002.122447.

[5.30] M. G. Signorini, G. Magenes, S. Cerutti, and D. Arduini, "Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings," IEEE Trans Biomed Eng, vol. 50, no. 3, pp. 365–374, 2003, doi: 10.1109/TBME.2003.808824.

[5.31] 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 Obstet Gynecol Scand, vol. 96, no. 11, pp. 1322–1329, Nov. 2017, doi: 10.1111/aogs.13228.

[5.32] N. Pini et al., “A Machine Learning Approach to Monitor the Emergence of Late Intrauterine Growth Restriction,” Front Artif Intell, vol. 4, Mar. 2021, doi: 10.3389/frai.2021.622616.

[5.33] A. Georgieva, P. Abry, I. Nunes, and M. G.Frasch, “Editorial: Fetal-maternal monitoring in the age of artificial intelligence and computer-aided decision support: A multidisciplinary perspective,” Front Pediatr, pp. 01–05, 2022, doi: 10.3389/fped.2022.1007799.

[5.34] V. J. King, L. Bennet, P. R. Stone, A. Clark, A. J. Gunn, and S. K. Dhillon, “Fetal growth restriction and stillbirth: Biomarkers for identifying at risk fetuses,” Front Physiol, vol. 13, Aug. 2022, doi: 10.3389/fphys.2022.959750.

[5.35] C. A. Lear et al., “Understanding Fetal Heart Rate Patterns That May Predict Antenatal and Intrapartum Neural Injury,” Semin Pediatr Neurol, vol. 28, pp. 3–16, Dec. 2018, doi: 10.1016/j.spen.2018.05.002.

[5.36] D. Hoyer et al., “Monitoring fetal maturation - Objectives, techniques and indices of autonomic function,” Physiological Measurement, vol. 38, no. 5. IOP Publishing Ltd, pp. R61–R88, Apr. 18, 2017. doi: 10.1088/1361-6579/aa5fca.

[5.37] M. W. Gillman, “Developmental Origins of Health and Disease,” N Engl J Med, vol. 27, pp. 1848–1850, 2005.

[5.38] N. Montano et al., “Heart rate variability explored in the frequency domain: A tool to investigate the link between heart and behavior,” Neurosci Biobehav Rev, vol. 33, no. 2, pp. 71–80, Feb. 2009, doi: 10.1016/j.neubiorev.2008.07.006.

[5.39] F. Shaffer and J. P. Ginsberg, “An Overview of Heart Rate Variability Metrics and Norms,” Front Public Health, vol. 5, no. September, pp. 1–17, 2017, doi: 10.3389/fpubh.2017.00258.

[5.40] M. Romano, L. Iuppariello, A. M. Ponsiglione, G. Improta, P. Bifulco, and M. Cesarelli, “Frequency and Time Domain Analysis of Foetal Heart Rate Variability with Traditional Indexes: A Critical Survey,” Comput Math Methods Med, vol. 2016, pp. 16–19, 2016, doi: 10.1155/2016/9585431.

[5.41] J. G. Nijhuis ’, H. F. R. Prechtl, C. B. ’ Martin, and R. S. G. M. Bots, “Are there behavioural states in the human fetus? *,” 1982.

[5.42] J. O. E. H. Van Laar, M. M. Porath, C. H. L. Peters, and S. G. Oei, “Spectral analysis of fetal heart rate variability for fetal surveillance: Review of the literature,” Acta Obstet Gynecol Scand, vol. 87, no. 3, pp. 300–306, 2008, doi: 10.1080/00016340801898950.

[5.43] A. M. Ponsiglione et al., “A Comprehensive Review of Techniques for Processing and Analyzing Fetal Heart Rate Signals,” 2021, doi: 10.3390/s21186136.

[5.44] A. Bauer et al., “Phase-rectified signal averaging detects quasi-periodicities in non-stationary data,” Physica A: Statistical Mechanics and its Applications, vol. 364, pp. 423–434, 2006, doi: 10.1016/j.physa.2005.08.080.

[5.45] S. M. Lobmaier et al., “Phase-rectified signal averaging as a new method for surveillance of growth restricted fetuses,” Journal of Maternal-Fetal and Neonatal Medicine, vol. 25, no. 12, pp. 2523–2528, Dec. 2012, doi: 10.3109/14767058.2012.696163.

[5.46] S. M. Lobmaier et al., “Influence of gestational diabetes on fetal autonomic nervous system: a study using phase-rectified signal-averaging analysis,” Ultrasound in Obstetrics and Gynecology, vol. 52, no. 3, pp. 347–351, 2018, doi: 10.1002/uog.18823.

[5.47] E. M. Graatsma et al., “Average acceleration and deceleration capacity of fetal heart rate in normal pregnancy and in pregnancies complicated by fetal growth restriction,” Journal of Maternal-Fetal and Neonatal Medicine, vol. 25, no. 12, pp. 2517–2522, Dec. 2012, doi: 10.3109/14767058.2012.704446.

[5.48] 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 J Biomed Health Inform, vol. 17, no. 5, pp. 959–966, 2013, doi: 10.1109/JBHI.2013.2268423.

[5.49] M. W. Rivolta, T. Stampalija, M. G. Frasch, and R. Sassi, "Theoretical Value of Deceleration Capacity Points to Deceleration Reserve of Fetal Heart Rate," IEEE Trans Biomed Eng, vol. 67, no. 4, pp. 1176–1185, 2020, doi: 10.1109/TBME.2019.2932808.

[5.50] A. Bauer et al., "Deceleration capacity of heart rate as a predictor of mortality after myocardial infarction: cohort study," Lancet, vol. 367, no. 9523, pp. 1674–1681, 2006, doi: 10.1016/S0140-6736(06)68735-7.

[5.51] Q. Pan et al., "Do the deceleration/acceleration capacities of heart rate reflect cardiac sympathetic or vagal activity? A model study," Med Biol Eng Comput, vol. 54, no. 12, pp. 1921–1933, Dec. 2016, doi: 10.1007/s11517-016-1486-9.

[5.52] S. J. Gordijn et al., "Consensus definition of fetal growth restriction: a Delphi procedure," Ultrasound Obstet Gynecol, vol. 48, no. 3, pp. 333–339, Sep. 2016, doi: 10.1002/uog.15884.

[5.53] A. Caughey and M. Turrentine, "ACOG Practice Bulletin: Gestational Diabetes Mellitus," Obstetrics and gynecology, vol. 130, no. 1, pp. e17–e31, 2017, doi: https://doi.org/10.1097/AOG.0000000000002501.

[5.54] N. E. Huang et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 454, no. 1971, pp. 903–995, 1998, doi: 10.1098/rspa.1998.0193.

[5.55] C. Torrence and G. P. Compo, "A Practical Guide to Wavelet Analysis," Bull Am Meteorol Soc, vol. 79, no. 1, pp. 61–78, 1998, doi: https://doi.org/10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2.

[5.56] M. A. Colominas, G. Schlotthauer, and M. E. Torres, "Improved complete ensemble EMD: A suitable tool for biomedical signal processing," Biomed Signal Process

Control, vol. 14, no. 1, pp. 19–29, 2014, doi: 10.1016/j.bspc.2014.06.009.

[5.57] J. W. Kantelhardt et al., “Phase-rectified signal averaging for the detection of quasi-periodicities and the prediction of cardiovascular risk,” Chaos, vol. 17, no. 1, 2007, doi: 10.1063/1.2430636.

[5.58] R. Sassi, T. Stampalija, D. Casati, E. Ferrazzi, A. Bauer, and M. W. Rivolta, “A methodological assessment of phase-rectified signal averaging through simulated beat-to-beat interval time series,” Comput Cardiol (2010), vol. 41, no. January, pp. 601–604, 2014.

[5.59] C. O. Fritz, P. E. Morris, and J. J. Richler, “Effect size estimates: Current use, calculations, and interpretation,” J Exp Psychol Gen, vol. 141, no. 1, pp. 2–18, 2012, doi: 10.1037/a0024338.

[5.60] M. Ferrario, S. Maria G, and G. Magenes, “Complexity analysis of the fetal heart rate variability: early identification of severe intrauterine growth-restricted fetuses,” Med Biol Eng Comput, vol. 47, pp. 911–919, 2009, doi: 10.1007/s11517-009-0502-8.

[5.61] M. Romano, P. Bifulco, M. Cesarelli, M. Sansone, and M. Bracale, “Foetal heart rate power spectrum response to uterine contraction,” Med Biol Eng Comput, vol. 44, no. 3, pp. 188–201, 2006, doi: 10.1007/s11517-006-0022-8.

[5.62] M. David, M. Hirsch, J. Karin, E. Toledo, S. Akselrod, and A. S. An, “An estimate of fetal autonomic state by time-frequency analysis of fetal heart rate variability,” J Appl Physiol, vol. 102, pp. 1057–1064, 2007, doi: 10.1152/japplphysiol.00114.2006.-In.

[5.63] M. David, M. Hirsch, J. Karin, E. Toledo, S. Akselrod, and A. S. An, “An estimate of fetal autonomic state by time-frequency analysis of fetal heart rate variability,” J Appl Physiol, vol. 102, pp. 1057–1064, 2007, doi: 10.1152/japplphysiol.00114.2006.-In.

[5.64] F. Bos and C. Einspieler, “Intrauterine growth retardation, general movements, and neurodevelopmental outcome: a review,” Dev Med Child Neurol, vol. 43, p. 61, 2001, doi: https://doi.org/10.1111/j.1469-8749.2001.tb00388.x.

Chapter 6

Deep learning methods for the classification of FHR tracings.

As anticipated at the end of **Chapter 3**, The incorporation of Artificial Intelligence (AI) into the analysis of CTG data has the potential to enhance the discriminative capabilities of this approach. AI models inherently possess the capacity to uncover hidden patterns and trends within the data that might escape detection through visual analysis alone, thereby offering valuable insights for the management of high-risk pregnancies and the informed decision-making process.

In recent years, major attentions are turning towards Deep Learning, which at time, seems to be the most promising approach. However, the performances of DL algorithms are strongly dependent on the availability of large amounts of labeled training data. Inadequate data volume or imprecise labeling can impede real-world applications, potentially leading to an overestimation of a model's ability to generalize and an elevated risk of misclassification. Additionally, these techniques require well-structured data to achieve optimal performance. As stressed through

Chapters 3 and 4, at present, there is a shortage of large CTG datasets purposefully tailored for deep learning applications, making them less readily available. In **Chapter 4** we illustrated how we managed to address the aforementioned constraints, building up NAPAMI, a novel large structured CTG dataset, specifically designed for the application of all that AI techniques that rely on substantial data volumes to attain satisfactory outcomes. In the present chapter we will detail how we exploited this new abundance of CTG data to develop three innovative DL solutions for the classification of FHR tracings. In particular, we illustrate a hybrid net which is able to deal with heterogeneous data (i.e., quantitative information describing the FHR tracings and a set of images coding the tracings themselves) to increase the performances of classical architectures in classifying CTG tracings.

6.1. Background: Artificial Intelligence in CTG analysis and the shift towards Deep Learning.

Several AI approaches have been proposed in the literature since the introduction of the computerized CTG analysis, which allows to quantify the FHR behavior by means of both linear and non-linear aspects. These aspects consider the indices used in traditional diagnostics, novel and advanced regressors coming from quantitative frequency analysis, nonlinear parameters, and are integrated with maternal information. The different sets of features are used for the classification of the occurrence of pathological states or simply for the assessment of the maintenance of the healthy condition. The results strongly depend on the number of cases, the used database, the considered features and the performance of the classifiers

Fergus et al. **[6.1]** utilized Machine Learning models to classify caesarean section and normal vaginal deliveries based on cardiotocographic traces. In this study 552 FHR signal recordings, of which 506 controls and 46 pathological, were used as dataset, from which features like baseline, accelerations, decelerations, Short-Term Variability (STV) and many others have been extracted. The models adopted in this work are multi-Layer feedforward neural network, Fisher's Linear Discriminant Analysis (FLDA) and Random Forests (RF).

These methods, based on predictive learning classifiers, are known to suffer from the limitation of relying on the extraction of complex hand-crafted features from the signals. Therefore, research in this field has been moving in the direction of deep learning techniques. Petrozziello et al. **[6.2]** make use of raw signals from Electronic Fetal Monitoring (EFM) to predict fetal distress. They fed a Long Short-Term Memory (LSTM) and a CNN network with both FHR and UC signals, reaching a predictive accuracy of respectively 61% and 68%. It is worthnoting that their dataset consisted of 35,429 recordings, but contained 33,959 healthy newborns, while only 1,470 compromised, resulting to be strongly unbalanced.

Iraji et al. **[6.3]** explored other soft computing techniques to predict fetal state using cardiotocographic recordings. Neuro-fuzzy inference system (MLA-ANFIS), Neural Networks and deep stacked sparse auto-encoders (DSSAEs) were implemented. Iraji used a limited dataset composed of 2,126 selected recordings that were divided in three classes: 1655 normal, 295 suspect, and 176 pathologic. On the full dataset, the best performing approach was deep learning with an accuracy of 96.7%, followed by ANFIS that reaches an accuracy of 95.3%.

Zhao et al. **[6.4]** used FHR signals transformed into images by using Continuous Wavelet Transform. Their models consist of an 8-layer Convolutional Neural Network (CNN) with a single Convolutional Layer. Their dataset was the open-access database (CTU-UHB), with 552 intrapartum

FHR recordings, containing a noticeable percentage (about 20%) of scalp electrode recordings. Their model reaches a 98.34% of accuracy with an AUC of 97.82%.

More recently, Rahmayanti et al. **[6.5]** propose a comparison between ML methods for the classification of fetal well-being using 21 attributes from the measurement of FHR and UC. They report excellent levels of accuracy. The dataset used was obtained from the University of California Irvine Machine Learning Repository, which is a public dataset. It consisted of 2,126 data on pregnant women who are in the third trimester of their pregnancy collected through the system Sys Porto. In their study, the application of deep learning methods did not produce satisfactory and improved results compared to the ML approach The authors consider that using a more representative dataset and perfecting the set of variables can improve performance.

The contribution of Su Liu et al. **[6.6]** stresses the importance of integrating echo images with CTG data for improving the classification of fetal states. The goal of the study was to improve the feasibility and economic benefits of an artificial intelligence based medical system when Doppler ultrasound (DUS) imaging technology are combined with fetal heart detection to predict the fetal distress in pregnancy-induced hypertension (PIH).

Finally, the review by Ki Hoon Ahn, et al. **[6.7]** presents a comprehensive overview of the possible application of AI, DL and ML in Obstetrics for the early diagnosis of various maternal-fetal conditions such as preterm birth and abnormal fetal growth. The work purpose was to review recent advances on the application of artificial intelligence in this medical field. The work summarizes in table form the main characteristics of the different studies using AI, ML, DL methods. From this study we understand the pervasiveness in the field of fetal medicine of AI methodologies. There is also a perception of the complexity of the work still to be done to build reliable and validated classification systems. The data illustrated show the great variety in terms of applications and the number of data collected/used that define the final performance of the analysis tools.

As a general remark, it is possible to notice that the global accuracy and performance of AI methods for perinatal medicine so far published in the literature are almost inversely proportional to the number of cases: the best results are obtained with limited and selected datasets.

In fact, the main limitation imposed by deep learning techniques is the huge number of data needed to train the neural architectures. Hence, the use of an inadequate number of records could lead to an overestimation of the generalization capabilities of the model.

To overcome this intrinsic limitation, as detailed in **Chapter** 4, we succeeded in structuring NAPAMI, a vast CTG dataset, whose dimension and class balance opens the way to all that AI solutions (e.g., DL ones) which require a huge number of records to achieve acceptable results.

Leveraging this new abundance of labelled CTG data, as will be illustrated throughout the chapter, we developed and compared three different DL solutions to classify CTG tracings, such to disambiguate a reassuring trace from a suspicious one, which could likely reflect a condition of fetal distress.

The first solution includes a Multi-Layer Perceptron (MLP) net, which is fed with a set of quantitative indices, selected from the wider group described in **Chapter 4**.

The second one consists of a Convolutional Neural Network (CNN) which receives as input a set of images, obtained from the FHR tracings through ad hoc transformation techniques, as will be detailed in **Section 6.2.2**.

Finally, we propose a new neural model with two branches respectively consisting of the two nets introduced above. This MLP+CNN neural architecture receives heterogeneous input data, i.e., a set of parameters and images. The aim is to exploit the neural network's generalization capacity by integrating FHR quantitative regressors, known to summarize the pathophysiological condition of the fetus, (either in time, frequency, and non-linear domains) with some new features implicitly learned from images, consisting in various representations of the raw FHR signal (time-frequency, recurrent patterns, etc.).

6.2. Preparing Training and Test sets.

The dataset used to train and test the classification performance of the proposed neural architectures derives from some modifications to NAPAMI (**Chapter 4**).

First records relating to FHR tracings with an excessive level of corruption by noise and signal losses were excluded. Then, to avoid any potential polarization induced by the difference in the length of the FHR sequences, we decided to consider sequences of exactly 20 minutes each, corresponding to 2,400 points in time (Fs = 2Hz). Records including tracings lasting less than 20 minutes were directly discarded, while traces with a duration lasting longer than 20 minutes were split such to consider the last 2,400 points.

After these steps, the dataset contains 17,483 entries, with 7,733 healthy and 9,750 pathological tracings.

Pathological group included tracings of subjects with different diseases both of maternal and fetal origin, such as diabetes, malformations, intrauterine growth restriction (IUGR).

As the goal of the study aimed at the separation between healthy and pathological fetuses, each entry of the dataset was binary categorized. Recordings belonging to the physiological pregnancy group (Normal), i.e., those with codes 01 and 02 in **Table 4.1** of **Chapter 4**, were denoted with 0 and those presenting a disease condition (Pathologic), i.e., codes from 03 to 11, were denoted with 1.

Furthermore, to obtain a perfect balance between Normal and Pathologic records, a subset of 7,000 recordings was selected in both the groups, on the basis of the FHR signals quality.

Thus, the final dataset contains 14,000 of fetal records, equally balanced by category.

Having a large set of balanced data is fundamental to avoid polarized and inconsistent results which are both weaknesses affecting DL method applications.

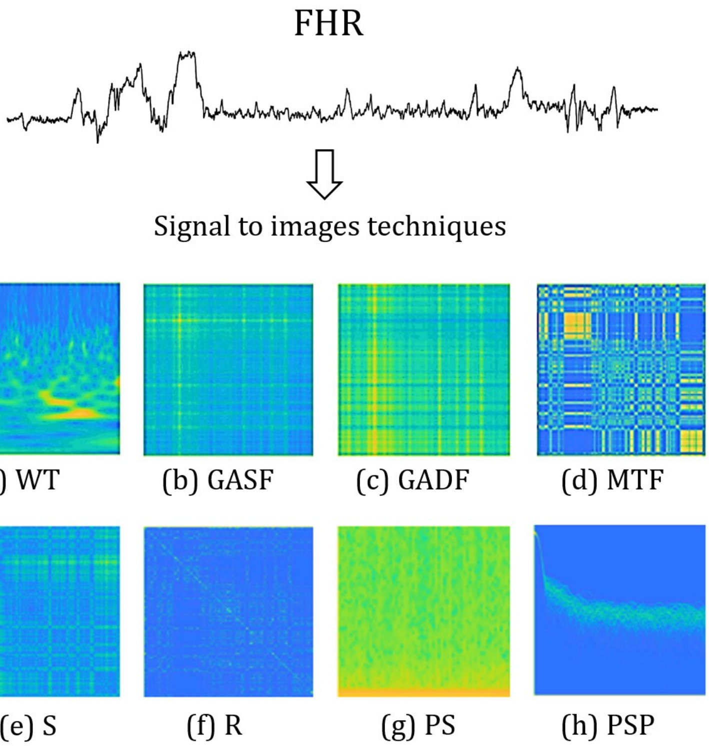
6.2.1. Parameters selection.

With the intent of feeding the MLP branch with a set of quantitative regressors describing the statistical characteristics of recorded signals, we considered a set of parameters for each of the 20 minutes split and processed FHR signals in our dataset.
Among the features calculated from the FHR signal, the subgroup that constitutes the selection of the ones included in the analysis, was made based on the literature study. This was followed by a process of feature selection and correlation analysis, starting from a wide group of more than 30 regressors, as introduced in **Chapter 4**, commonly used in fetal monitoring, and known to provide pathophysiological meaning related with the control mechanisms of heart. In particular, we decided to include in our study all the parameters evaluated in **[6.8]**, which provided good results in the classification of Normal and IUGR fetuses, although with a small dataset.
The final set consists of 15 quantitative parameters. The parameter set includes 4 linear parameters describing signals in time domain, namely: DELTA, Interval Index (II), Short-Term Variability (STV), Long-Term Irregularity (LTI) computed as described in **[6.9]**, 3 linear parameters related to frequency domain signal content i.e., Low Frequency (LF), Movement Frequency (MF), High Frequency (HF) and the complex, non-linear parameter Approximate Entropy (ApEn) **[6.10]**. These parameters were automatically extracted by the 2CTG2 software. Moreover, we also included the FHRB that is the mean value of baseline, extracted with a modified version of Mantel’s Algorithm **[6.11]**, the ratio in the power spectrum bands (LF/(MF+HF)), the number of small accelerations (>10 bpm and <15 bpm for 15 s), the number of large accelerations (>15 bpm for 15 s), the number of decelerations (>20 bpm for 30 s or > 10 bpm for 60 s) **[6.9]**. Two more indices were considered as input values for the MLP branch, i.e., the gestational week and mother’s age.

This parameter set covers most of the information the FHR signal contains as it considers time domain changes, frequency domain linear components and complexity signal characteristics associated to nonlinear dynamic evolution.
Since the considered parameters have different scales, before providing them as inputs to the MLP, we applied a normalization procedure by scaling all parameters in the range 0-1 using the min-max normalization.

6.2.2. Signal to images techniques.

Among the goals of our approach was to exploit the implicit ability of neural networks to learn complex features directly from the available data, without summarizing them by means of any statistical regressor. Thus, we developed a convolutional neural network (CNN), because CNNs have already shown great abilities in extracting important features from images and in image classification tasks **[6.12]**.
To realize this approach, we needed to shift from a 1-D representation of the FHR tracings to a bi-dimensional one, through the use of ad-hoc techniques. In other words, we had to encode the denoised FHR signals, obtained after preprocessing, into a set of images representing the FHR behavior by means of various computational transformations. Signal to image transformations are becoming more and more common since the recent successes got by DL in the field of computer vision.
In particular, for our purpose we decided to use eight transformation techniques, whose results are represented in **Figure 6.1**, with the same *parula* colormap, and briefly described in the following subsections, to allow the system to automatically grasp different aspects about the nature of the FHR signal from the different images provided.



(a) WT (b) GASF (c) GADF (d) MTF

(e) S (f) R (g) PS (h) PSP

Figure 6.1: The present figure illustrates the conversion of an exemplifying FHR tracing into 8 images, obtained with 8 different transformation techniques. The latter comprehend: Wavelet transform (WT), Gramian Angular Summation Field (GASF), Gramian Angular Difference Field (GADF), Markov Transition Field (MTF), Recurrence Plot (R), Distance Matrix for Recurrence Plot (S), Power Spectrum (PS), Persistence Spectrum (PSP).

The choice of the particular set of techniques employed to encode the FHR signals into images has followed an in-depth literature search. Our intent was to exploit the intrinsic capacity of CNNs to automatically select the most relevant features, starting from the images provided as input. For that aim we selected a set of transformation techniques to obtain a group of images that could allow a description of the FHR signals, from different points of view, as much complete as possible. To provide a time-frequency view of FHR tracings, both spectrograms and scalograms were considered. Scalograms represent the analog to power spectrums when dealing with wavelet transforms; they generally provide a better time localization for rapid, high frequency events and a better frequency localization for low-frequency, longer-duration events. However, since the best time-frequency representation depends on the specific application, both spectrograms and scalograms were included in our study. The use of scalograms to encode FHR tracings into images, used as inputs for a CNN net, were already proposed in the work by **[6.13]**. However, their dataset was limited to 552 records (of which 447 normal and 105 pathological), so that a process of data augmentation was necessary to obtain a sufficient number of records for the training of the proposed neural model.

Moreover, persistence spectrums were included in our study since they provide information about the persistence of a certain frequency in a signal during its evolution.

Together with the aforementioned techniques, which are used to obtain time-frequency representations of signals, other methods were employed to explore different aspects of FHR tracings, such as their evolutional dynamics. Among these, Markov Transition Fields (MTF) were taken into account. The latter allow to obtain a visual representation of the transition probabilities, for each time point in the sequence, that maintains their sequentially, in order to preserve information in the temporal dimension.

To explore the presence of recurrent patterns or irregular cyclicities in the FHR tracings, recurrence plots (RP) were also considered, as they provide visual representations that reveal all the times when the phase space trajectory of a dynamical system visits roughly the same area in the phase space.

Another transformation technique that we included in our work is Gramian Angular Field (GAF), which provides a description of the temporal correlation structure of a time series, through the use of a polar coordinate system.

MTF, RP and GAF have already been employed as methods to transform time series into images, as illustrated, for example, in the work from **[6.14]**. However, none of the studies in the literature reports the use of these techniques to encode FHR tracings.

The different transformations were applied to encode each 20 minutes of clean FHR signal into corresponding images, as illustrated as follows.

Even in this case, the images generated through the different methods were characterized by different scales and were so mapped in [0,1] range through min-max normalization. Moreover, all images so far obtained, were reduced in size to a dimension of 64x64x1.

Signals to images encodings were implemented by using the software MATLAB 2022a (The Math Works, Inc.).

From here on, the mentioned techniques used to convert the FHR series into images will be detailed.

Continuous Wavelet Transform (CWT)

The wavelet transform (WT) **[6.15]** is a mapping from $L^2(R) \rightarrow L^2(R^2)$, with superior time-frequency localization as compared to the Short Time Fourier Transform (STFT). This characteristic opens up the possibility of a multiresolution analysis.

WT have been extensively employed in biomedical engineering to analyze non-stationary and nonlinear signals over the last decades. CWT presents great abilities, such as its flexible capacity to extract general and fine-grained feature information from the input signal.

CWT is a formal tool that provides a hyper-complete representation of a signal by performing the convolution of a signal with a rapidly decaying oscillating finite-length, waveform, called mother wavelet, whose translation and scaling varies continuously.

The result of these convolutions is a series of coefficients, obtained for each time point, that are used to create a 2D representation of the signal, called scalogram. The x-axis coincides with the time axis and the y-axis with the scaling factor of the mother wavelet.

Each point of this 2D map represents the intensity of the corresponding (associated) coefficient and it is shown using a particular colormap.

The CWT is defined as the summation of the overall signal spectrum multiplied by the compressed and translated "mother wavelet", which can be expressed mathematically by the following **equations**:

$$CWT(\tau, s) = \psi(\tau, s) = \int_{-\infty}^{\infty} f(t)\phi_{\tau}^{*} \ , \ f(t) \epsilon L^{2}(R) \qquad (6.1)$$

$$\phi_{\tau,s}(t) = \frac{1}{\sqrt{|S|}}\phi\left(\frac{t-\tau}{s}\right) \qquad (6.2)$$

$$\tau, s \ \epsilon \ R, s \neq 0 \ , \int_{-\infty}^{+\infty} \phi(t)\, dt = 0$$

In **equation 6.1** *f(t)* represents the input signal, *φ(t)* is the wavelet basis, and *ψ (τ, s)* represents the wavelet coefficient, a function of two variables. *τ* is the translation factor that reflects temporal and spatial information and represents the translation diameter of time shifting; s is the scaling factor that determines the degree to which the wavelet is compressed or stretched.

An example of image obtained by applying the CWT to 20 minutes FHR signal of our DB is shown in **Figure 6.1 (a)**. The yellow portions in the image represent the coefficients with the higher intensity values while the blue portions define the coefficients with the lower intensities.

The primary reason for applying the CWT in this research is that the CWT can provide a better method than others for observing and capturing the local characteristic information which is hidden in the FHR signal both in time and frequency domains.

Gramian Angular Field (GAF)

Gramian Angular Field (GAF) **[6.14]** generates an image, obtained from a time series, which shows the temporal correlations between each time point in the time signal. GAF images represent a time series in a polar coordinate system instead of the typical Cartesian coordinates.

GAF images depict the relationship between every point and each other in the time series, that is, it displays the temporal correlation structure in the series. The greatest advantage of GAF is that it can preserve temporal dependencies and leading to bijective encodings.

It is possible to obtain two different kinds of GAF, i.e., the Gramian Angular Difference Field (GADF) and the Gramian Angular Summation Field (GASF).

First, let's recall that the dot product (or inner product) is a measure of similarity between two arrays. I.e., if we consider two vectors (*x1* and *x2*), the dot product between *x1* and *x2* is defined as follows:

$$\langle \boldsymbol{x_1}, \boldsymbol{x_2} \rangle = \|\boldsymbol{x_1}\| \|\boldsymbol{x_2}\| \cos \boldsymbol{\theta} \qquad (6.3)$$

where θ is the angle between *x1* and *x2*. If we don't consider the magnitude of the arrays, we can state that if the angle between *x1* and *x2* is small (i.e., close to 0) then the cosine of that angle will be nearly 1. If *x1* and *x2* are perpendicular, the cosine of the angle is 0. If the two vectors are pointing in opposite directions, the cosine will be -1. Starting from this consideration it is possible to obtain the Gram Matrix *G,* as:

$$\boldsymbol{G} = \begin{pmatrix} \langle \boldsymbol{x_1}, \boldsymbol{x_1} \rangle & \cdots & \langle \boldsymbol{x_1}, \boldsymbol{x_N} \rangle \\ \vdots & \ddots & \vdots \\ \langle \boldsymbol{x_N}, \boldsymbol{x_1} \rangle & \cdots & \langle \boldsymbol{x_N}, \boldsymbol{x_N} \rangle \end{pmatrix} \qquad (6.4)$$

Gramian Angular Field (GAF) is obtained by introducing "special" inner product as follows.

Let's suppose we are given a time series X= $\{x_1, x_2, \ldots, x_N\}$, normalized to be in [-1,1]. The first step is to convert each value in X into polar coordinates, through the transformation:

$$\boldsymbol{\phi_i} = \mathbf{arccos}(\boldsymbol{x_i}) \qquad (6.5)$$

Finally, the GAF method defines its "special" inner product as:

$$\langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle = \mathbf{cos}(\boldsymbol{\phi_i} + \boldsymbol{\phi j}) \tag{6.6}$$

to obtain GASF, and:

$$\langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle = \mathbf{sin}(\boldsymbol{\phi_i} - \boldsymbol{\phi_j}) \tag{6.7}$$

to obtain GADF.

An example of a GASF and a GADF images, obtained from an FHR signal of 20 mins of our dataset, are reported in **Figure 6.1 (b) and (c)** respectively.

Markov Transition Field (MTF)

Markov Transition Field (MTF) **[6.14]** provides an image which is obtained from a time series. The image contents represent a field of transition probabilities for a discretized time series. For an n-length time series, MTF is a $n \times n$ matrix containing the probability of a one-step transition from the bin for x_k to the bin for x_l, where x_k and x_l are two points in the time series at arbitrary time steps k and l.
More in detail, given a time series $X = \{x_1, x_2, \ldots, x_n\}$, a data point x_i is assigned to its corresponding quantile bin q_j ($1 \leq j \leq Q$) where Q is the number of bins, i.e., the number of states. In this way we can construct, from X, a Markov chain, deriving the $Q \times Q$ Markov transition matrix (W) where wij ($1 \leq I, j \leq Q$) in W is the frequency with which a data point in the state q_j is followed by a data point in state q_i. After normalization, W is the Markov Transition Matrix, where w_{ij} represents the transition probability of $q_i \rightarrow q_j$ in the MTF. By assigning the probability from the quantile at time step i to the quantile at time step j at each pixel of M_{ij}, the MTF, denoted as M (**equation 6.8**), encodes the multi-span transition probabilities of the time series. The main diagonal M_{ii} captures the probability from each quantile to itself (the self-transition probability) at time step i.

$$\boldsymbol{M} = \begin{bmatrix} w_{ij}|x_1 \epsilon\, q_i\,, x_1 \;\epsilon\, q_j & \cdots & w_{ij}|x_1 \;\epsilon\, q_i, x_n \;\epsilon\, q_j \\ \vdots & \ddots & \vdots \\ w_{ij}|x_1 \;\epsilon\, q_i, x_1 \;\epsilon\, q_j & \cdots & w_{ij}|x_1 \;\epsilon\, q_i, x_n \;\epsilon\, q_j \end{bmatrix} \quad (6.8)$$

To make the image size manageable and the computation more efficient, we reduced the MTF size by averaging the pixels in each non-overlapping $m \times m$ patch, that is we aggregate the transition probabilities in each subsequence of length *m* together.

Figure 6.1 (d) shows an example of an MTF image obtained from a 20-minute FHR sequence belonging to the dataset used in our work.

Recurrence Plot (RP)

Recurrence plots (RP) were introduced as a visualization tool to measure the time constancy of dynamical systems **[6.14]**.

Natural processes can have distinct recurrent behaviors like periodicities (as seasonal cycles) or irregular cyclicities. A RP, generally defined as *R*, depicts all the time instants when the phase space trajectory of a dynamical system visits the same area in the phase space. A recurrence of a state at time *i* at a different time *j* is marked within a two-dimensional squared matrix where both axes represent time.

A RP can be generated by first computing a distance matrix *S* that contains each distance from one point in the time series with each other and then applying a threshold ε to binarize the values.

In mathematical terms, the RP is defined by matrix *R*, as in **equation 6.9**:

$$\boldsymbol{R(i,j)} = \begin{cases} \boldsymbol{1} \;\; \boldsymbol{if} \; \left\|\vec{x}(i) - \vec{x}(j)\right\| \le \varepsilon \\ \quad \boldsymbol{0} \quad\;\; \boldsymbol{otherwise} \end{cases} \quad (6.9)$$

$\vec{x}(i)$ and $\vec{x}(j)$ represent the states, ‖.‖ is the first norm and ε is a chosen threshold. $R(i,j)$ is 1 if $\vec{x}(i) \approx \vec{x}(j)$ up to an error ε. This is fundamental since systems often do not recur exactly to a previously visited state but tend to visit roughly the same area in the phase space. Similarly, a RP can be generated by first computing a distance matrix S (**equation 6.10**), that contains each pair of distances and then applying the threshold ε.

$$\boldsymbol{S} = \left|\left|\vec{\boldsymbol{x}}(\boldsymbol{i}) - \vec{\boldsymbol{x}}(\boldsymbol{j})\right|\right| \qquad (\mathbf{6.10})$$

An example of S and R images obtained from an FHR series included in our dataset is shown in **Figure 6.1**, respectively **(e) and (f)**.

Power Spectrogram (PS)

Power spectrogram (PS) **[6.16]** is a visual representation of the frequency spectrum of a signal (y-axis) as it varies with time (*x*-axis). The most common way to show a spectrogram is using a heat map which uses a system of color-coding to represent different intensity values.

Given a time series, we can estimate spectrograms with methods based on Fourier transform (FT) or by using filter banks. Our choice was to adopt a FFT approach. This method splits data into chunks, which usually overlap, and proceed to compute the Fourier transform of each chunk to calculate the relating frequency spectrum magnitude. Each vertical line in the image corresponds to a chunk, a measurement of magnitude versus frequency for a specific moment in time. These so-called spectra are then put sequentially to form the image. Hence, given a time series *s(t)*, to retrieve the image we need to apply a short-time Fourier transform (STFT) on the signal *s(t)* and window width *ω* (**equation 6.11**)

$$\boldsymbol{spectrogram}(\boldsymbol{t}, \boldsymbol{\omega}) = |\boldsymbol{STFT}(\boldsymbol{t}, \boldsymbol{\omega})|^2 \qquad (\mathbf{6.11})$$

The spectrogram of a 20-minute FHR signal from the DB used in this work is shown in **Figure 6.1 (g)**.

Persistence Spectrum (PSP)

The persistence spectrum of a signal is a time-frequency representation that shows the percentage of time a given frequency is present in a signal.

The persistence spectrum is a histogram in power-frequency space. The longer a particular frequency persists in a signal as it evolves, the higher its time percentage and thus the “brighter” or "hotter" its color in the display.

The calculation of the persistence spectrum is obtained by first computing the spectrogram for a time segment. After that, power and frequency values are partitioned into 2-D bins. For each time value, a bivariate histogram of the logarithm of the power spectrum is computed. For every power-frequency bin where there is signal energy at that instant the corresponding matrix element is increased by one. The persistence spectrum is obtained through the computation of the sum of all the histograms related to every time value. The image obtained presents the Frequency (Hz) on the x-axis and the Power Spectrum (dB) on the y-axis. **Figure 6.1 (h)** shows an example of Persistence Spectrum extracted from an FHR signal belonging to the dataset used in our study.

6.3. The proposed MLP net.

The proposed MLP net is the resulting architecture obtained after testing different combinations of layers and parameters. As anticipated, the MLP is fed with a set of 15 quantitative indices describing various aspects of the FHR signals (denoised and with a length of 20 minutes), in both time and frequency domains. This net is trained to learn to distinguish whether the provided set of input parameters describe a reassuring tracing or a pathological one.

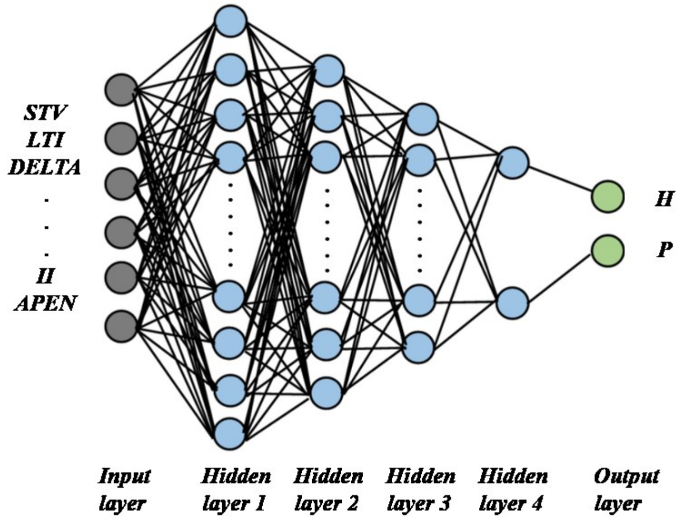


Figure 6.2: The present figure illustrates the scheme of the proposed MLP net. The Input layer is composed of 15 neurons, each receiving one of the 15 indices describing the i^{th} 20-min FHR signal. The Input layer is then followed by 4 Hidden Layers, the last of which is connected to an Output Layer formed by two neurons (H-Healthy, P-Pathological).

An illustration of the proposed MLP is shown in **Figure 6.2**, while the details of each layer composing the architecture are reported as follows:

Input layer: composed of 15 neurons, one for each quantitative parameter passed in input. These neurons are fully connected to the ones of the first hidden layer.

Hidden layer 1: Composed of 500 neurons with ReLU activation function, followed by a Dropout layer with a probability of 0.4, to avoid overfitting.

Hidden layer 2: Composed of 250 neurons with ReLU activation function, followed by a Dropout layer with a probability of 0.4.

Hidden layer 3: Composed of 150 neurons, with ReLU activation function. L1 and L2 regularization penalty is applied. The value for L1 is set to 10 −5, for L2 is 10 $^{-4}$.

Hidden layer 4: Composed of 50 neurons, with ReLU activation function followed by a Dropout layer with a probability of 0.4.

Output layer: The 50 nodes of the fourth hidden layer are fully connected to the 2 last neurons of the output layer, with Softmax activation function.

6.4. The proposed CNN net.

As previously stated, the idea behind the choice of a CNN architecture is trying to leverage the inherent capability of these kind of network to learn complex features from the available data, without condensing them through the use of any statistical regressor. Since the most suitable inputs for CNNs are shaped to be in the form of two-dimensional arrays, we proceeded to fed the proposed net with a set of images, obtained through the techniques described in **Section 6.2.2**, which furnish an alternate 2-D representation of FHR signals. Particularly, we decided to reshape the inputs to be in the form of 64 x 64 gray level images.
Even in this case, this net is trained to learn to distinguish whether the provided set of images describe a reassuring tracing or one coming from a likely pathologic fetus.
A complete description of each layer is reported as follows.

Input layer: The CNN input layer receives as input an array 64 x 64 x 1 x *n*, where *n* stands for the number of images fed to the net. The array is created by concatenating n images on the fourth dimension, with n=1…8. The images building up the input array are the ones described in **Section 6.2.2**.

Convolutional 2D Layer: The input layer nodes are convoluted by using 16 filters of 5 x 5 kernel, with no padding and ReLU activation function. A Batch Normalization layer is then used to re-scale and recenter the input layer to make the network more stable and faster.

Max Pooling 2D Layer: The first convolutional layer is followed by a Max Pooling Layer with pool size 2 x 2. The pooling operation reduces the eigenarrays of the convolution output and the number of parameters, so it can lower the model complexity and speed up the computation while preventing overfitting.

Convolutional 2D Layer: The second convolutional layer is formed by 32 filters with 5 x 5 kernels, ReLU activation function and no padding, followed by a Batch Normalization layer.

Max Pooling 2D Layer: After the second convolutional layer, a Max Pooling Layer with pool size 2 x 2 is added. Dropout is applied with probability 0.8.

Flatten Layer: To unroll the output of the convolutional layers, a Flatten layer is applied.

Dense Layer: Each neuron of the Flatten Layer is fully connected to the 64 neurons of the successive Dense Layer with ReLU activation function, followed by a Batch Normalization layer. A Dropout with 0.8 rate is then applied.

Dense Layer: The 16 neurons with ReLU activation function are fully connected to the last 2 neurons of the output layer.

Output Layer: Consists of 2 neurons, one per class, with Softmax activation function.

A visual illustration of the described CNN is depicted in **Figure 6.3**.

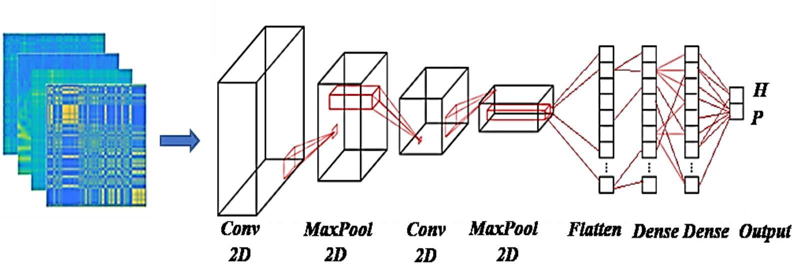


Figure 6.3: The present figure illustrates the scheme of the proposed CNN net. The Input layer, which receives in input a set of images in the shape 64 x 64 x n, with n = 1, …, 8, is followed by two blocks, each composed by a 2D convolution layer and a 2D Max Pooling layer. These are followed by a Flatten Layer and by two Dese layers, the last of which is connected to an Output Layer formed by two neurons (H-Healthy, P-Pathological).

6.5. The hybrid MLP+CNN net: a novel mixed data type approach.

As reported at the beginning of the chapter, the core idea was to design a neural network capable of dealing with heterogeneous data, i.e., a set of scalar values summarizing a signal processing pipeline and a set of images which represent the whole FHR signal in different domains (time-frequency, recurrent periodicities).

Our aim is to integrate the information automatically grasped by two connected branches, each of which is provided with a different kind of input (i.e., images and arrays of values).

This type of approach combines parameters already known to provide information about the physiological mechanisms responsible of the FHR signal, with other characteristics obtained from an implicit understanding made by the model itself. More precisely, the network we designed was fed with an array of 15 quantitative regressors and a set of images, obtained from each FHR sequence of 2,400 samples (20 min length, as reported in the previous sections).

The proposed hybrid net is composed by two branches, which exactly correspond to the MLP and CNN architectures presented in **Sections 6.3** and **6.4**.

The CNN+MLP net was hence obtained by concatenating the outputs of the two separate branches to form a single output array, which is passed to the subsequent fully connected layers, through a concatenation layer. From a structural viewpoint, the terminal neurons of MLP and CNN branches are connected to form a flatten layer so that the input to the final set of layers is the output of the layer where MLP and CNN branches are concatenated. This one is followed by a Dense layer of 128 neurons with ReLU activation function. The nodes of the Dense layer are then fully connected to each of the 2 neurons of the output layer, which use a Softmax activation function. These ones give back the probability of the input passed to the artificial network belonging to one of the 2 possible classes (healthy or unhealthy fetus).

An illustration of the proposed hybrid net is furnished in **Figure 6.4**.

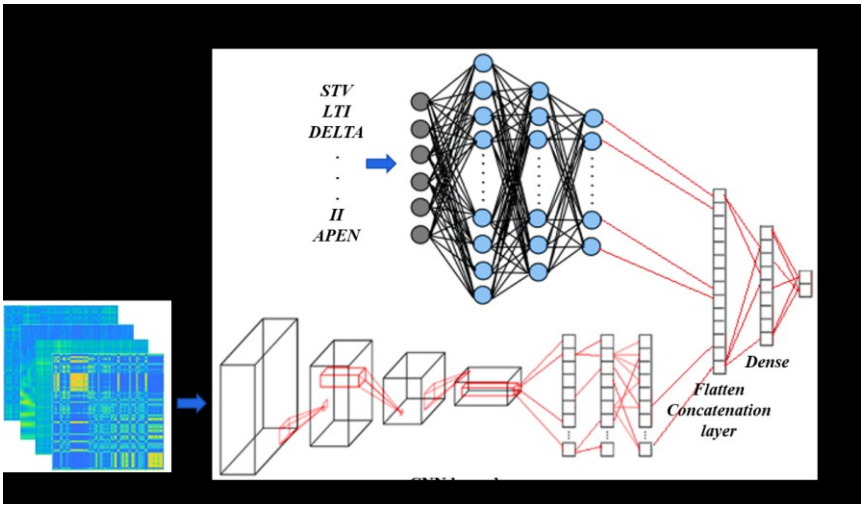


Figure 6.4: The present figure illustrates the scheme of the proposed mixed CNN + MLP net. The CNN and MLP branches are connected through a Flatten Concatenation Layer. The latter is fully connected to a Dense Layer, which is connected

to an Output Layer formed by two neurons (H-Healthy, P-Pathological)

6.6. Training and testing.

The dataset used to train and test the performances of our proposed neural classifier is the one described in **Section 6.2**. It is formed by 14,000 labelled examples, of which 7,000 correspond to healthy fetuses and 7,000 to pathological ones. Each example, which relate to a diverse FHR trace, consists of a set of 15 quantitative parameters and a group of images. 80% of the dataset (i.e., 11,200 data) was used to train the neural networks, while 20% (i.e., 2,800 data) was used for testing the performances of the trained nets.

The neural model setup was carried out by using Python 3.7 and for the training phase the online virtual machines provided by Kaggle were used (https://www.kaggle.com/).

We adopted the Adam optimizer with a learning rate of 10^{-4} and a decay rate of $10^{-4/200}$ for the training of the network.

Binary-cross-entropy was designed as the loss function to be optimized.

Early Stopping technique, with a patience of 2, was employed as an overfitting prevention technique. We always analyzed the relation between the accuracy and loss curves obtained in the different training sessions, in order to verify that overfitting was not occurring. For example, **Figure 6.5** shows the accuracy and loss trends on the training examples for the CNN+MLP net, as functions of the epochs, compared with the accuracy and loss curves on the validation examples. The use of Early stopping, interrupts the training phase at the 90th epoch, preventing the model from excessively adapting to the training data. The crossing point of the red dashed lines in both diagrams of **Figure 6.5** identifies the point where the training is interrupted by the stop criterion.

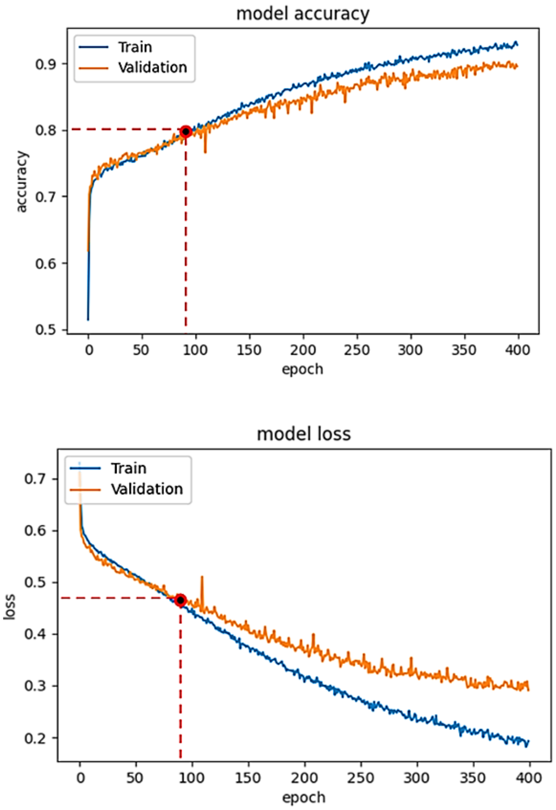


Figure 6.5: CNN+MLP model: Accuracy curve on training and validation set vs. epochs and loss curve on training and validation set vs. epochs. After about 90 epochs the training is interrupted by the anti-overfitting stop criterion.

6.7. Results.

We separately evaluated the performance of the proposed MLP and CNN nets and compared them with the ones obtained by the CNN+MLP mixed model, in order to state if the latter could over perform as respect to the MLP and CNN branches singularly.
To provide robustness to the analysis, we repeated the training process 30 times for each of the considered nets. After each training phase was completed, accuracy (ACC),

sensitivity or True Positive Rate (TPR), specificity or True Negative Rate (TNR), precision or Positive Predictive Value (PPV), negative predictive value (NPV), False Positive Rate (FPR), False Negative Rate (FNR), False Discovery Rate (FDR) and Area Under the ROC Curve (AUC) were computed. The definition of each of these performance metrics is reported as follows:

$$\boldsymbol{ACC = \frac{TP + TN}{TP + TN + FP + FN}} \tag{6.12}$$

$$\boldsymbol{TPR = \frac{TP}{TP + FN}} \tag{6.13}$$

$$\boldsymbol{TNR = \frac{TN}{TN + FP}} \tag{6.14}$$

$$\boldsymbol{PPV = \frac{TP}{TP + FP}} \tag{6.15}$$

$$\boldsymbol{NPV = \frac{TP}{TP + FN}} \tag{6.16}$$

$$\boldsymbol{FPR = \frac{FP}{FP + TN}} \tag{6.17}$$

$$\boldsymbol{FNR = \frac{FN}{TP + FN}} \tag{6.18}$$

$$\boldsymbol{FDR = \frac{FP}{TP + FP}} \tag{6.19}$$

To infer significant statistical differences in terms of average classification accuracy, between the three architectures proposed, T-test was applied.

The average confusion matrices for the three proposed models are reported in **Table 6.1**.

The average values, for the different performance metrics computed for the three nets, are instead summarized in **Table 6.2**.

Table 6.1: Confusion Matrix for the MLP, CNN, CNN+MLP models obtained on the 2800 examples of test. TP = True Positive, TN = True Negative, FN = False Negative, FP = False Positive, TP = True Pathological, TH = True Healthy, PP = Predicted Pathological, PH = Predicted Healthy.

	MLP		**CNN**		**CNN+MLP**	
	PP	**PH**	**PP**	**PH**	**PP**	**PH**
TP	TP = 998	FN = 427	TP = 681	FN = 596	TP = 960	FN = 431
TH	FP = 258	TN = 1117	FP = 298	TN = 1225	FP= 109	TN = 1300

Table 6.2: Performance metrics for the MLP, CNN, CNN + MLP models. It reports: True Positive Rate (TPR = TP/(TP+FN)), even called Recall or Sensitivity, True Negative Rate (TNR = TN/(TN+FP)) or Specificity, Positive Predictive Value (PPV = TP/ (TP + FP)) or precision, Negative predictive value (NPV = TP/ (TP + FN)), Fall out or false positive rate (FPR = FP/ (FP + TN)), False negative rate (FNR = FN/ (TP + FN)), False discovery rate (FDR = FP/ (TP + FP)).

	Performance metrics							
	TPR	**TNR**	**PPV**	**NPV**	**FPR**	**FNR**	**FDR**	**AUC**
MLP	0.7	0.81	0.79	0.72	0.18	0.29	0.2	0.76
CNN	0.53	0.80	0.69	0.53	0.19	0.46	0.3	0.67
CNN+MLP	0.69	0.92	0.90	0.75	0.08	0.31	0.1	0.81

The mean accuracy reached by the single MLP, over the 30 train trials, on the 2’800 test examples, was 75.5%, i.e., 2’115 correct classifications against 685 misclassifications.

For what concerns the CNN branch, we firstly had to choose which combination of images provided as input to the net, could lead to the best results. For that aim, we tested all the different combinations of images, and for each of them we trained the CNN a number of 15 times, computing the classification accuracy at each step.

At the end of the process, we selected the combination of images providing in average the highest accuracy. By looking at the results achieved, the most impactful set of images results to be composed by GADF, PS and PSP.

After selecting the most performing CNN architecture, we trained the latter a number of 30 times and for each phase we computed all the performance metrics. The overall accuracy obtained is 68.1%, i.e., 1'906 correct classifications against 894 misclassifications.

After evaluating MLP and CNN branches separately, we tested the performances of the combined CNN+MLP model, that concatenates MLP and CNN nets in a single mixed architecture.

As for the single CNN case, we had to select the top performing combination of images to feed the CNN branch of the combined model. Even in this case, the most impactful images have been proved to be GADF, PS and PSP.

After selecting the most suitable inputs for the CNN branch of the combined model, we repeated the training phase of the CNN+MLP net 30 times and as for the single CNN case, we computed all the performance metrics for each phase.

A summary of the overall accuracy achieved by the different models evaluated is reported in **Table 6.3**, while the corresponding boxplot and ROC curves are illustrated in **Figure 6.6 and 6.7** respectively.

Table 6.3: Summary of overall accuracy achieved for the MLP, CNN and CNN+MLP models.

	Mean ACC	Number of correct classifications	Number of wrong classifications
MLP	75.7%	2120	680
CNN	68.1%	1907	893
CNN + MLP	**80.1%**	**2260**	**540**

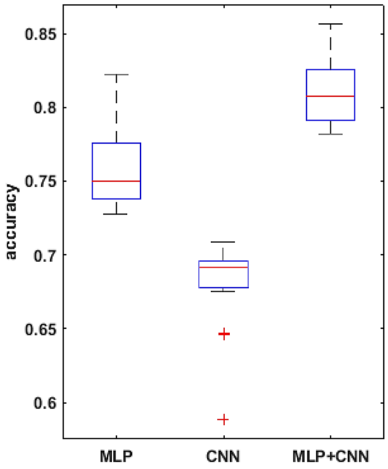


Figure 6.6: Boxplots for the mean accuracy values reached, over the 30 replications of the training phase, for the 3 models compared, i.e., MLP, CNN, combined CNN+MLP. The hybrid CNN+MLP data outperforms, showing an average accuracy of 0.801, against the 0.757 and 0.681 respectively achieved by the MLP and CNN.

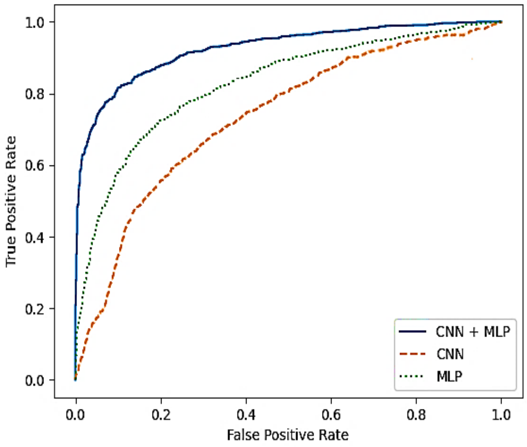


Figure 6.7: ROC curves for MLP, CNN and CNN+MLP combined neural model developed.

To prove significant statistical differences among the three nets explored, T-test was applied to challenge the null hypothesis (H0) of equality, in terms of average classification accuracy, between the three models. With a level of significance of 0.01, H0 was refused for every comparison performed.
From the observation of the obtained results, it appears how the use of the convolutional branch alone does not allow to reach an adequate classification accuracy, showing lower performances than those obtained with the single MLP branch. However, the results achieved with the combined CNN+MLP model show a significant increase in the classification capacity of the model, compared to the MLP and CNN architectures individually considered. The combined CNN+MLP model, in fact, reached an overall classification accuracy of 80.1%. This corresponds to a total number of 2260 correct classifications against 540 misclassifications.
The combined model proposed, hence, seems to be able to exploit the good accuracy of the MLP to influence and boost the performance of the CNN on the provided images, confirming how the combined use of known quantitative

regressors and features, implicitly learnt from the neural model, could increase the classification capabilities.

There is however to point out that the neural model realized tends to better classify the signals related to healthy fetuses (FP = 109, FN = 431). In fact, the CNN-MLP model presents a high specificity (TNR) of 92%, but its sensitivity (TPR) is of 69%. This means that the proposed architecture misclassifies a signal related to a healthy subject the 8% of times while misses the classification 31% of times when dealing with a signal referred to an unhealthy subject.

6.8. Discussions and conclusions

The possibility to identify early signs of fetal sufferance antepartum still remains a dream in the Ob-Gyn management of pregnancies. An accurate disambiguation between healthy and suffering fetuses can allow obstetricians to intervene in a timely manner and take appropriate actions to prevent permanent damages to the fetus. Among the prenatal exams, the CTG represents the major source of information on the correct development of the fetus.

Despite the fast increase of the digital technology in medical devices, in the clinical practice, the analysis of CTG signals, both antepartum and during labor, is mostly carried out by visual analysis of the tracings. This procedure is obviously affected by significant inter-observer and intra-observer variability, which often causes erroneous interpretations of real fetal conditions.

The introduction of computerized CTG analysis decreased the qualitative and subjective interpretation of the CTG exam, but didn’t lead to a reliable clinical decision-making strategy, despite the great effort produced in the past twenty years for extracting significant quantitative indices from the FHR signal.

Artificial Intelligence techniques, with a particular focus on Deep Learning, represent a further tool to investigate the information content of CTG tracings, although they need huge datasets in order to provide reliable conclusions. As we had available a considerable amount of annotated CTG exams, we decided to approach the problem of classifying normal and pathological fetuses by means of those methods. The availability of a very large and structured database, consisting of real labeled data that were collected in the same clinical department, represents the first important aspect of this work. This feature is difficult to find in the field of fetal monitoring. It has made possible to exploit machine learning and deep learning methods to the best of their abilities. In fact, it is known that the classification power of AI methods is best expressed only with large amounts of data, which was not allowed until now for the analysis of the fetal heart variability signal.

A second factor is the correspondence between the quantitative values of the parameters used for classification and the fetal and maternal physiology. Each parameter we have employed (and the 15-feature set is an example), can contribute to the understanding of the physiological mechanisms that controls fetal heart. These features make readable and interpretable the data set in terms of control developed by physiological systems.

The classification proposed in this work benefits from the information contained in these parameters. Therefore, it is possible to formulate a classification between healthy and pathological fetuses that is interpretable according to involved pathophysiology, whose measurements take place through the parameters extracted in the FHR.

We designed and implemented a neural architecture able to deal with heterogeneous data, i.e., images and quantitative parameters describing the statistical characteristics of the FHR signal. The neural network consists of two branches, a MLP receiving an array of 15 regressors and a CNN one fed with a set of 64x64 images. The latter have been obtained through several transformations (e.g., MTF, GADF, RP, etc.) applied to the pre-processed denoised FHR signal.

To understand if the novel mixed-type architecture outperforms the MLP and CNN branches singularly, we compared the results obtained from the three neural architectures in terms of overall classification accuracy. After the hyperparameters' optimization for each NNs, the MLP, CNN and MLP-CNN architecture have been trained and tested on a set of 14 K data (split in 80% for the training and 20% for the test). The results obtained have shown that the MLP-CNN network is the best performing architecture. Hence, with the best set of hyperparameters, this mixed-type net achieved an overall classification accuracy of 80.1%.

The major limitation of the Method still lies in the sensitivity, which is not yet fully satisfactory. In fact, with the combined model (CNN+MLP) the TPR reached is of 69%, that corresponds to a probability of erroneous classification of an unhealthy subject of 31%.

There is, however, to consider that unhealthy subjects contained in the database and used for this work, are a heterogeneous group, which includes several types of diseases: intra uterine growth restriction (IUGR), metabolic alterations, fetal malformations, and even maternal pathological conditions, such as diabetes.

The decision to include all the different categories of disease in the unhealthy class made it possible to balance the number of healthy and unhealthy subjects with a sufficient numerosity to allow the use of Deep Learning techniques. This could lower the performance of our neural model, both in terms of accuracy and specificity, since different pathologies could show different behaviors in the FHR signals, reducing the classification capacities and increasing the variability of FHR features. These analyses must be considered as a starting point in the direction of more complex studies, that look at the different classes of pathologies separately, once the amount of data for each pathology will reach an acceptable value for Deep Learning methods.

Nevertheless, the obtained results are promising, since they have been achieved by using a noticeable amount of clinical data, whose variability closely represents the real population. Although this fact may reduce the classification performance, as compared to other existing works, it can

however increase the robustness and the generalization ability of the model.
Further developments for this work include the search for other techniques for converting the CTG signals into images, in order to provide new kinds of inputs to the CNN branch. In addition, more quantitative parameters, to feed the MLP branch, will be investigated. Moreover, other mixed-type neural architectures will be explored and will include other types of neural branches such as Recurrent NN or Temporal CNN.

6.9. References for Chapter 6.

[6.1] P.Fergus, A.Hussain et al., “Classification of caesarean section and normal vaginal deliveries using foetal heart rate signals and advanced machine learning algorithms”., BioMedical Engineering OnLine 16, 2017.
[6.2] A.Petrozziello, I.Jordanov et al. “Deep learning for continuous electronic fetal monitoring in labor”, 2018, 5866-5869.
[6.3] M.S.Iraji, “Prediction of fetal state from the cardiotocogram recordings using neural network models”, Artificial Intelligence in Medicine 96, 2019.
[6.4] Z.Zhao, Y.Deng et al. “Intelligent prediction of fetal acidemia using fetal heart rate signals based on convolutional neural network”, BMC Med Inform Decis Mak 19, 2019.
[6.5] N.Rahmayanti et al. “Comparison of machine learning algorithms to classify fetal health using cardiotocogram data”, Procedia Computer Science 197 162-171, 2022.
[6.6] S.Liu S et al. “Doppler Ultrasound Imaging Combined with Fetal Heart Detection in Predicting Fetal Distress in Pregnancy-Induced Hypertension under the Guidance of Artificial Intelligence Algorithm”, J Healthc Eng, 2021.
[6.7] K.H.Ah, K.S.Lee, “Artificial intelligence in obstetrics”, Obstet Gynecol Sci 65(2) 113-124, 2022.

[6.8] M.G.Signorini, N.Pini et al. “Integrating machine learning techniques and physiology based heart rate features for antepartum fetal monitoring. Computer Methods and Programs in Biomedicine 185 (2020).

[6.9] D.Arduini, A.Rizzo et al. Computerized analysis of fetal heart rate. Journal of Perinatal Medicine 22 (1994) 22-27.

[6.10] M.G.Signorini, G.Magenes et al., “Linear and Nonlinear Parameters for the Analysis of Fetal Heart Rate Signal from Cardiotocographic Recordings”, IEEE Trans Biom Eng. 50 365-374, 2003.

[6.11] R.Mantel, H.P.van Geijn et al. “Computer analysis of antepartum fetal heart rate: 1. Baseline determination”, International Journal of Bio-Medical Computing 25 261-272 1990.

[6.12] A. Krizhevsky, I.Sutskever, G.E.Hinton, “ImageNet classification with deep convolutional neural networks”, Proceedings of NIPS, Lake Tahoe, Nevada 1097-1105 2012.

[6.13] Z.Zhao, Y.Deng et al., “Intelligent prediction of fetal acidemia using fetal heart rate signals based on convolutional neural network”, BMC Med Inform Decis Mak 19 286, 2019.

[6.14] Z.Wang, T.Oates ,”Imaging time-series to improve classification and imputation”, IJCAI International Joint Conference on Artificial Intelligence 3939-3945, 2015.

[6.15] Daubechies I. The wavelet transform, time-frequency localization and signal analysis. J Renew Sustain Energy 36 (2015) 961-1005.

[6.16] T.Sawa, T.Yamad et al., “Power spectrum and spectrogram of EEG analysis during general anesthesia: Python-based computer programming analysis”. J. Clin. Monit. Comput, 2021.

Overall discussions and conclusions.

This thesis collects and illustrates the research carried out during my PhD program. The activities included in this work are part of the European PRIN project ICT4MOMs, which aims to develop a system for the intelligent monitoring of mother and fetus, during pregnancy.

In this frame, my role was to develop diverse artificial intelligence (AI) solutions to face and eventually overcome some of the limitations affecting traditional CTG analysis.

The latter is the most widely employed technique to monitor the well-being of the fetus during pregnancy, based on the inspection of fetal heart rate variability (FHR) series, which are obtained as explained in **Chapters 1** and **2**.

The effectiveness of CTG, has been proven in labor; however, its usefulness in the ante-partum period remains questionable. This is primarily due to challenges in interpreting the intricacies of CTG recordings and the absence of consensus on evaluation criteria. While visual examination of FHR tracings is the current standard method, however it fails to encompass all the valuable information present in the signals. Efforts to address this through computerized CTG (cCTG) have aimed to identify a single indicator but have fallen short in delivering an adequate evaluation. Recognizing the necessity of a multi-dimensional approach, researchers have sought to integrate various methodologies.

As detailed in **Chapter 3**, there is considerable potential for utilizing AI techniques to analyze CTG data and overcome the limitations associated with visually inspecting CTG tracings. However, a significant hindrance currently exists due to the limited availability of reliable, well-organized, and labeled clinical datasets. This obstacle is especially noteworthy because the most effective AI methods, such as deep learning algorithms (DL), necessitate a substantial volume of data to achieve accurate and high-quality outcomes.

For these reasons, the first step towards an effective implementation of AI in the field of CTG is the creation of a large and structured CTG dataset. **Chapter 4** illustrates the steps that, starting from an unorganized set of CTG recordings, collected at the Federico II Hospital in Napoli, Italy, have taken to the setup of a large (more than 30k records), balanced, structured dataset of labelled CTG recordings, which we have baptized with the name NAPAMI. At date, the latter is one of the largest datasets on antepartum monitoring of healthy and high-risk pregnancies. It is an organized and continuous gathering of clinical data on pregnant patients and their fetuses, spanning from pregnancy to delivery. In high-risk pregnancies, fetal monitoring data can be obtained as early as the 26th week, and in some cases, information on delivery and neonatal outcomes is also accessible. The thorough collection of clinical data provided by NAPAMI, combined with a dedication to staying up to date with clinical evidence and the ability to continually assess, offers fetal medicine healthcare professionals a more precise understanding of fetal well-being. This knowledge is particularly valuable in making informed decisions regarding optimal delivery timing, especially in instances where delivery is necessary before the 34-week mark and there are risks associated with prematurity and fetal pathology.

The dataset resulting from the various stages of data processing outlined in the manuscript allows for the integration of existing clinical knowledge with findings from experimental research. This integration verifies the practical applicability of the gathered information in clinical settings and generates new evidence where gaps currently exist.

NAPAMI provides a reliable representation of the pregnant population over a substantial period in a standard clinical environment. Furthermore, recognizing that the mother-fetus system should be viewed as a whole, pregnancy can be seen as a continuously evolving entity that can be analyzed using time-varying approaches. Our dataset contains longitudinal data from both healthy and high-risk pregnancies.

Furthermore, the computed set of parameters within NAPAMI (detailed in the **Appendix** section) offers the possibility to examine the progression of fetal CTG features over time in all monitored patients without any limitations. This allows for continuous auditing and analysis, providing valuable insights. These features also enhance the analysis of FHR signal for both clinicians and researchers, regardless of the methodological approach they choose. The dataset serves as a tool to improve the diagnostic reliability of CTG. Moreover, the availability of a wide range of case histories grants clinicians access to a wealth of information, including cases involving maternal conditions such as diabetes that could influence fetal outcomes. Consequently, clinicians can gain a deeper understanding of these scenarios, identifying subtle patterns associated with specific conditions and making well-informed decisions even in challenging circumstances. By analyzing a vast set of CTG data, clinicians can identify patterns and correlations that may indicate potential risks or complications at an earlier stage of pregnancy, facilitating more accurate diagnoses and timely interventions.

This extensive collection of organized CTG data hence paves the way for new groundbreaking possibilities in the analysis of CTG signals.

Particularly, as illustrated in **Chapter 5**, we took advantage NAPAMI's wealth to investigate the understanding and analysis of FHR patterns, examining them in both the temporal and spectral domains.

Specifically, for what concerns temporal analysis, **Section 1 of Chapter 5** introduces a novel methodology based on a multivariate categorical Hidden Markov Model (HMM), designed for the automatic identification of fetal behavioral stages within FHR tracings. The obtained results suggest that the developed model can effectively identify quiet and activity states, since the best model's predictions matched an expert clinician's interpretations with a 90% agreement. One desirable feature of the proposed method is the ease of interpreting the obtained results, since the use of a small set of parameters, already extensively employed in cCTG, provides explainability and clear intrinsic meaning. Another appealing aspect is that the learning process is fully unsupervised. The model automatically identifies the clusters using the Baum-Welch algorithm, associating them with the "Active" and "Quiet" states after the fact. This approach eliminates reliance on expert annotations and ensures an objective evaluation based solely on the signal's intrinsic characteristics. However, it should be noted that the approach was tested with limited data due to time constraints, and future efforts will focus on expanding the testing data and the enrollment of more expert clinicians to enhance the reliability and generalizability of the results. It should also be noticed that the inclusion of the FMP signal in the analysis only marginally improves the model's performance, suggesting that the HMM is able to accurately identify fetal stages regardless from maternal perception of fetal movements. Future developments for this work include, e.g., leveraging knowledge of the actual state and the hysteresis of active/quiet state transition as a biomarker of development and fetal heath.

On the other hand, **Section 2 of Chapter 5** concerns FHR signals' spectral analysis and serves a dual purpose. First, we exploited the abundance of information contained in NAPAMI to compare various existing methods of spectral analysis; secondly, we propose a new approach for evaluating the spectrum of the phase-rectified signal averaging (PRSA), comparing the obtained results with those of the more classical spectral analysis methods.

Our study has shown that using the PRSA method before computing the spectra significantly improves the ability to distinguish between uncomplicated and high-risk pregnancies, compared to classical spectral analysis methods, while still providing physiologically interpretable results. On the other hand, the classical methods of spectral analysis did not show substantial differences.

While time-frequency analysis allows for localizing results in time and investigating fetal reactions to stimuli, the results suggest that the PRSA method presents a clear advantage over other methods when an aggregate measure is sought. This is attributed to its ability to reduce noise, capture non-phase synchronized oscillations, and its advantageous scaling behavior. One notable difference observed between control and high-risk groups is a relative increase in power in the mid-frequency range for the latter. This frequency band has been previously associated with fetal movements, but our study did not observe a significant increase in movement occurrence perceived by the mother in high-risk groups, suggesting other pathophysiological mechanisms may be involved. The physiological interpretation of this frequency band requires further study. Notably, a previous study found very little power in this band in uncomplicated low-risk pregnancies. The reduction in high-frequency power observed is consistent with a decrease in respiratory movements in pathological groups, which is supported by other studies demonstrating lower respiratory movement speed, power, intensity, and overall quality in intrauterine growth restriction (IUGR) fetuses. We did not observe significant differences in low-frequency power with the PRSA method, although minimal differences were identified with other methods. This may be attributed to the scaling behavior of the PRSA and specific parameter choices.

Instead, we did not find substantial differences when considering the acceleration-related PRSA curve instead of the deceleration-related one.

The conclusive Chapter of this thesis (**Chapter 6**) addresses the challenge of classifying FHR signals through the use of DL techniques. More specifically, it illustrates how we managed to exploit the abundance of CTG data offered by NAPAMI to develop and test the classification performance of a novel neural architecture. This neural network (NN) has been specifically designed to handle heterogeneous data, comprising both images and quantitative parameters representing the statistical characteristics of the FHR signal. The presented NN is composed of two branches: a Multi-Layer Perceptron (MLP) branch that receives a set of 15 regressors, and a Convolutional Neural Network (CNN) branch that is fed with a series of 64x64 images. These images are generated through various transformations applied to the pre-processed denoised FHR signal, including MTF, GADF, RP, and others.

To evaluate the superiority of the novel mixed-type architecture over the standalone MLP and CNN branches, we conducted a comparison of the three neural architectures in terms of overall classification accuracy. After optimizing each NN's hyperparameters, the MLP, CNN, and MLP-CNN architectures were trained and tested on a dataset of 14,000 instances, split into 80% for training and 20% for testing. The results demonstrated that the MLP-CNN network outperformed the other architectures, achieving an overall classification accuracy of 80.1% with the best set of hyperparameters. However, the main limitation of this method lies in its sensitivity, which is not yet fully satisfactory. Specifically, with the combined CNN+MLP model, the True Positive Rate (TPR) reached only 69%, corresponding to a 31% probability of incorrectly classifying an unhealthy subject. It should be noted that the unhealthy subjects included in the database used for this study represent a diverse group encompassing various types of diseases, such as intrauterine growth restriction (IUGR), metabolic alterations, fetal malformations, and maternal pathological conditions like diabetes.

The decision to include all these different disease categories within the unhealthy class was made to balance the number of healthy and unhealthy subjects and enable the utilization of Deep Learning techniques. However, this approach may decrease the performance of our neural model in terms of accuracy and specificity, as different pathologies can exhibit distinct patterns in FHR signals, thereby reducing classification capabilities and increasing the variability of FHR features. It is necessary to consider these findings as a starting point towards more intricate studies that examine distinct pathology classes separately, once the quantity of data for each pathology reaches an acceptable level for DL. Nonetheless, the results obtained from this study are promising, as they were achieved using a substantial amount of clinical data that closely represents the real population's variability. While this may result in lower classification performance compared to other existing works, it enhances the model's robustness and ability to generalize. Future advancements in this research will involve exploring alternative techniques for converting CTG signals into images to provide novel inputs for the CNN branch. Additionally, we will investigate the incorporation of a wider set of quantitative parameters to feed the MLP branch and explore other mixed-type neural architectures that include Recurrent Neural Networks (RNNs) or Temporal CNNs, thus allowing for an increased classification accuracy, enhancing the diagnostic capabilities of the CTG, and therefore, also maximizing its power as a medical decision-support tool.

Appendix.

This section is designed to furnish insights about each parameter introduced in the corpus.

Table 1A summarizes the full set of computed quantitative indexes and indicates, for each of them, the domain it belongs to (time, frequency, non-linear) and the duration of the signal excerpt considered for its computation.

Table 1A: summary of all computed parameters

Name	Domain	Window-length
Delta (Δ)	Time	1 min
Interval Index (II)	Time	1 min
Short Term Variability (STV)	Time	1 min
Long-Term Irregularity (LTI)	Time	3 min
Low Frequency (LF)	Frequency	3 min
Movement Frequency (MF)	Frequency	3 min
High Frequency (HF)	Frequency	3 min
Total Power (PWT)	Frequency	3 min
Approximate Entropy (ApEn)	Non-Linear	3 min
Sample Entropy (SampEn)	Non-Linear	3 min
Sample Asimmetry	Non-Linear	20 min
Binary Lempel-Ziv complexity (LZ2)	Non-Linear	3 min/20 min

Ternary Lempel-Ziv complexity (LZ3)	Non-Linear	3 min/20 min
Acceleration Capacity (AC)	Non-Linear	20min
Deceleration Capacity (DC)	Non-Linear	20min
Deceleration Reserve (DR)	Non-Linear	20min
Acceleration Phase Rectified Slope (APRS)	Non-Linear	20min
Deceleration Phase Rectified Slope (DPRS)	Non-Linear	20min
LFprsa	Non-Linear/Frequency	20min
MFprsa	Non-Linear/Frequency	20min
HFprsa	Non-Linear/Frequency	20min

The FHR signal is read by the 2CTG2 software 2 times per second, resulting in a “sampling frequency” of 2 Hz. The signal which is originally expressed in beats per minute (bpm) is transformed in ms by applying **Equation A1**.

$$\boldsymbol{RR}[\boldsymbol{ms}] = \frac{\mathbf{60000}}{\boldsymbol{FHR}[\boldsymbol{bpm}]} \qquad \textbf{(A1)}$$

Time domain parameters are computed on a down sampled version of RR [ms], called T[ms], obtained by averaging the signal in non-overlapping windows of 5 samples, resulting in a “sampling frequency” of 0.4 Hz. This procedure makes it impossible to quantify beat-to-beat variability but makes the analysis more robust to noise.

For consistency with the analysis conducted automatically by the 2CTG2 software, most parameters are computed on windows of 3 minutes. Some of the time-domain parameters are computed on sub-windows of 1 minute. Some non-linear parameters, which benefit from the use of longer time-series, are computed on windows of 20 minutes, which is the minimum signal-length available for all recordings.

Time Domain indices

Delta (Δ), Short Term Variability (STV), Interval Index (II) and Long-Term Irregularity (LTI) are traditional time-domain parameters that quantify variability. Δ, STV, II and LTI are computed on windows of 1 minute, while LTI on windows of 3 minutes.

Δ is simply computed as the difference between the maximum and minimum value of T[ms] in the window (**equation A2**):

$$\boldsymbol{\Delta} = \mathbf{max}(\boldsymbol{T}(\boldsymbol{i})) - \mathbf{min}(\boldsymbol{T}(\boldsymbol{i})), \quad \text{i=1,…,24} \qquad \textbf{(A2)}$$

STV estimates variability on a short-time scale, and is the mean of the absolute value of the difference between consecutive values of T[ms] (**equation A3**):

$$\boldsymbol{STV} = \frac{\sum_{\boldsymbol{i}=\mathbf{1}}^{\mathbf{23}} |\boldsymbol{T}(\boldsymbol{i}+\mathbf{1})-\boldsymbol{T}(\boldsymbol{i})|}{\mathbf{23}}, \quad \text{i=1,…,23} \qquad \textbf{(A3)}$$

II is defined as in **equation A4**:

$$\mathbf{II} = \frac{\boldsymbol{std}[|\boldsymbol{T}(\boldsymbol{i}+\mathbf{1})-\boldsymbol{T}(\boldsymbol{i})|]_{\boldsymbol{i}}}{\boldsymbol{STV}}, \quad \text{i=1,…,24} \qquad \textbf{(A4)}$$

LTI quantifies variability on a longer time scale, and is defined as in **equation A5**:

$$\boldsymbol{LTI} = \boldsymbol{IQR}(\sqrt{\boldsymbol{T}(\boldsymbol{i}+\mathbf{1})^{\mathbf{2}} + \boldsymbol{T}(\boldsymbol{i})^{\mathbf{2}}}), \quad \text{i=1,…,71} \qquad \textbf{(A5)}$$

Frequency Domain parameters

Frequency domain parameters are used to quantify the activity of the sympathetic and parasympathetic branches of the autonomic nervous systems. They are computed on windows of three minutes of RR[ms].

Spectra are estimated using autoregressive models of order 12, and the parameters are estimated using the Levinson-Durbin algorithm.

PWT (total signal power) is the integral of the entire spectrum and corresponds to the variance of the signal RR[ms]. LF, MF and HF are defined as the integral in their respective frequency bands and can be expressed in natural units (ms^2) or as percentage of PWT (e.g., $LF[\%] = \frac{\mathrm{LF[ms^2]}}{\mathrm{PWT[ms^2]}} \cdot 100$).

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

The Movement-Frequency (MF) band is defined between 0.15Hz and 0.5Hz and has been related to maternal breathing and fetal movements.

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

Non-Linear parameters

Non-linear parameters capture characteristics of the signal which cannot be properly quantified in time or frequency domain: regularity and complexity. Non-linear parameters are computed on RR[ms]

The **Approximate Entropy (ApEn)** is a family of statistics used to compute an approximate value of the signal entropy based on measuring the signal regularity: the less predictable the states are, the more complex the system is. In particular, “ApEn(m, r, N) is approximately equal to the negative average natural logarithm of the conditional probability that two sequences that are similar for m points remain similar, that is, within a tolerance r, at the next point”. N is the number of sampled of the input series.

The **Sample Entropy (SampEn)** is a variation over the approximate entropy that solves the bias effect of ApEn caused by self matches. In detail, “SampEn(m, r, N) is precisely the negative natural logarithm of the conditional probability that two sequences similar for m points remain similar at the next point, where self-matches are not included in calculating the probability”.

ApEn and SampEn are computed on windows of 3 minutes (N=360) with the following combinations of m and r: ApEn(1,0.15,360), ApEn(1,0.1,360), ApEn(1,0.2,360), SampEn(1,0.15,360), SampEn (1,0.1,360), SampEn (1,0.2,360). r is multiplied by the standard deviation of the signal in the 3-minutes window.

The **Lempel-Ziv Complexity** asses the minimum amount of information needed to define a string, i.e., it quantifies algorithmic complexity. Before computing the Lempel-Ziv complexity RR[ms] is transformed into a symbolic sequence of a low-dimensional alphabet.

Two metrics are computed, which differ in the strategy employed to code RR[ms] into a symbolic sequence. Both are computed on windows of three minutes. The values are normalized by the asymptotic behavior of a random string.

For the computation of the Binary Lempel-Ziv complexity (LZ2), RR[ms] is encoded as follows (**equation A6**):

$$\begin{cases} \mathbf{1}, & \boldsymbol{if\ RR[ms]_{n+1} > RR[ms]_n + p \cdot RR[ms]_n} \\ \mathbf{0}, & \boldsymbol{if\ RR[ms]_{n+1} \le RR[ms]_n + p \cdot RR[ms]_n} \end{cases} \tag{A6}$$

Where p indicates the smallest quantization level for a symbol alteration within the encoded strings. p is set to 0.02

On the other hand, the Ternary Lempel-Ziv complexity (LZ3), relies on the following encoding of RR[ms] (**equation A7**):

$$\begin{cases} \mathbf{1}, & \boldsymbol{if\ RR[ms]_{n+1} > RR[ms]_n + p \cdot RR[ms]_n} \\ \mathbf{0}, & \boldsymbol{if\ RR[ms]_{n+1} < RR[ms]_n - p \cdot RR[ms]_n} \\ \mathbf{2}, & \boldsymbol{if\ RR[ms]_n - p \cdot RR[ms]_n \le RR[ms]_{n+1}} \\ & \qquad \boldsymbol{\le RR[ms]_n + p \cdot RR[ms]_n} \end{cases} \tag{A7}$$

Where p is set to 0.01

The **Sample Asymmetry** describes changes in the shape of the histogram of RR intervals, allowing to separate the contribution of accelerations and decelerations. To compute the Sample Asymmetry the FHR [bpm] considered over windows of 20 minutes is detrended removing the moving median computed over 3 minutes (360 samples) and is computed as in **equation A8**:

$$\boldsymbol{Sample\ Asymmetry} = \frac{\sum_{i=1}^{n}(\widetilde{FHR}|\widetilde{FHR} > 0)^2}{\sum_{i=1}^{n}(\widetilde{FHR}|\widetilde{FHR} < 0)^2} \qquad \text{(A8)}$$

Where $\widetilde{FHR}$ indicates the detrended signal.

The **phase rectified signal averaging (PRSA)** is a non-linear signal processing technique that allows to identify quasi periodicities in non-stationary signals. It is a non-linear technique, but due to the importance it has assumed in the analysis of the FHR it has been dedicated a separate sub-chapter.

The PRSA method allows computing a compressed version of the signal, called PRSA curve, in which quasi periodicities are phase-synchronized while non-periodic components tend to be cancelled out.

To compute the deceleration-related PRSA curve (PRSAdec), the samples of the RR[ms] signal that satisfy the following condition are defined as deceleration anchor points (APdec) as in **equation A9**:

$$\mathbf{APdec} = \{\mathbf{t}\text{: } \frac{1}{T}\sum_{i=0}^{T-1} RR[t+i] > \frac{1}{T}\sum_{i=1}^{T} RR[t-i]\} \qquad \text{(A9)}$$

For each APdec, a window of length 2L is defined taking the values of the original signal that go from APdec-L to L to APdec+L-1.

L is set to 40 and T is set to 1.

A similar procedure is performed to obtain the acceleration-related PRSA curve (PRSAacc), inverting the direction of the inequality.

The **Deceleration Capacity (DC)** is defined as in **equation A10**:

$$\mathbf{DC} = \frac{1}{2s}\sum_{i=1}^{s} PRSAdec[L+i] - \frac{1}{2s}\sum_{i=0}^{s-1} PRSAdec[L-i] \qquad \text{(A10)}$$

s is the scale and is set to 1. DC quantifies the deceleration-related oscillations, and it has been linked to the activity of the parasympathetic nervous system.

The **Acceleration Capacity (AC)** is defined in a similar way, using PRSAacc instead of PRSAdec and has been hypothesized to quantify sympathetic activity.

The deceleration reserve is simply defined as in **equation A11**:

$$\boldsymbol{DR = DC + AC} \qquad \textbf{(A11)}$$

and quantifies asymmetries in the RR signal, since AC and DC have the same expected value for stochastic gaussian processes.

LFprsa, MFprsa and HFprsa were recently proposed to quantify the oscillations in PRSAdec, thus exploiting the abilities of the PRSA method to detect quasi oscillations and remove noise to improve spectral analysis of RR[ms]
The scalogram of PRSAdec is computed as in **equation A12**:

$$\boldsymbol{X_w^{PRSA}(s,p) = \sum_{k=-L}^{L-1} x_k^{PRSA} \cdot \frac{w[(k-p)/s]}{s}} \qquad \textbf{(A12)}$$

Where w is the analytic Morse wavelet with γ equal to 3 and time-bandwidth product equal to 60. The spectrogram is obtained as the square of the wavelet coefficients and is evaluated at k=0, thus obtaining a single spectrum, i.e. PRSA_Spt. LFprsa, MFprsa and HFprsa are obtained as the integral of PRSA_Spt between the frequency bands defined previously normalized by the total power of PRSA_Spt.

For the computation of **Acceleration Phase Rectified Slope** (**APRS**) and **Deceleration Phase Rectified Slope** (**DPRS**), PRSAacc and PRSAdec are computed as previously explained, but using FHR[bpm]. As a consequence, APdec is defined as follows (**equation A13**):

$$\mathbf{APdec} = \{\mathbf{t}\colon \frac{1}{T}\sum_{i=0}^{T-1} FHR[t+i] < \frac{1}{T}\sum_{i=1}^{T} FHR[t-i]\} \qquad \text{(A13)}$$

DPRS is defined as the slope of PRSAdec at the anchor point (**equation A14**):

$$\mathbf{DPRS} = \frac{\partial PRSAdec}{\partial i},\ \ i = \mathbf{APdec} \qquad \text{(A14)}$$

APRS is defined in the same way but on PRSAacc.