

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Analysis of the Association of Cardiac and Respiratory Signals

Miguel Dias



Mestrado em Engenharia Biomédica

Supervisor: Teresa Sarmento Henriques

Second Supervisor: Luísa Castro Guedes

August 8, 2023

Analysis of the Association of Cardiac and Respiratory Signals

Miguel Dias

Mestrado em Engenharia Biomédica

August 8, 2023

Abstract

Neonatal sepsis is a generalized infection that affects newborns with a high mortality rate. This disease manifests in different symptoms, sometimes making it difficult to diagnose. However, one of the best ways to combat the disease is to make a rapid diagnosis to anticipate the administration of antibiotics. Prematurity has been growing in developed countries and is a risk factor for the newborn to contract the infection. Therefore, the incidence of neonatal sepsis is increasing. There are different diagnostic methods for neonatal sepsis, including blood cultures. However, these methods are time-consuming and invasive. Artificial intelligence has recently been considered a possible alternative to making an early diagnosis. Several studies and projects, such as the Digi-NewB project funded by the European Union, have been investigating the usefulness of artificial intelligence in diagnosing infection with positive results. Heart and respiratory rates are two physiological signals that can indicate the presence of an infection in the body due to the symptoms associated with the disease. The present Master's Dissertation aims to develop a classifier for neonatal sepsis based on the changes recorded in cardiac and respiratory signals. The data studied in this dissertation comes from the Digi-newB project.

Entropy and compression measures were used, in order to understand how the complexity of the signal shifts thru time, with the objective of finding significant differences in the shifts recorded for control and for infected subjects. After computing the entropy and compression measures, several steps were performed to understand the dynamics of the signal. Boxplots were used to see how all the subjects measures obtained evolved thru time. It was noticeable, for certain measures (k2 kolmogorov entropy for the respiratory signal, sample entropy for the heart signal and cross permutation entropy for the coupling of both signals), that there were evident shifts in the signal's complexity, which could indicate the presence of the infection. Then, the distribution of the measures for the control and infected subjects was plotted. Some measures (fuzzy entropy for respiratory rate and compression rate for gzip in the heart rate) indicated differences in the complexity between an infected and a not infected subject. Lastly, the classification was performed, and for the univariate measures, the best performance (accuracy of 80%) was obtained for the heart signal using random forests.

The results highlighted the importance of studying the evolution of physiological signals associated with the disease. It could have a great impact when developing a non-invasive way of detecting neonatal sepsis. Diagnostic solutions like these could help decrease the disease's mortality rate and antibiotic resistance, often caused by the preventive medication administered to premature babies.

Keywords: complexity measures, heart rate, respiratory rate, classification, neonatal sepsis

Resumo

A sepsis neonatal é uma infecção generalizada que afecta os recém-nascidos com uma elevada taxa de mortalidade. Esta doença manifesta-se por diferentes sintomas, tornando por vezes difícil o seu diagnóstico. No entanto, uma das melhores formas de combater a doença é fazer um diagnóstico rápido para antecipar a administração de antibióticos. A prematuridade tem vindo a aumentar nos países desenvolvidos e é um factor de risco, para que o recém-nascido contraia a infecção. Por conseguinte, a incidência de sépsis neonatal está a aumentar. Existem diferentes métodos de diagnóstico para a sépsis neonatal, incluindo as culturas de sangue. No entanto, estes métodos são demorados e invasivos. Recentemente, a inteligência artificial tem sido considerada uma alternativa possível para efetuar o diagnóstico precoce. Vários estudos e projectos, como o projeto Digi-NewB financiado pela União Europeia, têm investigado a utilidade da inteligência artificial no diagnóstico de infecções com resultados positivos. A frequência cardíaca e respiratória são dois sinais fisiológicos que podem indicar a presença de uma infecção no organismo devido aos sintomas associados à doença. A presente Dissertação de Mestrado tem como objectivo desenvolver um classificador para a sépsis neonatal com base nas alterações registadas nos sinais cardíacos e respiratórios. Os dados estudados nesta dissertação são provenientes do projecto Digi-newB.

Foram utilizadas medidas de entropia e compressão, de forma a perceber como a complexidade do sinal se altera ao longo do tempo, com o objectivo de encontrar diferenças significativas nas alterações registadas para os sujeitos de controlo e para os sujeitos infectados. Após o cálculo das medidas de entropia e de compressão, foram efectuados vários passos para compreender a dinâmica do sinal. Foram utilizados *boxplots* para verificar a evolução temporal de todas as medidas obtidas pelos sujeitos. Verificou-se, para certas medidas (*k2 kolmogorov entropy* para o sinal respiratório, *sample entropy* para o sinal cardíaco e *cross permutation entropy* para o acoplamento de ambos os sinais), que havia mudanças evidentes na complexidade do sinal, o que poderia indicar a presença da infecção. Em seguida, foi traçada a distribuição das medidas para os sujeitos de controlo e para os sujeitos infectados. Algumas medidas (*fuzzy entropy* para a frequência respiratória e *compression rate for gzip* na frequência cardíaca) indicaram diferenças na complexidade entre um sujeito infectado e um não infectado. Por fim, foi efetuada a classificação e, para as medidas univariadas, o melhor desempenho (precisão de 80%) foi obtido para o sinal cardíaco utilizando *random forests*.

Os resultados evidenciaram a importância de estudar a evolução dos sinais fisiológicos associados à doença. Estes estudos poderão ter um grande impacto no desenvolvimento de uma forma não invasiva de detecção da sépsis neonatal. Soluções de diagnóstico como estas poderiam ajudar a diminuir a taxa de mortalidade da doença e a resistência aos antibióticos, muitas vezes causada pela medicação preventiva administrada aos bebés prematuros.

Keywords: medidas de complexidade, frequência cardíaca, frequência respiratória, classificação, sépsis neonatal

Acknowledgements

First, I would like to thank to Professors Teresa and Luisa, for accepting me in this project, and for always guiding me and supporting me.

I would also like to thank to all my friends and colleague, for making my academic experience better.

Finally, I would like to thanks my family, specially my brother, my grandfather, my grandmother and my mother, for always being supportive and helping me in everything I did. To my mother, for always being present and for transmitting me the mural values that made me the person I am today.

“Information is the resolution of uncertainty”

Claude Shannon

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Context	1
1.3	Goals	2
1.4	Outline	2
2	Neonatal Sepsis	3
2.1	Burden of Neonatal Sepsis	3
2.2	Clinical Signals and Symptoms of Neonatal Sepsis	4
2.3	Causative Agents	4
2.4	Diagnosis	4
2.4.1	Conventional Diagnosis	4
2.4.2	Culture-independent Diagnosis	5
2.4.3	Risk-based Approach Calculator	5
2.5	Treatment and Prevention	5
3	Heart and Respiratory Rate	7
3.1	Respiratory Rate	7
3.1.1	Neonatal Respiratory Rate	8
3.1.2	Recording Respiratory Rate	9
3.2	Heart Rate	9
3.2.1	Heart Rate Variability and Heart Diseases in Newborns	9
3.2.2	Recording ECG and Heart Rate	9
3.3	DIGI-NEWB	10
4	Complexity Measures	13
4.1	Entropy Methods and Mutual Information	13
4.1.1	Approximate and Sample entropy	13
4.1.2	Bubble and Permutation entropy	14
4.1.3	Attention entropy	15
4.1.4	Fuzzy Entropy	16
4.1.5	Kolmogorov Entropy and K2 Kolmogorov Entropy	17
4.1.6	Distribution Entropy	17
4.1.7	Conditional Entropy	18
4.1.8	Cross-entropies	18
4.1.9	Multiscale entropies	19
4.1.10	Mutual Information	19
4.2	Compressors and Normalized Compression Distance	19

4.2.1	Compressors	19
4.2.2	Normalized Compression Distance	20
4.3	Programming	21
5	Classifiers	23
5.1	Logistic Regression Classification	23
5.2	Tree Based Methods	25
5.2.1	Decision Trees	25
5.2.2	Random Forest Classifier	26
5.3	Programming	27
6	Results and Discussion	29
6.1	Preliminary results	29
6.2	Data	32
6.2.1	Sample Analysis	32
6.2.2	Signal Representation	33
6.3	Resampling	33
6.4	Complexity Measures Selection	37
6.5	Complexity Measures Analysis for Infected Subjects	40
6.6	Control and Infected Subjects	42
6.6.1	Classification	43
7	Conclusion	47
A	Boxplot	49
A.1	Boxplots for the Univariate Measures for Respiratory signal	49
A.2	Boxplots for the Univariate measures for Heart signal	53
A.3	Boxplots for the Bivariate measures	57
B	Control and Infected Subjects	59
B.1	Control and Infected Subjects for the Heart Rate signal	59
B.2	Control and Infected Subjects for the Respiratory Rate signal	62
C	Mann-Whitney test p-values	65
	References	73

List of Figures

3.1	Graphic representation of the respiratory cycle. Figure adapted from [75].	7
3.2	Schematic example of a part of an electrocardiogram to illustrate the RR interval, QRS complex and all the waves of the ECG. Adapted from [39].	10
5.1	Sigmoide function. Figure adapted from [91].	24
5.2	example of a decision tree classification. Figure adapted from [60].	26
5.3	Random Forest Classifier schematics. Figure adapted from [54].	27
6.1	Correlation HeatMap for Cross-Sample entropy (XSampEn), Multiscale Sample entropy (MsSamp_hr, CIS_hr), Multiscale Approximate entropy (MsAp_hr, CiA_hr), Multiscale Bubble entropy (MsB_hr, CiB_hr) and Multiscale Attention entropy (MsAt_hr, CiAt_hr) for Heart Rate signal	30
6.2	Correlation HeatMap for Cross-Sample entropy (XSampEn), Multiscale Sample entropy (MsSamp_hr, CIS_hr), Multiscale Approximate entropy (MsAp_hr, CiA_hr), Multiscale Bubble entropy (MsB_hr, CiB_hr) and Multiscale Attention entropy (MsAt_hr, CiAt_hr) for Respiratory Rate signal	31
6.3	First 10 minutes of Heart Rate signal for an infected individual.	33
6.4	First 10 minutes of Respiratory signal for an infected individual.	33
6.5	[Original heart rate signal (blue) VS resampled heart rate signal (red) for 1sample/s.] The horizontal axis represents the time (s) until infection. The vertical axis represents the signal	34
6.6	[Original respiratory rate signal (blue) VS resampled respiratory rate signal (red) for 1sample/1.5s. The horizontal axis represents the time (s) until infection.] The vertical axis represents the signal.	34
6.7	Correlation Heatmap for Heart Rate. Each correlation value, is representative of the similarity between the pairs of combinations of the files compared (previous list).	36
6.8	Correlation Heatmap for Respiratory Rate. Each correlation value, is representative of the similarity between the pairs of combinations of the files compared (previous list).	37
6.9	Correlation Heatmaps for the measures for Respiratory Rate	38
6.10	Correlation Heatmaps for the measures for Heart Rate	39
6.11	Complexity measures of heart rate and respiratory signal for the Infected Subjects.	40
6.12	K2 kolmogorov entropy boxplots for 30 minute windows until infection, for the respiratory signal.	41
6.13	Sample entropy boxplots for 30 minute windows until infection, for the heart rate signal.	41

6.14	Cross permutation entropy boxplots for 30 minute windows until infection, for the coupling of both signals.	42
6.15	Control (grey) vs Sepsis (red) values for the respiratory signal (1 sample/0.5s), for Fuzzy entropy.	42
6.16	Control (grey) vs Sepsis (red) values for the heart rate signal (1 sample/2.5s), for the compression rate for gzip.	43

List of Tables

3.1	Normal respiratory rate by age. Adapted from [86].	8
6.1	Median [minimum; maximum] values for the entropy index for the heart rate and respiration time series: Cross-sample entropy (XSampEn), Multiscale sample entropy (MsSamp, CiS), Multiscale approximate entropy (MsAp, CiA), Multiscale bubble entropy (MsB, CiB) and Multiscale attention entropy (MsAt, CiAt)	30
6.2	Clinical information of the 33 neonates.	32
6.3	Accuracy of the Classifiers. 23 subjects for training and 10 subjects for testing. .	44
6.4	Confidence Interval for the Accuracy of the Classifiers.	45
C.1	P-values for the measures of the original Heart Rate signal.	66
C.2	P-values for the measures of the Heart Rate signal, resampled for 2Hz.	67
C.3	P-values for the measures of the Heart Rate signal, resampled for 0.4Hz.	68
C.4	P-values for the measures of the original Respiratory Rate signal.	69
C.5	P-values for the measures of the Respiratory Rate signal, resampled for 2Hz. . . .	70
C.6	P-values for the measures of the Respiratory Rate signal, resampled for 0.4Hz. . .	71

Abbreviations

ANS	Autonomic Nervous System
BPM	Beats per Minute
CRP	C-reactive protein
ECG	Electrocardiogram
EU	European Union
GBS	Group B Streptococcus
HRM	Heart Rate Monitor
LZW	Lempel-Ziv Markov Chain Algorithm
NCD	Normalized Compression Distance
NICU	Neonatal Intensive Care Units
PCR	Polymerase Chain Reaction
PPG	Photoplethysmography
RESP	Respiratory Signal
RR	R-R Interval
WBC	White Blood Cells
WHO	World Health Organization

Chapter 1

Introduction

1.1 Motivation

Neonatal sepsis is a generalized infection that affects around 3,000,000 newborns annually [33]. According to the World Health Organization (WHO), neonatal sepsis significantly impacts neonatal morbidity, especially in low or middle-income countries. The WHO also states that sepsis increasingly affects hospitalized patients in high-income countries [71], representing 11% of neonatal and pediatric mortality [83].

Neonatal sepsis requires the medical community's attention once it significantly impacts the development of premature and neuronal problems, especially in vulnerable newborns [33, 71]. The best way to reduce the mortality associated with the disease is to diagnose early and administer a suited treatment to the newborn. However, the current conventional diagnosis methods of neonatal sepsis are not always accurate and cause delays in the diagnosis. It is known that several clinical factors impact the development of neonatal sepsis, such as prematurity and low-weight [33]. For instance, premature babies have a higher probability of suffering from neonatal sepsis, and once the number of premature births tends to increase in the modern world, the study of the disease becomes critical [28]. On the other hand, the symptoms associated with neonatal sepsis are usually not specific, such as fever, poor crying, and respiratory distress. Therefore, the detection of neonatal sepsis still requires a lot of improvements [33].

1.2 Context

Some changes in the newborn physiological signs occur as a response to the neonatal sepsis infection. Tracking vital signs after birth, when there is a risk associated with contracting sepsis, is crucial when trying to perform an early diagnosis [73].

Recently, new methods have been proposed to detect neonatal sepsis, for example, using complexity measures like entropy and compression indices of the heart rate [66].

The literature states that studying and researching neonatal sepsis using artificial intelligence could help improve the accuracy of antibiotics administration for treating this disease, helping

to reduce antibiotic resistance [45]. A detailed analysis of the clinical state of the patient, using predictive models, could also be helpful in predicting the infection [74].

1.3 Goals

The main goal of this project is to explore and study the dynamics of two physiological signals: heart and respiratory rate, in neonatal sepsis detection. This study uses complexity measures like entropy, mutual information, and normalized compression distance. Finally, a neonatal sepsis classifier will be developed considering the main risk factors and the heart and respiratory rate measures. The data for this study belongs to an international clinical research project.

1.4 Outline

This report is divided into nine main chapters. In Chapter 2, neonatal sepsis is defined, as well as its major symptoms. The most common methods of diagnosis are discussed, as well as how the disease is currently treated. This project studies two physiological signals, heart and respiratory rate. They are described in chapter number 3, as well as how these signals are usually collected. Then, in Chapter 4, the complexity measures that will be used to analyze the fluctuations in the signals will be explained, with their advantages and disadvantages pointed out. After describing the complexity measures used, the following chapter, number 5, will contain information about the classifiers that can be used when developing the neonatal sepsis predictor. Chapter 6 will summarise the main functions used in Python for the analysis. In chapter 7, the results obtained in the PTL (Projeto e Técnicas Laboratoriais) curricular unit will be briefly presented. Then, chapter 8 will bring a thorough description on the methods implemented to perform the study, and all the results obtained for each method. They will be in chronological order, and will explain all the decisions made throughout the project, until obtaining the final clasification results. Finally, the results obtained will be discussed in chapter number 9, and future improvements will be proposed.

Chapter 2

Neonatal Sepsis

Neonatal sepsis is highly deadly for newborns, and despite the existence of treatment, this disease represents the most common cause of neonatal death. The symptoms associated with sepsis are usually not specific and it is difficult to perform an early diagnosis. The current methods of diagnosis lack specificity and usually more than one method is required, in order to effectively detect neonatal sepsis [48]. An analysis and review performed by Fleischmann et al. to several databases regarding the period from January 1979-May 2019, estimated an incidence of around 2824 cases per 100000 live births, and, of those cases, approximately 17.6% resulted in death. [23]. There are 3 different classifications for neonatal sepsis, depending on when the clinical manifestations happen:

- If they occur in the first 3 days of life (<72hours), it is considered early-onset neonatal sepsis;
- Late-onset, when neonatal sepsis is diagnosed between 4 and 30 days of life;
- Very late-onset sepsis happens when the diagnosis only occurs after the 30th day of life.

However, these definitions may vary, as some researchers consider early-onset until the 7th day of life and late-onset from the 7th to the 30th day [53]. This disease has a wide range of clinical manifestations, ranging from sub-clinical infection to severe focal or systemic disease [53].

2.1 Burden of Neonatal Sepsis

As mentioned, neonatal sepsis is the most significant cause of neonatal deaths, representing more than half of them. The best way to treat this disease and avoid its mortality is by performing an early diagnosis and treating it right away. When babies with this condition are placed in the Neonatal Intensive Care Units (NICU), the risk of contracting hospital-acquired infections increases significantly [48].

2.2 Clinical Signals and Symptoms of Neonatal Sepsis

The treatment of the disease will be more effective with an early diagnosis. Therefore, it is essential to identify symptoms and signals of possible infection as early as possible. However, neonatal sepsis often presents nonspecific signals. This infection is a great threat to newborn babies, so recognizing the presence of neonatal sepsis on the organism is crucial. One way of detecting neonatal sepsis is by focusing on the clinical history of the baby and on the signs and symptoms that may be presented. As aforementioned, those signs and symptoms can vary significantly. Initial symptoms of the condition include hypothermia, fever, poor crying, respiratory distress, lethargy, refusal to suck, fast breathing, hypotonia, and tachycardia. On a more advanced stage, the signs and symptoms can include respiratory failure, pulmonary hypertension, cardiac failure, shock, thrombosis, and renal failure, among others [48, 70].

2.3 Causative Agents

Neonatal sepsis can occur during or before birth, causing early onset neonatal sepsis. In this situation, the condition is caused by the contamination of the amniotic fluid, placenta cervix or vaginal canal, by colonizers of the maternal genitourinary track. For the early onset neonatal sepsis, the infection occurs when the amniotic membran ruptures. There is a wide range of bacteria that can be in the origin of an early onset neonatal sepsis, like *Streptococcus*, *L. monocytogenes*, *E. faecalis*, *E. faecium*, group D *Streptococci*, -hemolytic *Streptococci* and *Staphylococci*, *S. pneumoniae* and *H. influenzae type B*.

Late onset neonatal sepsis is mostly caused by two bacteria, *E. Coli* and *Staphylococci species*, and it is usually associated with a low birth weight, contact with contaminated equipment in the hospital environment, assisted ventilation, endo-tracheal intubation, surgery, or use of intravascular catheters. [27].

2.4 Diagnosis

2.4.1 Conventional Diagnosis

All possible infections must be tested while the mother is pregnant to allow adequate pre-birth care and decrease the incidence of the disease [48].

Usually, in order to detect neonatal sepsis through conventional diagnosis, the causative agent is isolated from a sample of an expected sterile body site, like blood, cerebrospinal fluid, and urine, for example. Some neonatal sepsis causative agents are only detected in cerebrospinal fluid, so in some situations, a lumbar puncture should be included in the procedure for the assessment of the disease for symptomatic neonates [70]. Blood culture diagnosis is a commonly used diagnostic method. Although it is vastly used, it sometimes lacks on sensitivity, possibly due to neonatal blood contamination. It also has a delay on the results from 24h up to 72h [80]. Urine samples are

also considered for the diagnosis of neonatal sepsis, however urinary tract infections only occur after 72h of living, so it is more commonly used for the detection of late onset neonatal sepsis [70].

2.4.2 Culture-independent Diagnosis

Two commonly used culture-independent diagnostics methods are Polymerase Chain Reaction (PCR) and white blood cells (WBC) counts. PCR has an high sensitivity. Besides having high sensitivity, the results of the diagnose are usually obtained quickly, and the samples required for this method are usually of a small volume. One of the biggest disadvantages of the PCR test, is its difficulty to distinguish active infections from recently infections that no longer affect the system. It also has the possibility of detecting contaminants [70]. Although its sensitivity is low, white blood cells counts are also used for the diagnosis of neonatal sepsis. The main advantage of this method, is the negative predictive value of it, because in this case, it means the blood culture would not be infected [70].

Other culture-independent diagnosis can be used, like C-reactive protein (CRP), proteomic markers in amniotic fluid and cell surface markers, for example [70].

2.4.3 Risk-based Approach Calculator

Currently, there have been developed some online neonatal sepsis risk calculators. Marleen et al. [41] developed one of those calculators, considering five maternal risk factors: gestational age, highest maternal temperature during labor, duration of membrane rupture during delivery, group B Streptococcus (GBS) status at birth and, finally, intrapartum antibiotic treatment [41]. Alongside monitoring the newborn's clinical situation, the calculator allows a recommendation regarding antibiotic treatment. This recommendation is based on three levels for the risk of developing neonatal sepsis. This tool helps prevent the inappropriate usage of antibiotics treatment in the early life of the newborn.

Recently, Escobar et al. [21] conducted a study where a sepsis calculator was proposed. This calculator was based on a dataset information regarding objective maternal factors, demographics, specific clinical milestones, and vital signs during the first 24 hours after birth. The data was used to develop a risk classification method, using a combination of recursive partitioning and logistic regression [21]. The authors also compared the results and recommendations obtained by the calculator with the records of the medical team. This study's findings also implied that using a sepsis calculator could reduce the improper use of antibiotics when treating newborn babies [8, 88].

2.5 Treatment and Prevention

Neonatal sepsis treatment is typically composed of antibiotics and supportive care. Nevertheless, antibiotics still take between 12-14 hours to make a notable effect, so supportive care is critical in

the early stage of treatment. Depending on the suspected bacteria that causes the infection, the decision on which antibiotics to administer is made. A combination of ampicillin or penicillin with gentamicin is commonly used in situations where resistant strains are less likely to be causing the infection, like in community-acquired infection [48]. A combination of ampicillin and cloxacillin is normally used for hospital infections where resistance strains are more likely to happen [48].

Chapter 3

Heart and Respiratory Rate

3.1 Respiratory Rate

Understanding what respiratory rate means requires the understanding of a respiration/breathing cycle. A respiratory cycle comprises an inhalation followed by an exhalation, where the oxygen goes into the body while the carbon dioxide is excreted. Based on that, the respiratory rate consists of the number of breaths or respiratory cycles per period of time. It is usually represented in breaths per minute [9].

Figure 3.1 represents a respiratory cycle. SUM signal is obtained by the algebraic sum of the antero-posterior diameter of the rib cage (RC) with the antero-posterior diameter of abdominal excursions (Ab), and in the SUM signal, the tidal volume (V_t) is represented [76]. Tidal volume is a very important parameter for an healthy circulation on the lungs, and represents the maximum volume of air going in or out the lungs each respiratory cycle [31]. T_i is the inspiration time, and T_e is the expiration time [76].

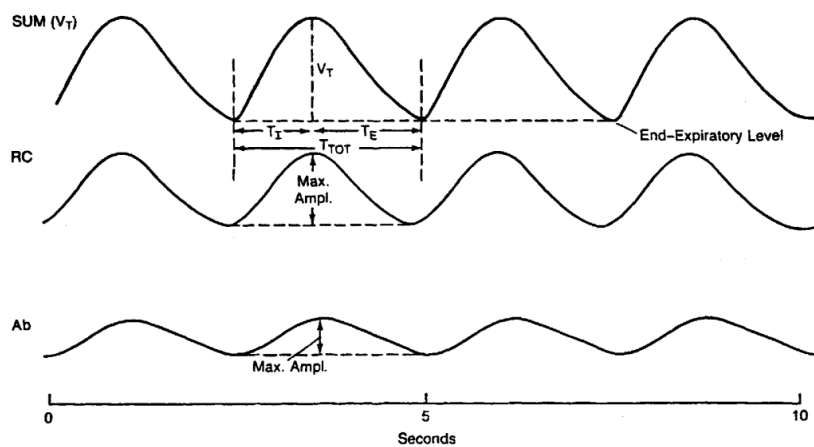


Figure 3.1: Graphic representation of the respiratory cycle. Figure adapted from [75].

The respiratory rate has been proven to be an essential signal, not only when diagnosing respiratory diseases but also when dealing with diseases that cause the human body to have a metabolic response, such as sepsis [16].

Respiratory rate is a measure that helps evaluate the level of metabolic demand generated by a particular disease. Generally, higher metabolic demand means a higher illness severity [86].

It is scientifically proven that variations in the respiratory rate are useful for diagnosing and managing respiratory and non-respiratory conditions. When newborn babies are hospitalized in intensive care units, respiratory rate is often not considered, although there is scientific evidence that this measure could be the most beneficial one [9]. That is why the respiratory rate is a signal used in some existing illness severity classification [86].

Different time counters can be used to estimate the respiratory rate, but it is usually presented in breaths per minute [86]. The World Health Organisation recommends that the respiratory rate is measured for counters of 60s [52].

The normal respiratory rate can vary for different ages, and table 3.1 presents the average resting respiratory rates sorted by age [86]:

Table 3.1: Normal respiratory rate by age. Adapted from [86].

Age	Normal Respiratory Rate
Birth to 6 weeks	30-60 breaths/minute
6 months	25-45 breaths/minute
3 years	20-30 breaths/minute
6 years	18-25 breaths/minute
10 years	15-20 breaths/minute
Adult	12-20 breaths/minute

3.1.1 Neonatal Respiratory Rate

As previously said, respiratory rate is a valuable vital signal. However, its proven importance, measuring respiratory rate represents a challenge, as it is affected by factors such as fever and agitation. This becomes an even greater challenge for neonatal respiratory rate, because newborn babies present an irregular breathing pattern until they mature their respiratory control mechanisms. Although there is a great challenge, there are some directories defined. For example, according to the Integrated Management of Newborn and Childhood Illness booklet, a respiratory rate above 60 breaths/minute means the child possibly has several bacterial infections or severe diseases, for the first two months after the birth. Although it could result in observer error, the strategy usually used is to measure the breath rate for shorter periods and then multiply it by a factor to estimate a per-minute rate. A study conducted in 2022 experimented with measuring respiratory rate for short periods, and concluded that measuring for short periods could have a small effect because of

inpatient variability. It also concluded that the measures for longer periods were not ideal. So, the best option, when possible, was the continuous monitoring of respiratory rate [52].

3.1.2 Recording Respiratory Rate

When a patient is not in a critical care unit, the standard way to measure respiratory rate, is by performing a manual count of the chest movements. Despite that, it is stated in the literature that this is an inefficient and inaccurate method and takes a lot of time. This way, an automated way of measuring this physiological signal is required, in order to continuously control the respiratory rate of patients in critical care units. Although still not commonly used in clinical practice, there have been developed respiratory rate estimators, based on electrocardiograms (ECG) and photoplethysmography (PPG). The advantages of this estimators are that the signals used can be continuously monitored, for example, using wearable sensors, and are more accurate than the traditional way. This would also allow the attainment of time-series, facilitating the analysis of the signal [57].

3.2 Heart Rate

The heart rate is a physiological signal used to diagnose heart conditions. This signal represents the heartbeat's pace, and it calculates the number of heartbeats per time unit, usually represented in beats per minute (bpm). Although it could differ, depending on the age or sex of a particular individual, the heart rate values are expected to be 60 to 100 beats per minute, for a healthy adult. For example, depending on the heart rate, an ECG signal can have three different classifications [62]:

- Normal - when the heart rate value is in between 60 and 100 bpm;
- Bradycardia - if the heart rate is below 60 bpm;
- Tachycardia – when the heart rate value exceeds 100 bpm.

3.2.1 Heart Rate Variability and Heart Diseases in Newborns

It is possible to measure the sympathetic and parasympathetic function of the autonomic nervous system (ANS) using the heart rate variability measure [17]. This way, it is possible to identify the adjustments made by the cardiovascular control mechanism, the ANS, to unexpected stimulus, as the heart rate variability is a representation of the state of that system [61]. It is described in the literature that heart rate variability can be a great indicator of certain medical conditions in newborns, like severe neonatal respiratory disease syndrome. Respiratory distress is another example, as aggravated respiratory distress is associated with increased heart rate variability [17].

3.2.2 Recording ECG and Heart Rate

Although some new heart rate monitoring technologies are coming up, like the heart rate monitor (HRM), the ECG is still the most reliable way to measure the heart rate. In the ECG, each heartbeat is represented by the RR interval, the time elapsed between two successive R waves [18].

Figure 3.2 shows an example of an ECG signal, and the P, QRS, T, and U waves. The P wave is usually of small voltage and represents the depolarization of the atria, prior to its contraction. The QRS complex is caused by ventricles polarization, also preceding their contraction. After this complex, comes the T wave, which represents ventricle repolarization. Finally, it is also shown the RR interval that equals a heartbeat. The horizontal axis represents the time, and the vertical axis represents the voltage [39].

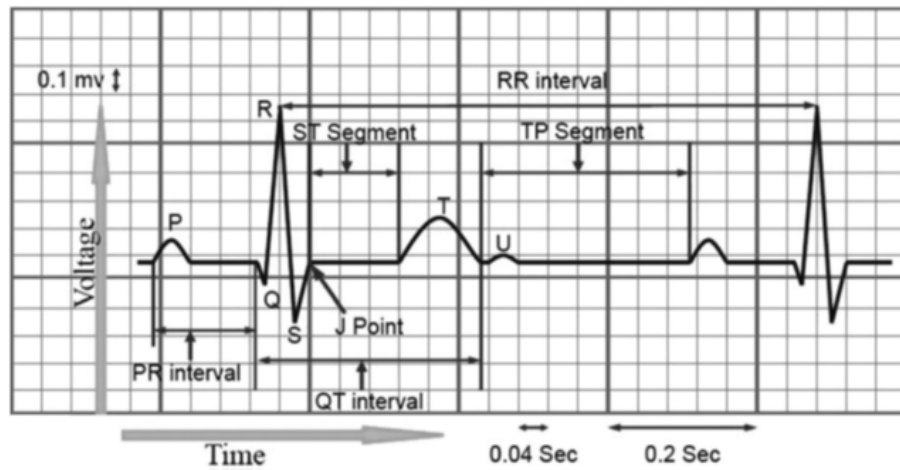


Figure 3.2: Schematic example of a part of an electrocardiogram to illustrate the RR interval, QRS complex and all the waves of the ECG. Adapted from [39].

3.3 DIGI-NEWB

The project Digi-NewB: monitoring system for sepsis risk analysis observational cohort is an European (EU) project regarding the early diagnosis of neonatal sepsis. It was supported by the EU program for Research and Innovation Horizon 2020 (NCT02863978, Horizon 2020 European project GA-689260). The Digi-NewB collected data regarding physiological signals, video, and clinical observations of the preterm infants. This project joined different Universities in Europe and aimed to develop decision support systems that could help clinicians in decision-making processes, aiding the early detection of sepsis. It was a large clinical study that resulted in several scientific publications regarding the topic [82].

One of the publications associated to the project regarded a study conducted in 2018 with 705 neonates by the University of Tampere and the Rennes University Hospital in France. The project aimed to develop an Artificial Intelligence algorithm to assess neonatal sepsis development risk. In this project, physiological signals, video, and sound information were recorded and then processed by an algorithm developed by three universities: the University of Porto, the University of Tampere, and the University of Rennes. Besides those features, the clinical characteristics were also included in the decision support system. Although the first tests did not include all the expected features, and the tests should be performed for more extensive datasets in order to get

clinical confirmation, the results showed by the system suggested that it could be a helpful tool in diagnosing sepsis [79].

As a result of the Digi-NewB project, a paper was published in 2022 to propose a new method of diagnosing neonatal sepsis, considering heart rate variability measures and neonatal gestational age. The heart rate variability was evaluated using the multiscale sample entropy and the multiscale compression methods. The RR traces were analyzed for 30 minutes segments, and five scales were used for the multiscale methods. The data was from 141 infants hospitalized in the neonatal units of 6 different University Hospitals in France. The control group had 118 infants, and complexity measures were computed for that group: sample entropy, multiscale entropy, compression ratio and multiscale compression curve. After computing the complexity measures for the control group, the same complexity measures were computed for 23 sepsis infected infants, and the values were compared. This study concluded that, as previously stated in the literature, the combined use of multiscale entropy and compressors was an efficient way to predict neonatal sepsis in a non-invasive way [65].

Chapter 4

Complexity Measures

4.1 Entropy Methods and Mutual Information

Entropy methods are a known way of measuring the uncertainty of a signal. The way entropy measures the uncertainty of the distribution of a signal is by recognizing pattern repetitions. The higher the entropy, the higher the level of complexity of the signal samples. A lower complexity would mean lower entropy. Entropy methods are widely used to reduce the uncertainty and subjectivity that is sometimes faced in the diagnosis process, for example [59].

4.1.1 Approximate and Sample entropy

Informally, given N points, the family of statistics $ApEn(m, r, N)$ 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. Thus a low value of $ApEn$ reflects a high degree of regularity. Importantly, the $ApEn$ algorithm counts each sequence as matching itself, a practice carried over to avoid the occurrence of $\ln(0)$ in the calculations. This step has led to the discussion of the bias of $ApEn$. In practice, it is established that this bias causes $ApEn$ to lack two important expected properties. First, $ApEn$ is heavily dependent on the record length and is uniformly lower than expected for short records. Second, it needs more relative consistency. If the $ApEn$ of one data set is higher than that of another, it should, but does not, remain higher for all conditions tested [56].

The family of statistics $ApEn(m, r, N)$ depends on three parameters N, m , and r that must be fixed for each calculation. N is the length of the time series, m is the length of sequences to be compared, and r is the tolerance for accepting matches [56].

Regardless, $ApEn(m, r, N)$ is biased and suggests more similarity than is present. It is important to note that $ApEn$ takes a template-wise approach to calculate this average logarithmic probability, first calculating a probability for each template. $ApEn$ statistics are biased, because the expected value of $ApEn(m, r, N)$ is less than the parameter $ApEn(m, r)$ [68].

A new family of statistics that does not count self-matches, the sample entropy ($SampEn$), was developed to reduce the bias of $ApEn$ [68]. $SampEn(m, r, N)$ 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. Thus a lower value of *SampEn* also indicates more self-similarity in the time series. In addition to eliminating self-matches, the *SampEn* algorithm is more straightforward than the *ApEn* algorithm, requiring approximately one-half as much time to calculate. *SampEn* is mainly independent of record length and displays relative consistency under circumstances where *ApEn* does not.

The computation of sample entropy starts by choosing two sequences of m consecutive points from the vector $X_N = \{x_1, x_2, \dots, x_N\}$, $X_m(i) = \{x_i, \dots, x_{i+m-1}\}$ and $X_m(j) = \{x_j, \dots, x_{j+m-1}\}$, ($i, j \in \{1, \dots, N-m\}, i \neq j$). The maximum distance between those two sequences is calculated and compared to the tolerance defined. Usually, the tolerance is defined by 0.1 – 0.2 times the standard deviation of the vector X_N [32]:

$$d[X_m(i), X_m(j)] = \max[|x_{i+k} - x_{j+k}|] \leq r \quad (k = 0, \dots, m-1; r \geq 0) \quad (4.1)$$

There are two major differences between *SampEn* and *ApEn* statistics. First, *SampEn* does not count the self-matches. It is justified to discount self-matches because entropy is conceived as a measure of the rate of information production. In this context, comparing data with themselves is meaningless. Second, *SampEn* does not use a template-wise approach when estimating conditional probabilities. To be defined, *SampEn* requires only one template to find a match of length $m+1$ [68].

4.1.2 Bubble and Permutation entropy

Sample and approximate entropy are very frequently used in the analysis of time series in many different scientific fields. The usage of both entropies relies on the definition of two parameters, m , the dimension of the vectors to study, and r , the threshold. Shifts in the parameters could drastically change the results obtained for both entropies. So, the scientific community has a great interest in reducing or even erasing the influence of these parameters on entropies analysis. Entropies like Bubble and Permutation, both arose as a solution for this problem. They both reduce the impact of the parameter m and erase the need for the definition of parameter r . They are both related and, in order to understand how bubble entropy works, it is essential to understand how permutation entropy and conditional Rényi entropy work first. Permutation entropy is free of parameters and measures a dynamic system's complexity. It calculates the probability distribution of sequences present in the time series analyzed. In permutation entropy, the one-dimensional time series is partitioned into a matrix with column vectors. In order to make this matrix, it is important to define the embedding dimension representing the length of each of the new column vectors, D , and the time delay that controls the number of periods between each element of the new column vectors, τ . After dividing the time series into a matrix with column vectors, the possible permutations for the defined embedding dimension of the vector are defined. For example, if $D = 3$, we know there are $3! = 6$ possible permutations. Then, the permutations are assigned to

each vector, the relative frequencies of each possible permutation are calculated, and the entropy is computed based on that [46].

The Rényi of order α for a vector X with n elements can be defined by the following equation [13]:

$$H_{\alpha}(X) = \frac{1}{1-\alpha} \log\left(\sum_{i=1}^n (p_i)^{\alpha}\right) \quad (4.2)$$

Bubble entropy is based on the permutation entropy method, but the effort required by the permutation process is calculated in this situation. Vectors are grouped by the number of swaps necessary to sort vectors, and then the Rényi entropy is computed for the groups made. Bubble entropy is based on the same principle as permutation entropy, however, instead of computing the relative frequencies of each vector, it calculates the relative frequency of the number of swaps required to get an ordered sub-sequence. So, in bubble entropy, first, the time series is partitioned into different column vectors. Then, the vectors resulting from the partition (X_i) are sorted by the number of necessary swaps (ni) to have an ordered sub-sequence. Finally, a histogram is made regarding the values of ni , normalized by $N - m + 1$, to obtain the probabilities (p_i) for each number of swaps (ni). In other words, the bubble entropy calculates how likely a certain number of swaps will be required [46].

4.1.3 Attention entropy

Attention entropy is computed by identifying the key patterns of a given time series. The key patterns are the peak points in the data. A point is a peak point (x) if it represents a local minimum or local maxima. So, a peak point needs to satisfy one of the following conditions:

- $x_{i-1} < x_i$ and $x_i > x_{i+1}$, for the local maxima being x_i ;
- $x_i < x_{i-1}$ and $x_i < x_{i+1}$, for the local minima being x_i

If each point of the time series is considered a representation of the system's current state, the change of one state/point to another can inform if and how the system is adjusting to the environment. This way, the peak points can also represent the top and bottom boundaries for the adjustments made by the system to the environment. When defining peak points as key patterns, four different situations are considered to calculate the intervals:

- intervals from local maxima to local maxima (max-max);
- intervals from local minima to local minima (min-min);
- intervals from local maxima to local minima (max-min);
- intervals from local minima to local maxima (min-max).

After defining the key patterns and which intervals to calculate, the intervals between two adjacent key patterns are calculated, and finally, the Shannon entropy is computed for the intervals. There are two different ways of calculating the attention entropy for the above mentioned intervals. We

can only consider one of the four situations for the possible intervals and calculate the entropy, or we can consider all four situations for the intervals and calculate the average of the four entropy values obtained. It is advisable to use the average of the four entropies, as it can reduce possible outliers in the data and does the calculation regarding a larger amount of information. However, the value of the entropies is usually similar for the four different cases. The most significant advantage of using attention entropy is that, given two time series with the same pattern frequency but different pattern distribution, attention entropy can distinguish both of them, unlike classical entropy methods [85].

4.1.4 Fuzzy Entropy

Fuzzy Entropy is a method that combines the fuzzy set theory with the entropy concept, as a randomness measure. Bringing the fuzzy set theory together with entropy, allows to deal with uncertainty in a better way [10]. In order to understand fuzzy entropy, it is important to primarily understand what fuzzy sets consist of. Fuzzy set theory can be seen as a way of modelling information. This method comes as different from the usual methods, since it uses a continuous classification (ranging from 0 to 1), instead of a discrete one, like 0 or 1, for the degree of membership of each element of the fuzzy set. If there is a set of data A, the degree of membership of each element of the set, can be given by the fuzzy set, which is presented in pairs, each one containing the element and its membership function [90]:

$$\tilde{A} = \{(x, \mu_{\tilde{A}}(x) \mid x \in X)\} \quad (4.3)$$

The membership function will be appropriate to the fuzzy set created, and its value for each element will range from 0 to 1, representing, as already mentioned, its degree of membership. So, for example, if the set has the following the six numbers:

$$A = \{1, 2, 3, 4, 5, 6\}$$

The fuzzy set should look something like this [90].:

$$\tilde{A} = \{(0, 0.1), (1, 0.3), (2, 0.7), (3, 1), (4, 0.7), (5, 0.3), (6, 0.1)\}$$

Fuzzy entropy was defined by De Luca and Termin, based on the membership function of the fuzzy set [10]:

$$H_A = -K \sum_{i=1}^n \{\mu_i \log(\mu_i) + (1 - \mu_i) \log(1 - \mu_i)\} \quad (4.4)$$

Where K is a positive constant, μ_i is the membership function, considering there are n membership functions. Besides the fuzzy entropy, also the 4 properties of the fuzzy entropy were defined:

- P-1 $H_A = 0$ if A is a crisp set ($\mu_i = 0$ or $1 \forall x_i \in A$);
- P-2 H_A is maximum if $\mu_i = 0.5 \forall x_i \in A$;

- P-3 $H \geq H^*$ where H^* is the entropy of A^* , a sharpened version of A ;
- P-4 $H = \bar{H}$ where $\bar{H}$ is the entropy of the complement set $\bar{A}$ [10].

4.1.5 Kolmogorov Entropy and K2 Kolmogorov Entropy

K2 kolmogorov entropy is a refined method of kolmogorov entropy, that is more suitable when studying the evolution on the complexity of experimented signals. When using the K2 entropy to measure the complexity of the system, its phase space is divided into small cells. Then, the probability that a trajectory in the system starts in one cell and ends up in another, after a certain amount of time, is estimated. K2 considers correlations between different cells, as opposed to kolmogorov entropy, which assumes that the system cells evolve independently [30]. By using K2 kolmogorov entropy, it is possible to measure the evolution of the system unpredictability over time. The use of this entropy is also suitable for systems where the initial conditions might prove to have a significant impact on its evolution, through time [38].

4.1.6 Distribution Entropy

Distribution Entropy is based on the probability distribution function, and studies the complexity of the signal, by estimating the distribution feature of inter-vector distances [42]. If there is a series defined by:

$$U(i), 1 < i < N, \quad (4.5)$$

Of N points, in order to compute the DistEn, the following steps are followed:

1. Form $N-m$ vectors defined as $X(i)$: $X(i) = \{U(i), U(i+1), \dots, U(i+m-1)\}$, with $1 \leq i \leq N-m$. The value of m is corresponding to the embedding dimension.
2. Define the distance matrix $D = d_{i,j}$ among the vectors $X(i)$ and $X(j)$ for all $1 \leq i, j \leq Nm$, wherein $d_{i,j} = \max\{|u(i+k) - u(j+k)|, 0 \leq k \leq m-1\}$ is the Chebyshev distance between $X(i)$ and $X(j)$.
3. By using the histogram approach, the probability density function of D is calculated. For each bin, there is a value of the probability/frequency of that bin, which is the value p_t , having t bins.
4. Finally, the probability density function is applied to the original formula of the Shannon Entropy, where M is the number of bins:

$$DistEn(m, M) = - \sum_{t=1}^M p_t \log_2(p_t) \quad (4.6)$$

By using DistEn, it is possible to, without depending on variance, study and identify the nonlinear behaviour of time series [43].

4.1.7 Conditional Entropy

If two systems are considered, the conditional entropy quantifies the uncertainty of the target system, when the other system is known. Imagining two time series, X and Y , conditional entropy, regarding Y , related to X , can be described as:

$$H(Y | X) = - \sum p(x, y) \log p(y | x) \quad (4.7)$$

Where $p(y)$ is the probability of $Y = y$ happening [22]. When considering one time series, conditional entropy can be defined by the following equation:

$$CE_n \left(\frac{N}{N-1} \right) = - \sum_{N-1} P_{N-1} \sum_{\frac{N}{N-1}} P_{\frac{N}{N-1}} \cdot \log \left(P_{\frac{N}{N-1}} \right) \quad (4.8)$$

Where P_{N-1} is the joint probability of $x_{N-1}(i)$, and $P_{N/N-1}$ is the conditional probability for the N th sample of $x_{N-1}(i)$ [14]. By using the Shannon Entropy formula,

$$SE_n(N) = - \sum_N P_N \log(P_N) \quad (4.9)$$

It is also possible to define the conditional entropy as [14]:

$$CE_n(N) = SE_n(N) - SE_n(N-1) \quad (4.10)$$

Conditional entropy is simple to compute, as simple as permutation entropy, for example [77]. This measure provides a quantitative measure of the relationship between two variables, or the dependency between one data point to the previous data point values, when considering one single time series [87].

4.1.8 Cross-entropies

It is possible to study the complexity associated with two different time series. For this kind of study, cross-entropy measures can be used. This method is resourceful, for example, when studying two coupled physiological signals is required. Cross-entropies use the same base function as the base entropies but consider two different time series. There are two main types of cross-entropies, direction-dependent and direction-independent. When a cross-entropy is direction-dependent, the result will differ, depending on which time series is considered first. The first of the two time series studied will serve as a template for the second to be compared. In a direction-independent cross-entropy, the same does not happen as the order in the time series does not change the results [89]. For example, cross-sample entropy is direction-independent, while cross-approximate entropy is direction-dependent [37].

4.1.9 Multiscale entropies

Multiscale entropies use the same base function as the base entropies, but in this case, the base functions are applied for several different time scales. This method is used to consider the possible variability of time series, like usually happens in physiological signals recordings [15]. Each different time scale is created based on the original time scale, and then the entropy base functions are applied to the new time series created [67].

4.1.10 Mutual Information

Mutual information (MI) is usually used to evaluate the relationship between two random variables simultaneously sampled, by calculating how much information is exchanged between them. As the name states, this method evaluates whether two variables share any mutual information [47]. Mutual information allows us to measure the dependence between two random variables. Given two random variables, X and Y , MI can be calculated and expresses in entropy :

$$MI(X;Y) = H(Y) - H(Y/X) \quad (4.11)$$

where $H(X/Y)$ is the uncertainty about Y when X is known, and $H(X)$ represents the Shannon entropy for variable X . If Y and X are independent, then the MI equals zero, as knowing one of the variables would not reduce the entropy of the other. The further away from zero MI is, the bigger the relation between X and Y is. So, if the value is closer to zero, it means that having information about one variable would not reduce the uncertainty related to the other [72].

4.2 Compressors and Normalized Compression Distance

4.2.1 Compressors

The increasing importance and usage of data in modern times results in a bigger demand for data compression methods to improve the efficiency of data-related studies and operations [35]. Compressors are a data-reduction scheme used to reduce the amount of data by reducing redundancies. This scheme is used by encoding the data and decoding it again [64]. Compression techniques can either be lossy or lossless, but both aim to have the data presented more compactly, reducing the number of bits. Lossless techniques eliminate statistical redundancy, leading to no information losses. On the other hand, lossy compression removes either unnecessary or less important information, compacting the data that way [40]. Throughout the next sections the techniques used in this project, that are commonly applied in physiological signals, will be briefly described.

4.2.1.1 Huffman Coding

Huffman coding, a tree-based approach, is a lossless data compression algorithm invented by David Huffman [34]. The first step is to define the set of leaf nodes, and each node has a code assigned according to its frequency on a given sequence of data inputs. Then, the nodes with the

smallest frequency/weight are selected and removed from the set of leaf nodes and combined to create a new internal node corresponding to the sum of the two selected components. This process is repeated until there is only one internal node left. Considering a data input with n characters, this cycle should be repeated for $n - 1$ iterations until the final internal node is obtained, corresponding to the sum of the original frequencies/weights, and the process is stopped [49].

4.2.1.2 Lempel Ziv

Lempel Ziv (*LZW*) algorithm is also a lossless data compression method. It uses a similar method to Huffman coding, as it replaces character counts in a sequence using a referencing dictionary. *LZW* dictionary has 256 entries, from 0 – 255. Having a data sequence of characters, *LZW* groups the characters into strings and replaces them with code. It is frequently used for text files but can also be used in other file formats [35].

4.2.1.3 Lempel-Ziv Markov Chain Algorithm

This compressor uses two different ways of compressing, statistical modeling and dictionary techniques. It also replaces frequent characters occurrences with pointers to similar sequences of characters [6].

4.2.1.4 Bzip2

Bzip2 compressor has a similar way of working as the Lempel-Ziv Markov Chain Algorithm [3]. This method was developed by Julian Seward in 1996, and it is an opensource program. Bzip2 can compress but cannot archive files. It is able to compress the files up to 10 to 15%, comparing to other available techniques. It transforms frequent characters sequences into identical letter strings [1].

4.2.1.5 Gzip

It is estimated that a gzip compressor is able to compress a file up to 80%, decreasing the page load time and bandwidth consumption [7]. It works slower than bzip2 compressor and is a combination of two different algorithms, LZ 77 and Huffman coding. LZ 77 is a lossless data compression technique, where the repeated characters sequences are replaced by a reference, throughout the file. LZ 77 is used to refer to repeated characters sequences, and to the distance between them. Then, Huffman coding is used to compress the data compressed with the LZ 77, and is used to generate variable length codes, for the length of the repeated sequences of characters [2].

4.2.2 Normalized Compression Distance

Generally speaking, Normalized Compression Distance (*NCD*) represents how different two data sequences are, considering that a smaller *NCD* would represent a bigger degree of similarity. Usually, the value representing the *NCD* between two different data sequences, r , is a non-negative

number and $0 \leq r \leq 1 + \epsilon$. The ϵ is associated with possible errors or imperfections in the compression techniques used [36].

Equations 4.12 and 4.13 refer to C as the chosen data compressor to compute the NCD . If it is known that $C(y) \geq C(x)$, it is possible to write the NCD equation [81]:

$$NCD(x, y) = \frac{C(xy) - C(x)}{C(y)} \quad (4.12)$$

The NCD method uses normalized compressors. A compressor is considered normalized if it satisfies these four conditions:

1. Idempotency: The compressor should detect exact repetitions and compress the empty string to an empty one.
2. Monotonicity: The compressor should respect the monotonicity property.
3. Symmetry: Not all compressors are precisely symmetric, but they should be symmetric up to the required precision.
4. Distributivity: The distributivity property means that the compressor should obey the following equation:

$$C(xy) + C(z) \leq C(xz) + C(yz) \quad (4.13)$$

Where $C(z)$ is the compressed version of the string z [81].

4.3 Programming

The language that will be used in this dissertation is Python [78], and the programming will be made in Spyder, using version 3.10.9 of Python. The entropy measures described in chapter 4 were computed using the Entropy Hub package, which is an open-source toolkit that is used to calculate the entropy values of time-series [24]. The Mutual Information was calculated using the Python' package *scikit-learn*, where the score of the Mutual Information between two variables is returned by the *mutual_info_score* function [4, 5].

For the compressors *bzip2* and *gzip*, the function *to_csv* from the *pandas* library [84] was used, specifying the type of compressor selected. The *lzma* compressor's corresponding library was employed, *lzma.py* [?]. To obtain the Normalized Compression Distance (NCD), the Python library *zgli* was used, which calculates the NCD between all the files in a certain folder [11].

Chapter 5

Classifiers

When diagnosing a disease, health professionals usually use traditional tests, like collecting samples from the patient and performing lab analysis. These processes are usually time-consuming, and the diagnosis could be performed too late to allow an efficient treatment [51]. However, in many situations in medicine, like the diagnosis of diseases, it is required for a more reliable but expeditious way of making decisions. According to Vili Podgorelec et al. [58], the construction of decision models with the ability to learn is one appropriate way of making a rapid and trustworthy decision in the field of medicine.

A classification problem is when a class label is predicted for a given input data. For example, when talking about medicine, class 0 is if the patient is not infected, and class 1 is if the patient is infected by a specific pathology, in the case of binary classification. Usually, this classification is based on previously studied data, where the class assigned to each data point is known [20].

It is also important to understand which of the features mentioned in chapter 4 should be included in the classification process. In order to decide that, statistical tests may be applied, like Mann-Whitney test. The Mann-Whitney test is used to study if two independent groups are from the same population. For this test, the null hypothesis is that the two group's distributions are equal, against the alternative hypothesis that they have a significant difference. The test statistic p-value is used to see if it is possible to reject the null hypothesis, and for a small value, it is possible to conclude that the two independent groups do not belong to the same population [44]. In this case, the p-value established was 0.2, so all the measures with a Mann-Whitney test p-value smaller or equal to 0.2 would be considered for the classification process.

5.1 Logistic Regression Classification

Logistic regression classification defines a mathematical model representing the relationship between two independent variables. In binary classification, the output should be either 0 or 1, so a function that fulfills that need is necessary. The most commonly used is the sigmoid function, and a graph example is seen in figure 5.1 and equation 5.1 [91].

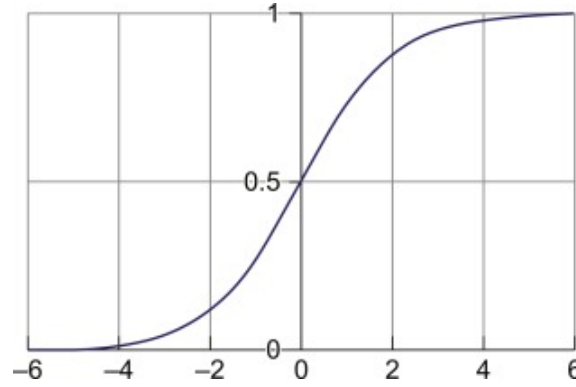


Figure 5.1: Sigmoide function. Figure adapted from [91].

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (5.1)$$

By using this sigmoid function, it would mean that any value of z under 0 would belong to the class 0 and any value of z over 0 would be classified as 1. To calculate z , each characteristic of the input data $(x_0, x_1, x_2, x_3, \dots, x_n)$, is multiplied by its regression coefficient $(w_0, w_1, w_2, w_3, \dots, w_n)$ [91]:

$$z = w^T x \quad (5.2)$$

Considering the utilization of the sigmoid function, the output data values are in between 0 and 1. If the output data is superior to 0.5, the result of the classification will be 1, if it is below 0.5, it will be classified as data 0. In order to define the regression coefficient, usually the gradient descent method is used, by reducing the value of the error, Y [91]:

$$Y = |y' - y| \quad (5.3)$$

Where y' is the prediction of the classifier and y is the actual category label. The gradient descent method is done by iterations of the following function [91]:

$$w = w - \alpha \frac{\partial Y}{\partial w} \quad (5.4)$$

The regression coefficient is w , α is the iteration step size and Y is the error function [91].

One advantage of the logistic regression is that it allows to understand the influence of each parameter of the data to the prediction, through the regression coefficients [20].

5.2 Tree Based Methods

5.2.1 Decision Trees

Decision Tree is a classification algorithm that started to be computed by the end of the 20th century, making it one of the oldest statistical methods be used for this process [50]. Generally, decision tree is a subsetting target data method, where the data is attributed to different subsets according to input fields or predictors. Then they are placed in different descendent data subsets that contain progressively similar intra-leaf or intra-node features or progressively dissimilar intra-leaf or intra-node features. This placement occurs throughout the different levels of the tree, associating each data point of the target data to its corresponding class and maximizing the separation between the data. The first node or leaf of the decision tree is the root node. Each node represents how the split of the data on that node will be made, for example, which parameter and for which values of that same parameter will the data be separated. Each node splits the data into different subsets, that should leave to the next node or leaf. This process is repeated until the final level or the decision node is reached, where the outcome of the classification is presented [19].

Usually, the criteria used to split the data is information gain, meaning that at each level of the tree, the entropy should decrease along the data splitting. A disadvantage associated with decision tree method is that at each step/node, the single best variable and optimal split point are selected, making the process susceptible to errors, for example, when the most significant variable is not selected correctly. A way to change this is to use a model that could consider different combinations of variables. A great advantage of the decision tree is that it is possible to know how the input is turned into an output easily, making it easier to understand and use this method in other fitting situations, and that is one of the reasons why these models are widely used in medicine [20].

Figure 5.2 represents a simple example of a decision tree. The first node of the tree, splits the data based on how is the outlook: sunny overcast or rain. Then on the next level of the tree, the data is split according to humidity or how windy it is, depending on which subset it was assigned to in the previous level. The final purpose of the decision tree is to classify whether the data belongs to the class P or N [60].

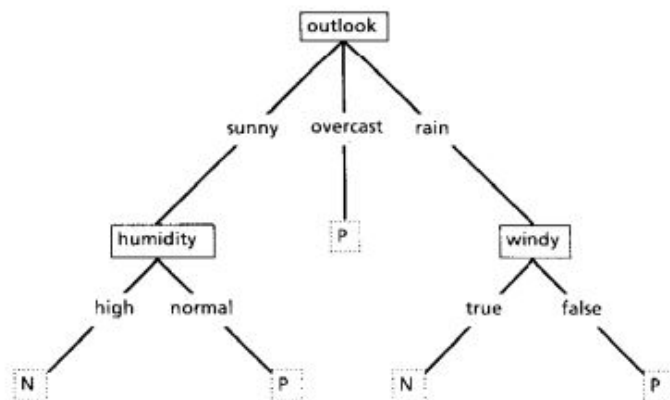


Figure 5.2: example of a decision tree classification. Figure adapted from [60].

5.2.2 Random Forest Classifier

Random Forest was first proposed at the University of California's Statistics Department, by Leo Breiman. This statistical method is based on the decision tree classification process. It uses multiple decision trees as an ensemble [12].

Ensemble classification is a way of getting a more precise classification. In this process, the model uses multiple classifiers and, each of them outputs its own class label for the data. Then, after a voting process, the class with the most votes become the model prediction [63].

One of the greatest advantages brought by the random forest classifier, is the Bootstrap Aggregation (Bagging). Bagging is a process that is proved to increase the effectiveness of some classifiers, and it consists of allowing different classifiers to use different subsets of the original data. By doing this, in the case of the random forest classifier, each tree will use different subsets of the data and therefore, each will have significant different structures from each other, allowing the classification to make predictions with higher variability, and allowing each decision tree to make predictions independently [26].

Another characteristic that keeps the multiple decision trees of the random forest independent, is the random features selection. Each decision tree will make a prediction, based on different subsets of the features to be used in the model [54]. The combination of bagging and random features has shown a good performance, especially for classification problems [12].

Figure 5.3 shows how the random Forest classifier model works. The first step is to divide both the original data and the features into different subsets. Afterwards, each subset is used in each decision tree, so that they perform the independent prediction. After all the predictions are obtained from each tree, the voting process starts, and the class with the most votes, becomes the final model prediction [54].

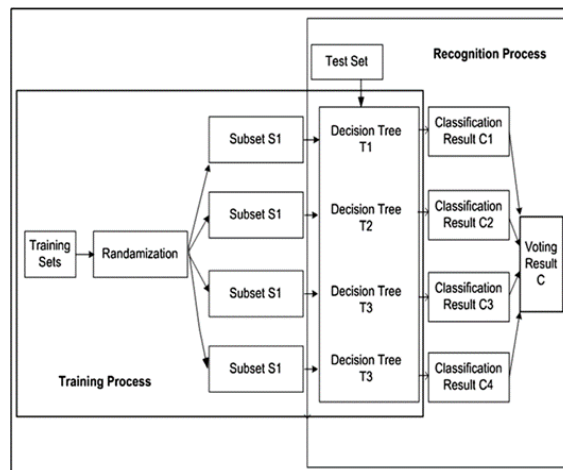


Figure 5.3: Random Forest Classifier schematics. Figure adapted from [54].

5.3 Programming

The classification algorithms, logistic regression classification, decision tree classification and random Forest classifier, can be computed using the *scikit-learn* package [5, 55], from Python [78].

Chapter 6

Results and Discussion

6.1 Preliminary results

During the last academic year, within the scope of the Curricular Unit of *Projeto e Técnicas Laboratoriais* (PTL), a study of complexity measures, entropy measures in this case, was performed for the signals of heart rate and respiratory rate. The data used for the project was from the open-source database, Physionet [29, ?].

First, the sample, approximate, bubble, attention, and cross-sample entropy were computed using the Entropy Hub package [25] on Python [78]. For the multiscale sample (MsSamp) and approximate entropy (MsAp), the parameters defined were: an embedding dimension of 2, a time delay was 1, and the radius distance threshold given by 0.2 times the standard deviation of the studied signal. For bubble multiscale entropy (MsB), the embedding dimension was 2, and the time delay was 1. Attention multiscale entropy (MsAt) only required the definition of the logarithmic base. For each multiscale entropy, the first scale was considered as the original scale of the time series. The complexity index (CI) summarized the multiscale information, computed as the sum of the entropy for the several scales. Table 6.1 presents the median, minimum and maximum values found for each entropy for both the heart rate and respiratory signals.

Table 6.1: Median [minimum; maximum] values for the entropy index for the heart rate and respiration time series: Cross-sample entropy (XSampEn), Multiscale sample entropy (MsSamp, CiS), Multiscale approximate entropy (MsAp, CiA), Multiscale bubble entropy (MsB, CiB) and Multiscale attention entropy (MsAt, CiAt)

Entropies	Heart Rate	Respiration
MsSamp	0.107 [0.003; 1.170]	0.290 [0.003; 0.966]
CiS	0.513 [0.015; 4.513]	1.368 [0.014; 3.493]
MsAp	0.145 [0.004; 1.244]	0.482 [0.022; 1.066]
CiA	0.710 [0.016; 4.055]	1.821 [0.089; 3.375]
MsB	0.190 [0.016; 0.615]	0.154 [0.014; 0.382]
CiB	0.888 [0.078; 1.866]	0.904 [0.082; 1.565]
MsAt	3.093 [1.703; 3.449]	3.148 [0.347; 3.403]
CiAt	8.095 [4.768; 9.349]	8.291 [1.040; 9.237]
XSampEn	0.260 [0.02; 3.170]	0.260 [0.02; 3.170]

To study the correlation between the uni and bivariate measures, heatmaps were computed (figures 6.1 and 6.2).

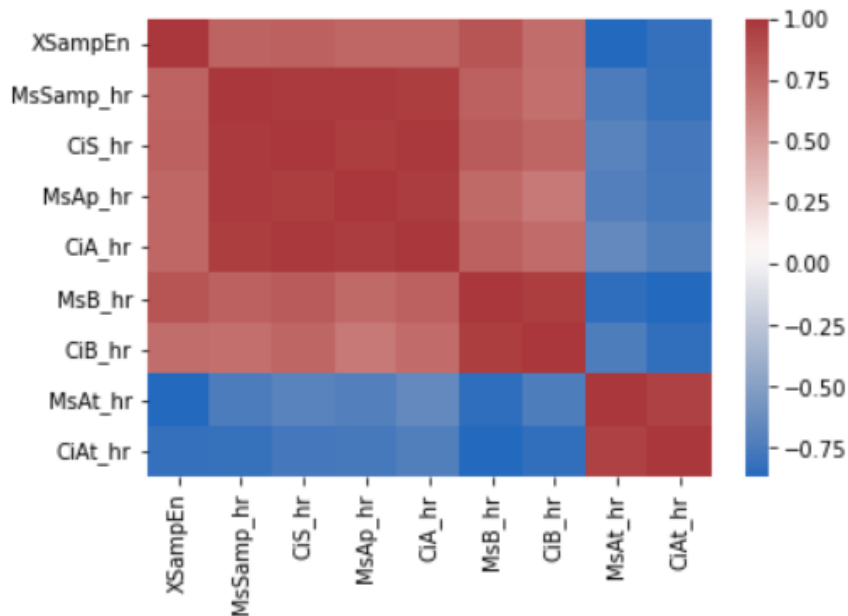


Figure 6.1: Correlation HeatMap for Cross-Sample entropy (XSampEn), Multiscale Sample entropy (MsSamp_hr, CiS_hr), Multiscale Approximate entropy (MsAp_hr, CiA_hr), Multiscale Bubble entropy (MsB_hr, CiB_hr) and Multiscale Attention entropy (MsAt_hr, CiAt_hr) for Heart Rate signal

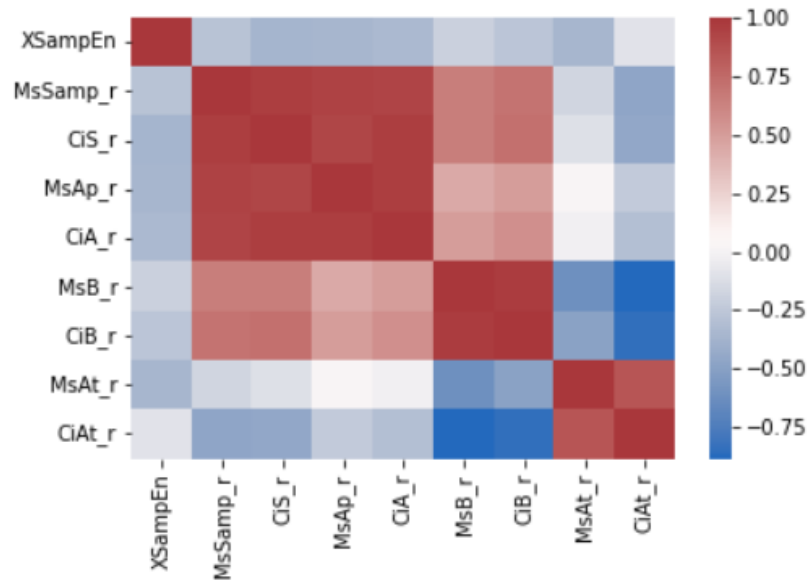


Figure 6.2: Correlation HeatMap for Cross-Sample entropy (XSampEn), Multiscale Sample entropy (MsSamp_hr, CiS_hr), Multiscale Approximate entropy (MsAp_hr, CiA_hr), Multiscale Bubble entropy (MsB_hr, CiB_hr) and Multiscale Attention entropy (MsAt_hr, CiAt_hr) for Respiratory Rate signal

The heatmaps presented allowed comparing the correlation between the multiscale entropies with the cross-sample entropy. For heart rate, all the multiscale entropies had, at least, a moderate positive correlation with cross-sample entropy. Only the attention entropy computed did not have a positive correlation. In fact, the results showed a strong negative correlation between multiscale attention entropy and all the other measures. These results would mean that, for a high value of the multiscale attention entropy, a small value of the cross-sample entropy should be registered for the heart rate signal.

For the respiratory signal, all the multiscale entropies negatively correlated with the Cross-sample entropy. An increase in the multiscale entropies would mean a decrease in the cross-sample entropy and vice-versa.

Therefore, it was concluded that the heart rate multiscale entropies had a positive correlation, and the respiratory rate multiscale entropies had a negative correlation with the cross entropy. These measures could help to understand the dynamics between both physiological signals, as well as how the variation on them would influence the cross entropy measures. By studying the shifts in the relation between the entropies of different physiological signals, it is also possible to predict the behavior of the organism. That should be the primary purpose of this Masters Dissertation: understand the reactions of the organism associated with neonatal sepsis infection and predict when the disease could occur.

6.2 Data

The clinical data used for the current project belongs to the previously mentioned project dataset, Digi-NewB [82]. It contains clinical information: gestational age, newborn weight, sex, and the result of the diagnosis for neonatal sepsis (infected or not infected). The dataset used also contained physiological signals records (heart and respiratory rate) from the infected and control neonates.

6.2.1 Sample Analysis

The clinical data is relative to 33 individuals, 17 males ($\approx 51.5\%$) and 16 females ($\approx 48.5\%$). Of the 33 individuals, 20 ($\approx 60.6\%$) were diagnosed with neonatal sepsis, and 13 ($\approx 39.4\%$) were considered controls (without sepsis). The median, minimum, and maximum of the weight and ages of the studied participants can be found in table 6.2.

Table 6.2: Clinical information of the 33 neonates.

	Sepsis (N=20)		Control (N=13)	
	Median	Minimum-Maximum	Median	Minimum-Maximum
Weight (g)	860	520 - 1690	1115	705 - 1490
Gestational Age (weeks)	27.6	24.7 - 30.4	29.6	25.7 - 32.1

6.2.2 Signal Representation

Figures 6.3 and 6.4 present the first 10 minutes of the heart rate and respiratory rate signals for one newborn infected with sepsis.

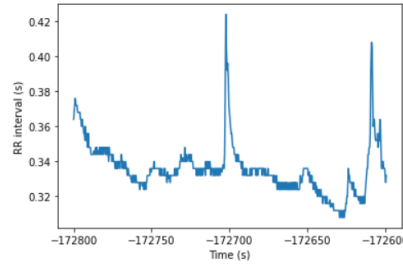


Figure 6.3: First 10 minutes of Heart Rate signal for an infected individual.

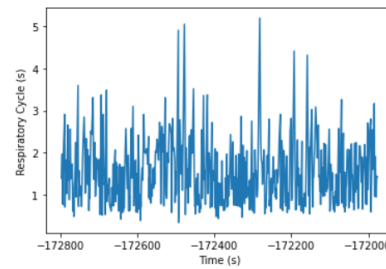


Figure 6.4: First 10 minutes of Respiratory signal for an infected individual.

6.3 Resampling

Most of the bivariate measures used in this thesis imply that both signals have the same length. However, the heart and respiratory rate signals do not have a constant (fixed) sampling frequency. Therefore, both signals were resampled at fixed frequencies. The resampling was experimented for the following frequencies: 1sample/0.25s; 1sample/0.5s; 1sample/s; 1sample/1.5s; 1samples/2s; 1sample/2.5s; 1sample/3s. The forward filling method from the *pandas* library was used to fill the missing values (nan) obtained for each resampling, where all the missing values (nan) were filled with the last non-missing value. The resampling number of samples obtained and the original number of samples for each signal and each frequency were compared, as can be seen in the figures 6.5 and 6.6.

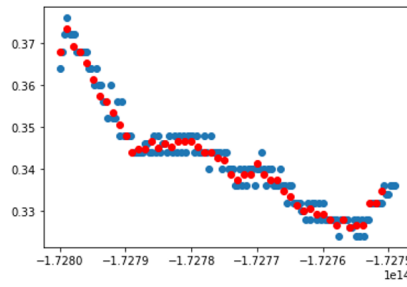


Figure 6.5: [Original heart rate signal (blue) VS resampled heart rate signal (red) for 1sample/s.] The horizontal axis represents the time (s) until infection. The vertical axis represents the signal

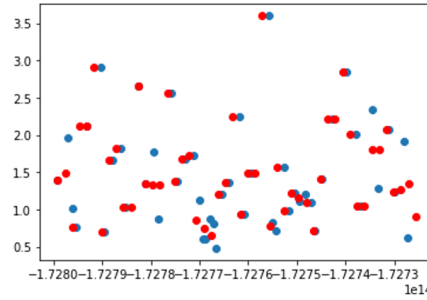


Figure 6.6: [Original respiratory rate signal (blue) VS resampled respiratory rate signal (red) for 1sample/1.5s. The horizontal axis represents the time (s) until infection.] The vertical axis represents the signal.

After studying the distribution of the original signal files comparing to the resampled signal, the chosen frequencies were 1sample/0.5s (2Hz); 1sample/s (1Hz); 1 sample/1.5s (0.66Hz) and 1sample/2.5s (0.4Hz).

After choosing these four sampling frequencies, the univariate and bivariate entropy and compression measures were computed for the original and resampled signals. All the entropies were computed using default values from the EntropyHub package, and the compression algorithms were implemented using the libraries mentioned in Chapter 6.

The correlation between each pair of the sampling frequencies considered and between the original files for all the different complexity measures was calculated, as displayed in the following two correlation heatmaps (figures 6.7 and 6.8). For both figures, each number along the correlation heatmap vertical axis, refers to the number of the the following list, with the combinations included for the heatmap. The following pairs of combinations were compared:

1. original file VS resampled for 1sample/0.25s
2. original file VS resampled for 1sample/0.5s
3. original file VS resampled for 1sample/s
4. original file VS resampled for 1sample/1.5s
5. original file VS resampled for 1sample/2s
6. original file VS resampled for 1sample/2.5s
7. original file VS resampled for 1sample/3s
8. resampled for 1sample/0.25s VS resampled for 1sample/0.5s
9. resampled for 1sample/0.25s VS resampled for 1sample/s
10. resampled for 1sample/0.25s VS resampled for 1sample/1.5s
11. resampled for 1sample/0.25s VS resampled for 1sample/2s
12. resampled for 1sample/0.25s VS resampled for 1sample/2.5s
13. resampled for 1sample/0.25s VS resampled for 1sample/3s
14. resampled for 1sample/0.5s VS resampled for 1sample/s
15. resampled for 1sample/0.5s VS resampled for 1sample/1.5s
16. resampled for 1sample/0.5s VS resampled for 1sample/2s
17. resampled for 1sample/0.5s VS resampled for 1sample/2.5s
18. resampled for 1sample/0.5s VS resampled for 1sample/3s
19. resampled for 1sample/s VS resampled for 1sample/1.5s
20. resampled for 1sample/s VS resampled for 1sample/2s
21. resampled for 1sample/s VS resampled for 1sample/2.5s
22. resampled for 1sample/s VS resampled for 1sample/3s
23. resampled for 1sample/1.5s VS resampled for 1sample/2s
24. resampled for 1sample/1.5s VS resampled for 1sample/2.5s
25. resampled for 1sample/1.5s VS resampled for 1sample/3s
26. resampled for 1sample/2s VS resampled for 1sample/2.5s
27. resampled for 1sample/2s VS resampled for 1sample/3s
28. resampled for 1sample/2.5s VS resampled for 1sample/3s

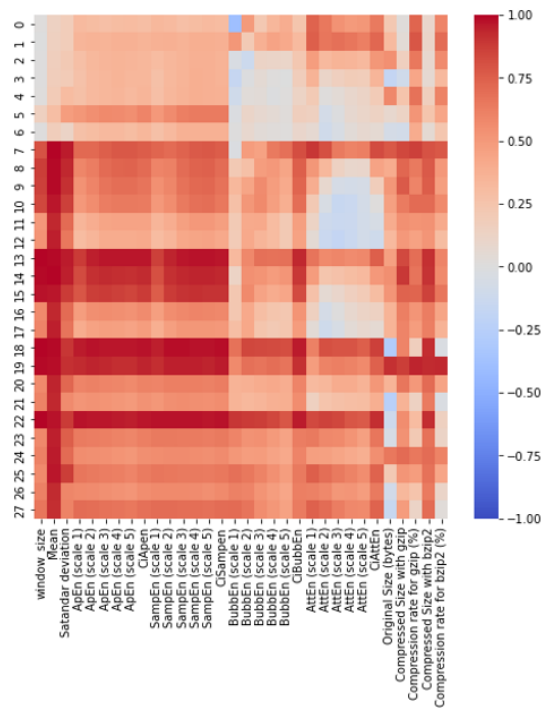


Figure 6.7: Correlation Heatmap for Heart Rate. Each correlation value, is representative of the similarity between the pairs of combinations of the files compared (previous list).

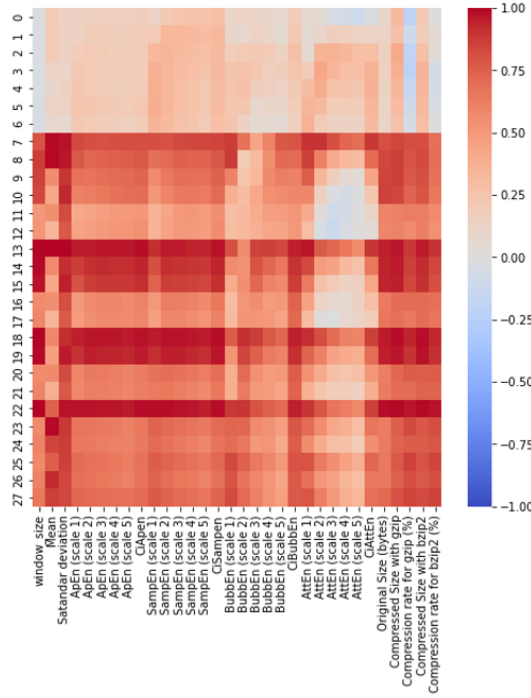


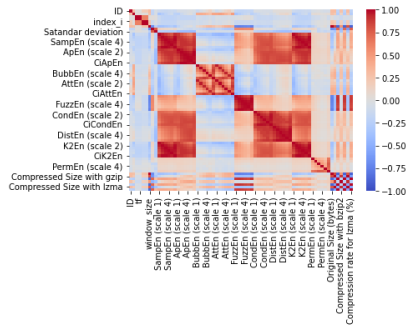
Figure 6.8: Correlation Heatmap for Respiratory Rate. Each correlation value, is representative of the similarity between the pairs of combinations of the files compared (previous list).

After observing the correlation heatmaps, the chosen frequencies remained the same: 1 sample/0.5s; 1 sample/1s; 1 sample/1.5s, and 1 sample/2.5s, as those seemed to be the sampling frequencies that showed more proximity to the original signal results. In addition, considering that the heart and the respiratory rate had significantly different numbers of values, different sampling frequencies could be more beneficial for only one of the signals. For example, the heart rate signal would lose fewer samples for a smaller sampling frequency, like 1 sample/0.5s, than the respiratory rate. Taking this into account, using a wide range of sampling frequencies for the study might be beneficial.

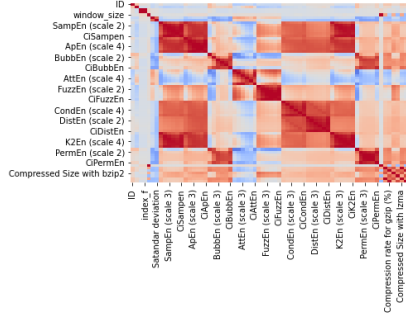
6.4 Complexity Measures Selection

For this work, we considered the Entropy Hub package's entropies that allowed cross-entropy: sample, approximate, fuzzy, conditional, distribution, K2 Kolmogorov, and permutation entropy. Besides these ones, bubble entropy was selected to be computed once it has been widely used. The multiscale entropies were computed for 5 different time scales when considering the univariate measures. The cross-entropies were computed for one-time scale only. For all the entropies computed, the parameters values selected were default. Some entropies present some similarities in their concepts and definitions. Therefore, after computing the measures for all the different sampling frequencies, the correlation heatmap between all the complexity measures was obtained

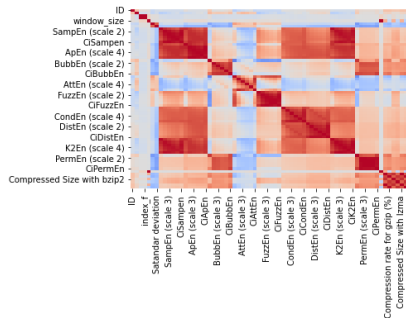
for each sampling frequency. These correlations allowed to decide which entropies should be considered for the study in order to create a more efficient algorithm. Figures 6.9 and 6.10 show the correlation heatmaps.



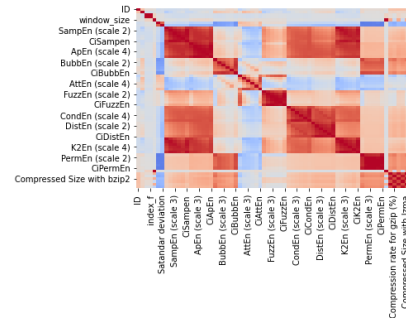
(a) Original Signal



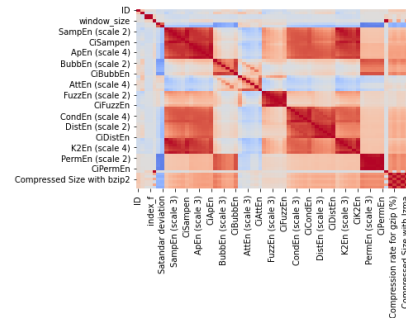
(b) Sampling frequency of 1sample/0.5s



(c) Sampling frequency of 1sample/s

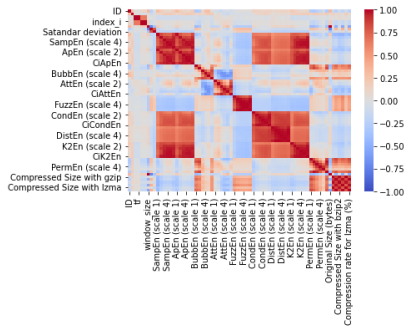


(d) Sampling frequency of 1sample/1.5s

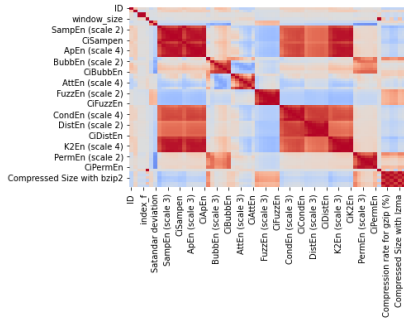


(e) Sampling frequency of 1sample/2.5s

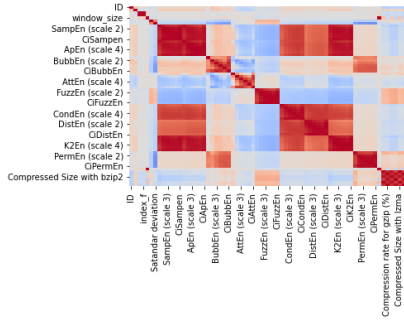
Figure 6.9: Correlation Heatmaps for the measures for Respiratory Rate



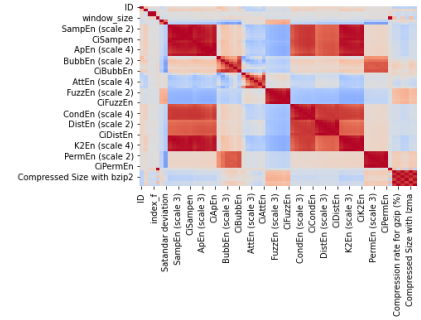
(a) Original Signal



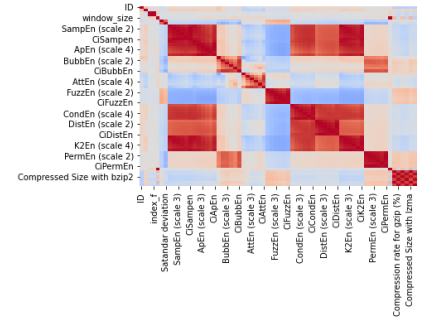
(b) Sampling frequency of 1 sample/0.5s



(c) Sampling frequency of 1 sample/s



(d) Sampling frequency of 1 sample/1.5s



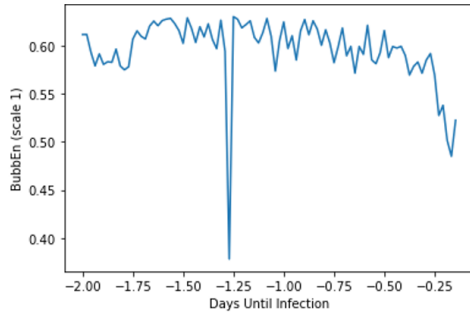
(e) Sampling frequency of 1 sample/2.5s

Figure 6.10: Correlation Heatmaps for the measures for Heart Rate

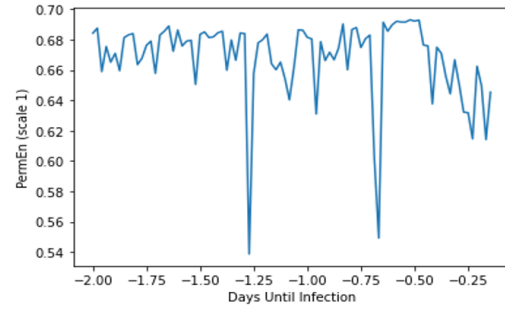
After observing the correlation heatmaps, approximate entropy was excluded because it had similar values to sample entropy. The conditional entropy was also removed, as it had a high correlation with K2 Kolmogorov and distribution entropy for both signals. Finally, attention entropy was also not considered for the project's subsequent phases, as it had a high positive correlation with bubble entropy or fuzzy entropy, in different sampling frequencies, for respiratory rate. Attention entropy also presented a high negative correlation with bubble entropy for the heart rate signal, for most sampling frequencies.

6.5 Complexity Measures Analysis for Infected Subjects

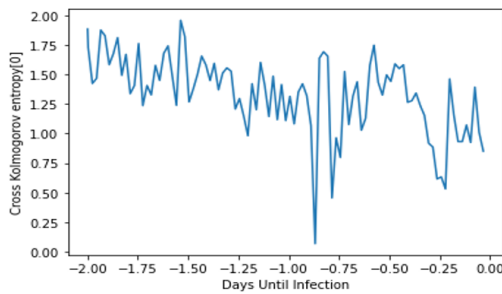
After obtaining the preliminary results for the infected subjects, it was possible to study the evolution of the complexity measures selected to predict the disease. Figure 6.11 shows examples of the complexity measures computed. Note that, the horizontal axis is presented in days until infection and the vertical axis contains the values of the complexity measure.



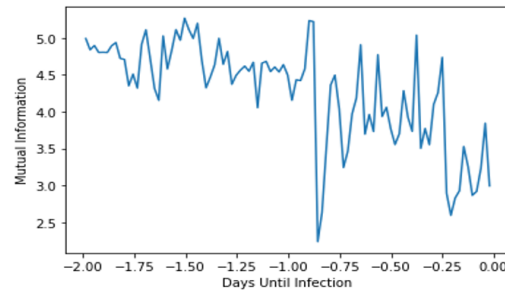
(a) Bubble entropy (scale 1) for subject 110035, for a sampling frequency of 1 sample/second, for the Heart Rate signal.



(b) Permutation entropy (scale 1) for subject 110035, for a sampling frequency of 1 sample/second, for the Respiratory Rate signal.



(c) Cross Kolmogorov entropy for subject 5005, for a sampling frequency of 1 sample/2.5 seconds.



(d) Mutual Information for subject 110034, for a sampling frequency of 1 sample/2.5 seconds.

Figure 6.11: Complexity measures of heart rate and respiratory signal for the Infected Subjects.

The evolution of the computed measures of the signal behaved in similar way. Additionally, in some entropy measures like Bubble (Figure 6.11a) entropy, it was possible to observe an abrupt increase in the complexity of the signal when reaching the time of the diagnostic. For some

subjects, this increase was registered at an earlier stage. The preliminary results suggest that a pattern in the evolution of the physiological signals could be found when using complexity measures to study them. It is important to consider that the results obtained in Figure 6.11 are regarding infected patients. So, the project's next phase would be to compare these values with healthy subjects. The pattern for infected and healthy subjects was expected to be significantly different, allowing the classifier to discriminate between both classes.

A boxplot is also an essential tool to observe the data distribution. For this plot, the subjects' results were combined in each box, for each feature, and for each 30-minute window studied. The plot was performed for every sampling frequency. It can be observed a slight shift in the median, from a particular instant, like for the K2 Kolmogorov entropy on the univariate measures for the respiratory signal. Also, some boxplots can show shifts for a specific period, like the sample entropy, on the univariate measures for the heart rate signal. In the bivariate measures, Cross-permutation shows a slight shift in the median for the 30-minute windows. Figures 6.12, 6.13 and 6.14 show the examples of the boxplots for the mentioned measures, all the other boxplots can be consulted in Appendix A.

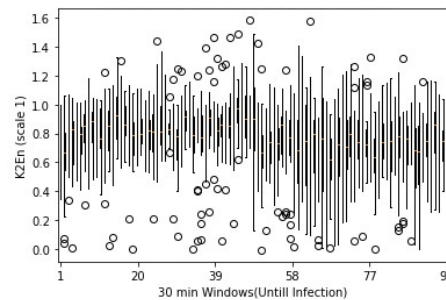


Figure 6.12: K2 kolmogorov entropy boxplots for 30 minute windows until infection, for the respiratory signal.

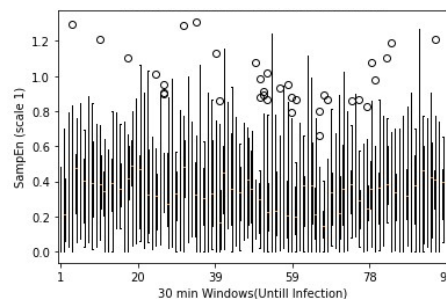


Figure 6.13: Sample entropy boxplots for 30 minute windows until infection, for the heart rate signal.

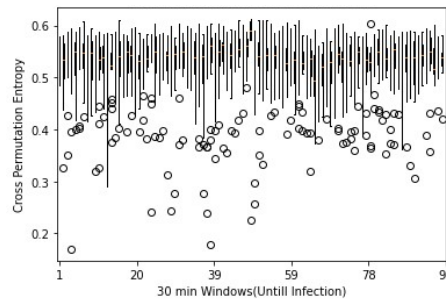


Figure 6.14: Cross permutation entropy boxplots for 30 minute windows until infection, for the coupling of both signals.

After observing all the boxplots regarding every feature and sampling frequency, it was noticed that the four different sampling frequencies used behaved similarly. Considering that, for the next phase of the study, only the highest (1 sample/0.5s) and the lowest (1 sample/2.5s) sampling frequencies were used.

6.6 Control and Infected Subjects

After deciding which features to use, based on the results for the infected subjects, the previously selected complexity measures were computed also for the control group.

For each of the original files and each frequency, the values of each measure were computed in windows of 30 minutes and plotted against the newborns' gestational age. It is possible to notice slight differences in the distribution of the measure's values, as in some representations, some areas have significantly more values belonging to infected subjects than control subjects and vice-versa. Figures 6.15 and 6.16 show examples of two measures where it is possible to see areas where either the infected or control subjects' results are more predominant. The red cross represents the 30-minute window for an infected newborn, and the light grey circle represents the 30-minute window for a control subject.

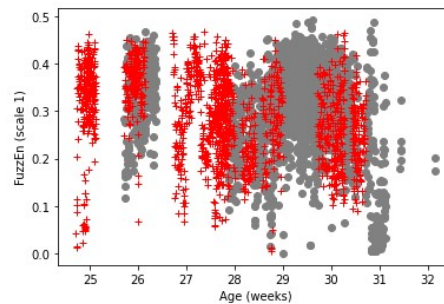


Figure 6.15: Control (grey) vs Sepsis (red) values for the respiratory signal (1 sample/0.5s), for Fuzzy entropy.

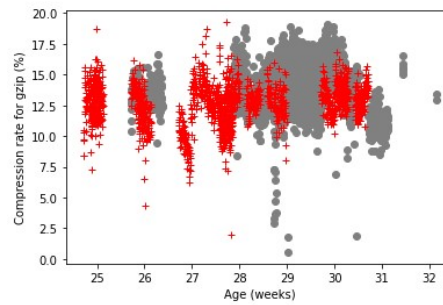


Figure 6.16: Control (grey) vs Sepsis (red) values for the heart rate signal (1 sample/2.5s), for the compression rate for gzip.

6.6.1 Classification

The slope of all the 30-minute windows for each subject and each feature was computed. This process was performed for both sampling frequencies mentioned earlier and the original signal in the univariate measures case. Before the classification process was executed, it was essential to understand which features were more associated with the outcome to improve the classification process. The method used was the Mann-Whitney test, considering relevant measures the ones with a p-value ≤ 0.2 . Here are the selected complexity measures for the next step of the project (the p-value table for all the computed measures can be consulted in Appendix C):

- Univariate Measures for the Heart Rate Signal:

Original file: PermEn (scale 1), PermEn (scale 3), PermEn (scale 4), PermEn (scale 5), Compression rate for gzip (%).

Resampled for 1sample/0.5s: SampEn (scale 1), PermEn (scale 3), PermEn (scale 4), PermEn (scale 5), Compression rate for gzip (%), Compression rate for bzip2 (%).

Resampled for 1sample/2.5s: PermEn (scale 2), CiPermEn, Compression rate for gzip (%).

- Univariate Measures for the Respiratory Rate Signal:

Original file: SampEn (scale 2), SampEn (scale 3), SampEn (scale 4), SampEn (scale 5), CiSampEn, BubbEn (scale 4), FuzzEn (scale 1), FuzzEn (scale 2), FuzzEn (scale 3), FuzzEn (scale 4), CiFuzzEn, DistEn (scale 1), DistEn (scale 2), K2En (scale 1), K2En (scale 2), K2En (scale 3), K2En (scale 4), K2En (scale 5), CiK2En, PermEn (scale 3), PermEn (scale 4).

Resampled for 1sample/0.5s: SampEn (scale 1), SampEn (scale 2), SampEn (scale 3), SampEn (scale 4), SampEn (scale 5), CiSampEn, FuzzEn (scale 1), FuzzEn (scale 2), FuzzEn (scale 3), FuzzEn (scale 4), CiFuzzEn, DistEn (scale 1), DistEn (scale 2), DistEn (scale 3), DistEn (scale 4), DistEn (scale 5), CiDistEn, K2En

(scale 1), K2En (scale 2), K2En (scale 3), K2En (scale 4), K2En (scale 5), CiK2En, PermEn (scale 5).

Resampled for 1sample/2.5s: SampEn (scale 1), SampEn (scale 2), SampEn (scale 3), SampEn (scale 4), SampEn (scale 5), CiSampEn, FuzzEn (scale 1), DistEn (scale 1), DistEn (scale 2), DistEn (scale 5), CiDistEn, K2En (scale 1), K2En (scale 2), K2En (scale 3), K2En (scale 4), K2En (scale 5), CiK2En, PermEn (scale 1), Compression rate for bzip2 (%).

The data split was performed in order to have 70% (23 subjects) of the data for training and 30% (10 subjects) for testing. The division was seeded in order to have a better comparison among classification algorithms. The parameters used were the default ones for each package's functions. Table 6.3 presents the accuracy score for the univariate measures for each classifier, separated by signal and sampling frequency. The accuracy score estimates if the labels predicted for the test data corresponded to the true values of the test data labels. The values can range from 0 to 1, 1 representing that the model predictions match exactly the real labels [55].

Table 6.3: Accuracy of the Classifiers. 23 subjects for training and 10 subjects for testing.

Signal	Signal File (original or resampled)	Logistic Regression Accuracy	Decision Tree Accuracy	Random Forest Regressor Accuracy
Respiratory Rate Signal	Original	0.70	0.50	0.50
	1sample/0.5s	0.70	0.50	0.50
	1sample/2.5s	0.70	0.60	0.50
Heart Rate Signal	Original	0.70	0.50	0.60
	1sample/0.5s	0.70	0.60	0.80
	1sample/2.5s	0.70	0.60	0.80

The confidence interval for the accuracy score is also very important to understand how the error associated with the model can vary. The distribution of the data was assumed as normal [69]. Considering that, for each of the results presented in table 6.3, all the confidence intervals were calculated, with the confidence of 95%, meaning $z = 1.96$. Table 6.4 shows the accuracy's confidence intervals obtained. The confidence interval for the accuracy is representative of between which values the accuracy can vary. For example, in the next table, the confidence interval for the heart rate signal of the logistic regression is 0.16, which would mean that, given the 0.7 accuracy, it is possible to affirm with 95% confidence, that the value of the accuracy should be in the interval (0.7-0.16, 0.7+0.16).

Table 6.4: Confidence Interval for the Accuracy of the Classifiers.

Signal	Signal File (original or resampled)	Logistic Regression Accuracy	Decision Tree Accuracy	Random Forest Regressor Accuracy
Heart Rate Signal	Original	0.16	0.17	0.17
	1sample/0.5s	0.16	0.17	0.17
	1sample/2.5s	0.16	0.17	0.17
Respiratory Signal	Original	0.16	0.17	0.17
	1sample/0.5s	0.16	0.17	0.14
	1sample/2.5s	0.16	0.17	0.14

Predominantly, the classifiers performed better for the univariate measures of the Heart Rate signal. Random forest had a significantly better accuracy score for the heart rate signal, especially for the original time series and the sampling frequency of 1 sample/0.5s and 1 sample/2.5s, as they had better accuracy and smaller confidence interval. For the respiratory signal, logistic regression performed better, either for the original signal or the different sampling frequencies.

Chapter 7

Conclusion

The goal of the project was to study the dynamics of heart and respiratory signals in the detection of neonatal sepsis. Additionally, the project aimed to compare the coupling of both signals to understand the advantages of using both signals simultaneously instead of just one. Bearing that in mind, several complexity measures were computed to describe the evolution of the signal complexity, with the end goal of obtaining an effective neonatal sepsis classifier.

As previously mentioned, the results for the infected subject's complexity measures suggested the possibility of finding a pattern for the infected subject signal, which is useful for training the classification model. Afterwards, it was possible to observe some significant and slight shifts in the median and quartiles of the complexity measures throughout time. Those shifts were more evident for the univariate respiratory signal results than the heart signal results. The boxplots could guide the observer to think that, for the univariate measures, the respiratory signal could show a more evident pattern for the infected subjects, resulting in better classification.

Then, the differences between the infected and the control subjects appear similar for each signal. Finally, the classification was performed. The logistic regression showed the same accuracy score of 70%. On a different note, for heart rate signal the random forest had a better performance, having an accuracy score of 60%, 80%, and 80% again for the original file, the resampled signal for 1sample/0.5s and the resampled signal for 1 sample/2.5s, respectively. These results could be expected. Random forest is a classifier better suited for data where the randomness should be foreseen. The boxplot observations indicate a higher randomness in the evolution of the signal complexity for the heart signal.

Resampling was also a crucial stage of the project. The resample had to be performed to compute the cross entropies and the mutual information, as these measures require that both signals have the same length. As described in Chapter 8, respiratory and heart rate files did not have the same sampling frequency or signal length. The results suggest that the resampling process was not crucial for the classification accuracy using the univariate

measures, only for the case of the random forest using the heart rate signal. As already stated, this signal seemed more complex for infected subjects, so the random forest classifier could be the best fit. Considering that the heart rate signal had a vast number of samples compared to respiratory rate, a lot of the samples seemed to have almost the same value, and this could make the shifts in the signal complexity less evident, as for each 30-minute window, a lot of the values would be very similar. This makes the signal less complex, confirmed by observing the entropy and compression rate values. After performing the resample and having samples less frequently, the shift in the values might become more evident, as each 30-minute window would have fewer samples, probably more or less similar to each other, which could lead to a better performance.

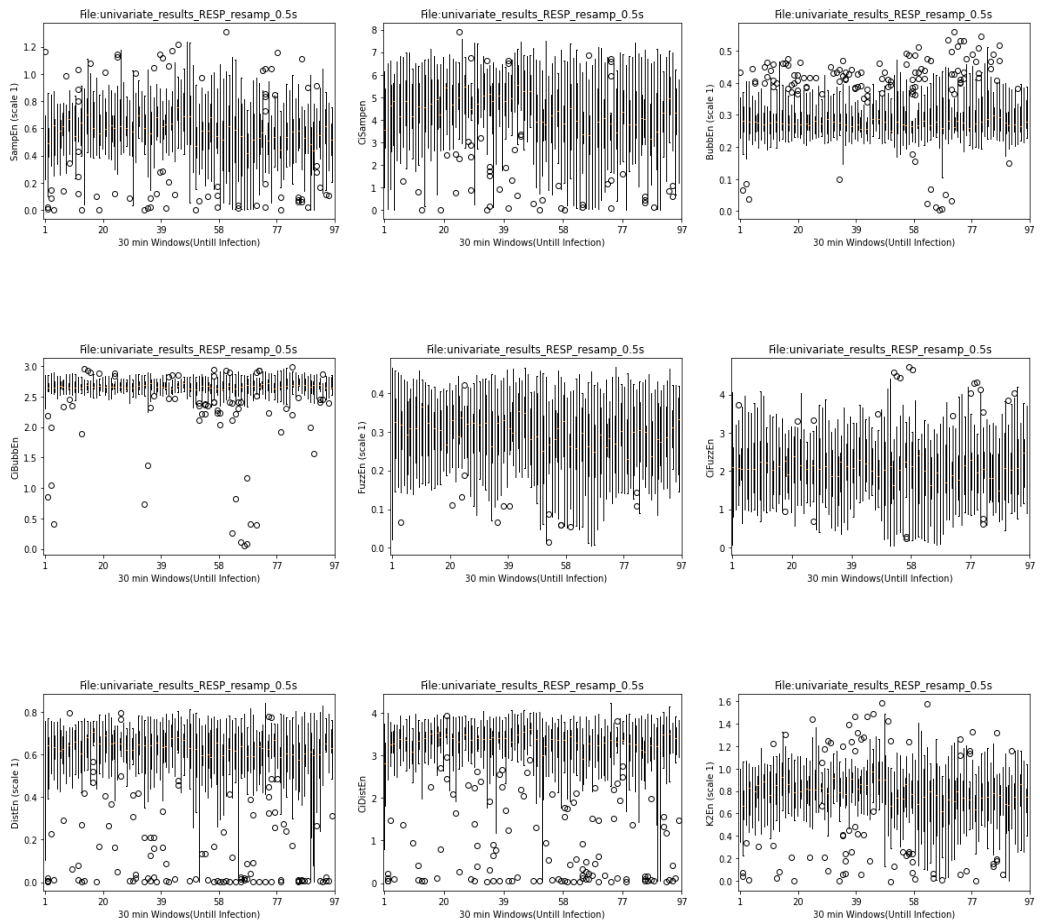
In the future, some improvements should be made. Features for the model training could be added, like conditional and attention entropy, that were removed from the study, not only due to the strong correlation they had with other entropies but also due to implementing problems during the project, which, due to lack of time, were not solved. The parameters of the features could also be changed to see if they could perform better. Other parameters and values could be experimented for the complexity measures, classifiers, and train/test data split. One of the essential improvements would be performing the study for a more extensive dataset, improving the model's training, and avoiding over-fitting more efficiently. It would also be important to obtain the results for the classification, based on the bivariate measures, comparing them to the univariate measures.

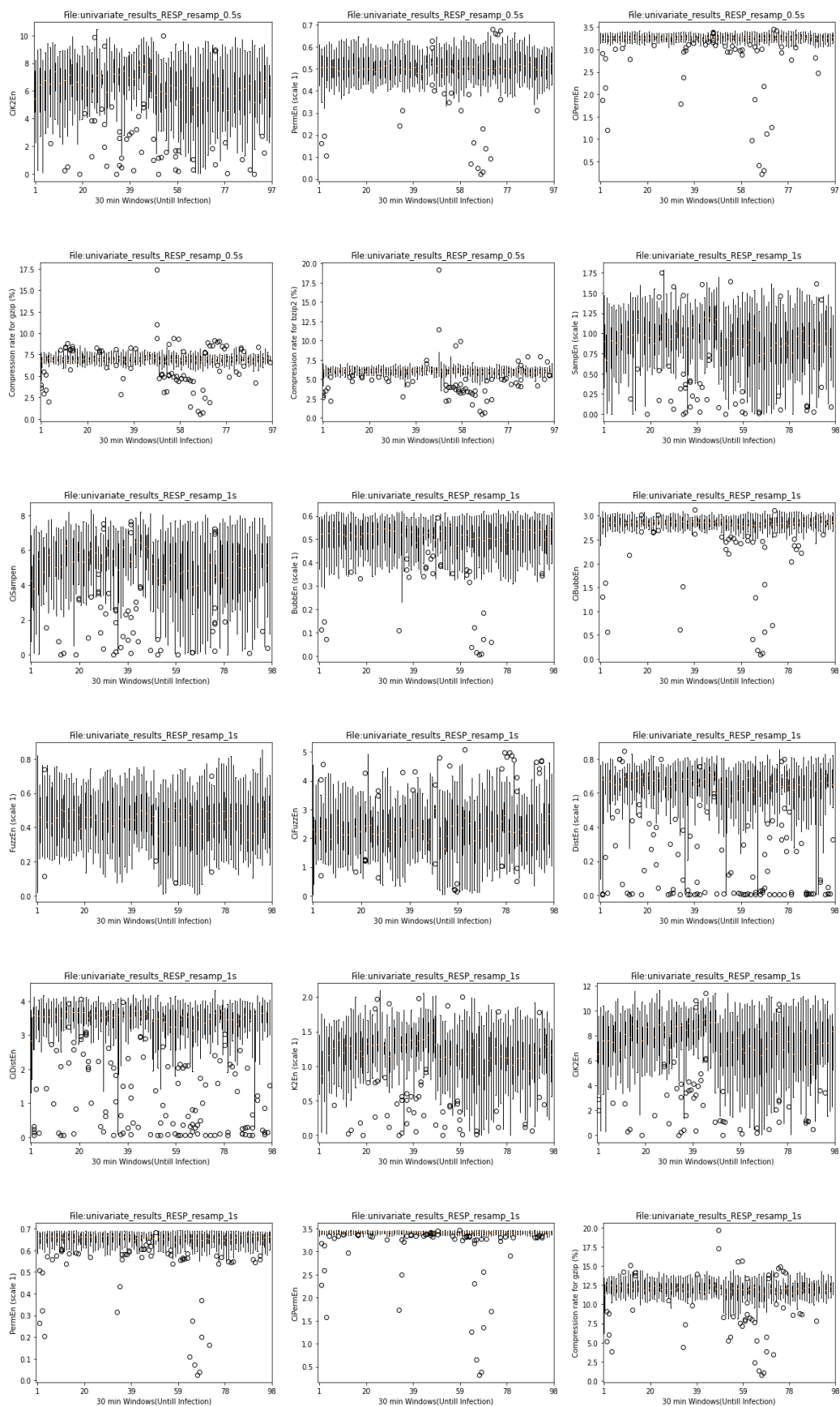
Overall, studying the heart and respiratory signals through time could have a huge role in developing non-invasive tools for the early diagnosis of neonatal sepsis. By performing an early detection, a more suited treatment could be administered to the patient, resulting in a higher probability of a positive outcome and the decrease of the mortality rate of the disease, as well as the antibiotic resistance provoked by it.

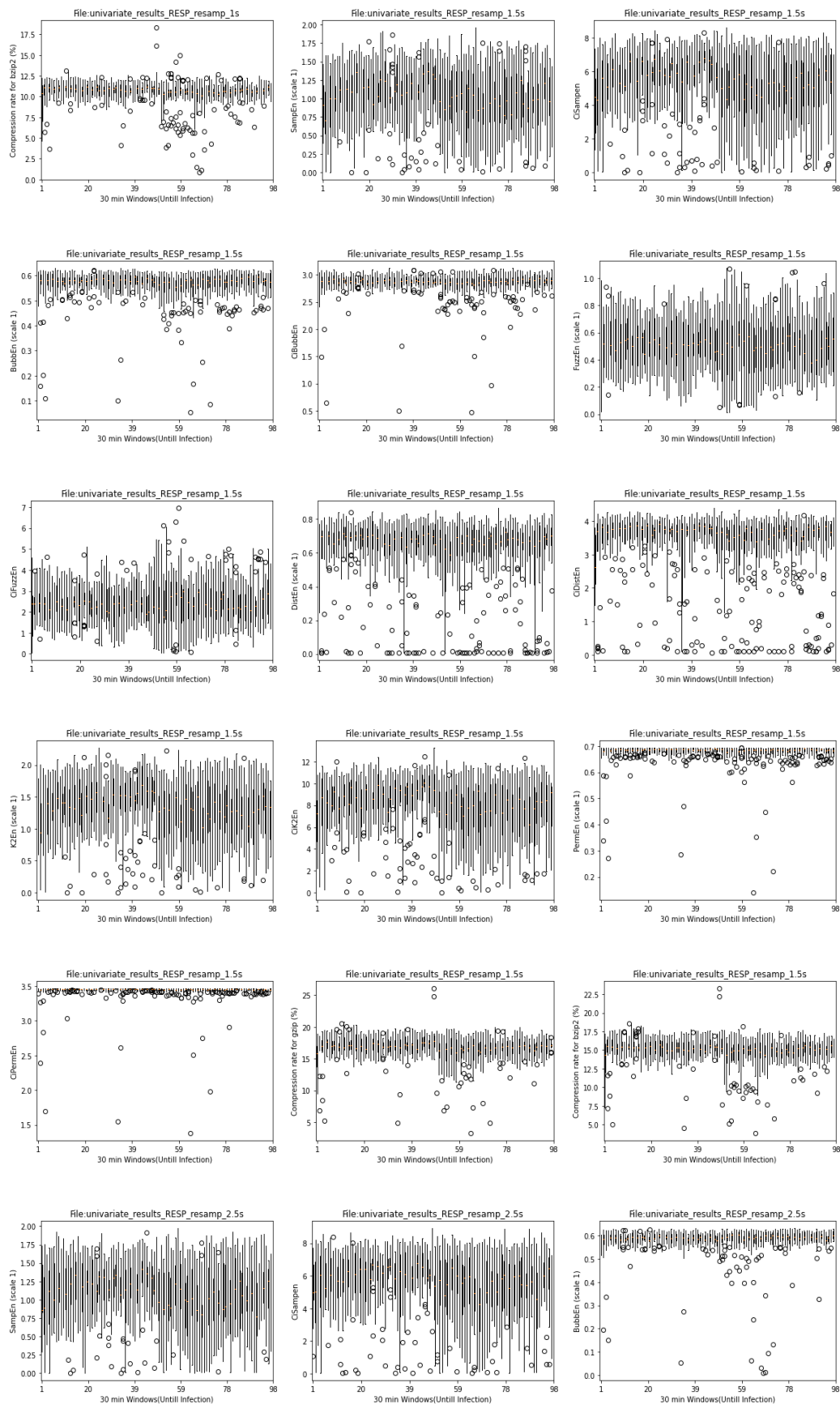
Appendix A

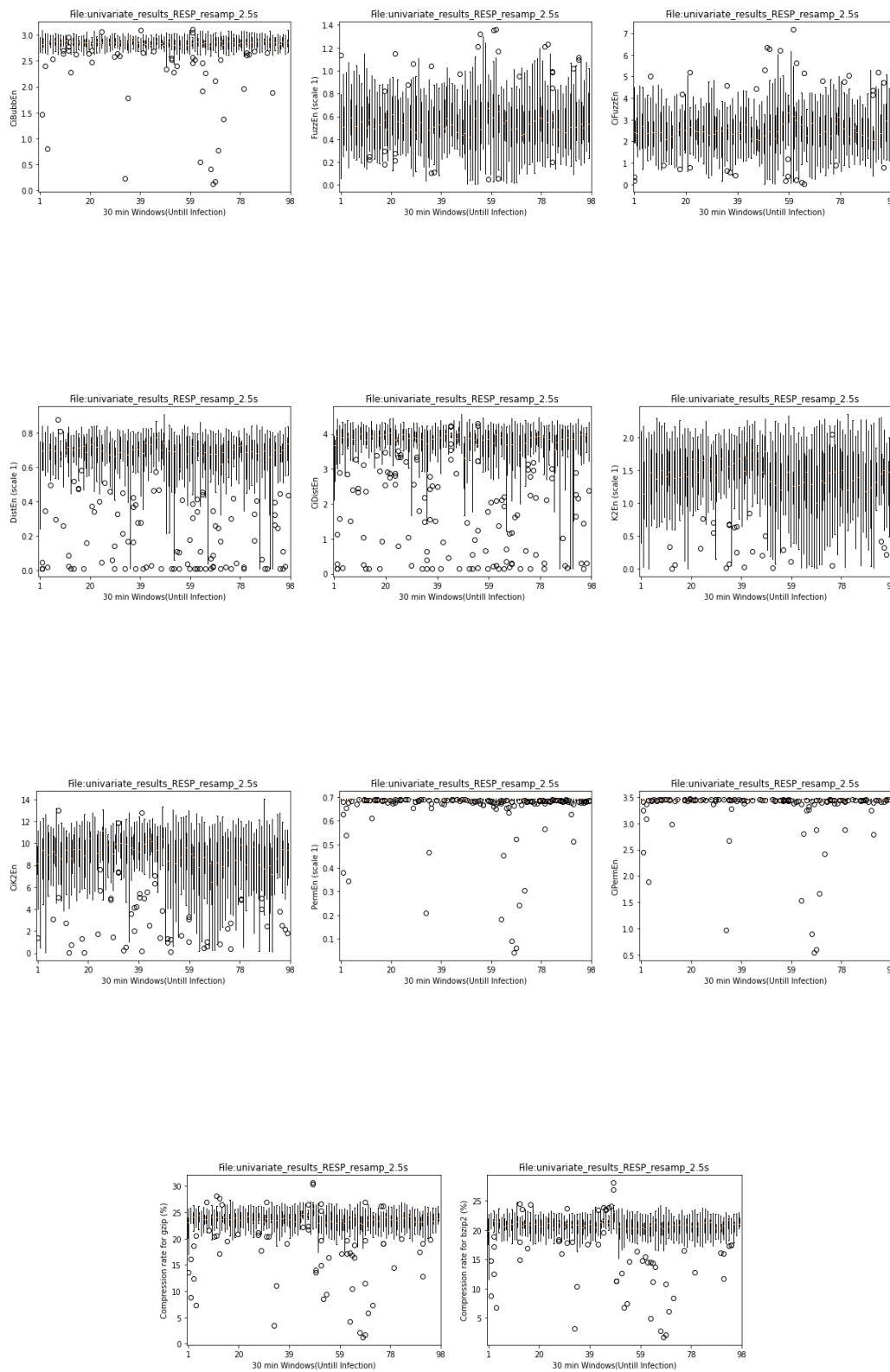
Boxplot

A.1 Boxplots for the Univariate Measures for Respiratory signal

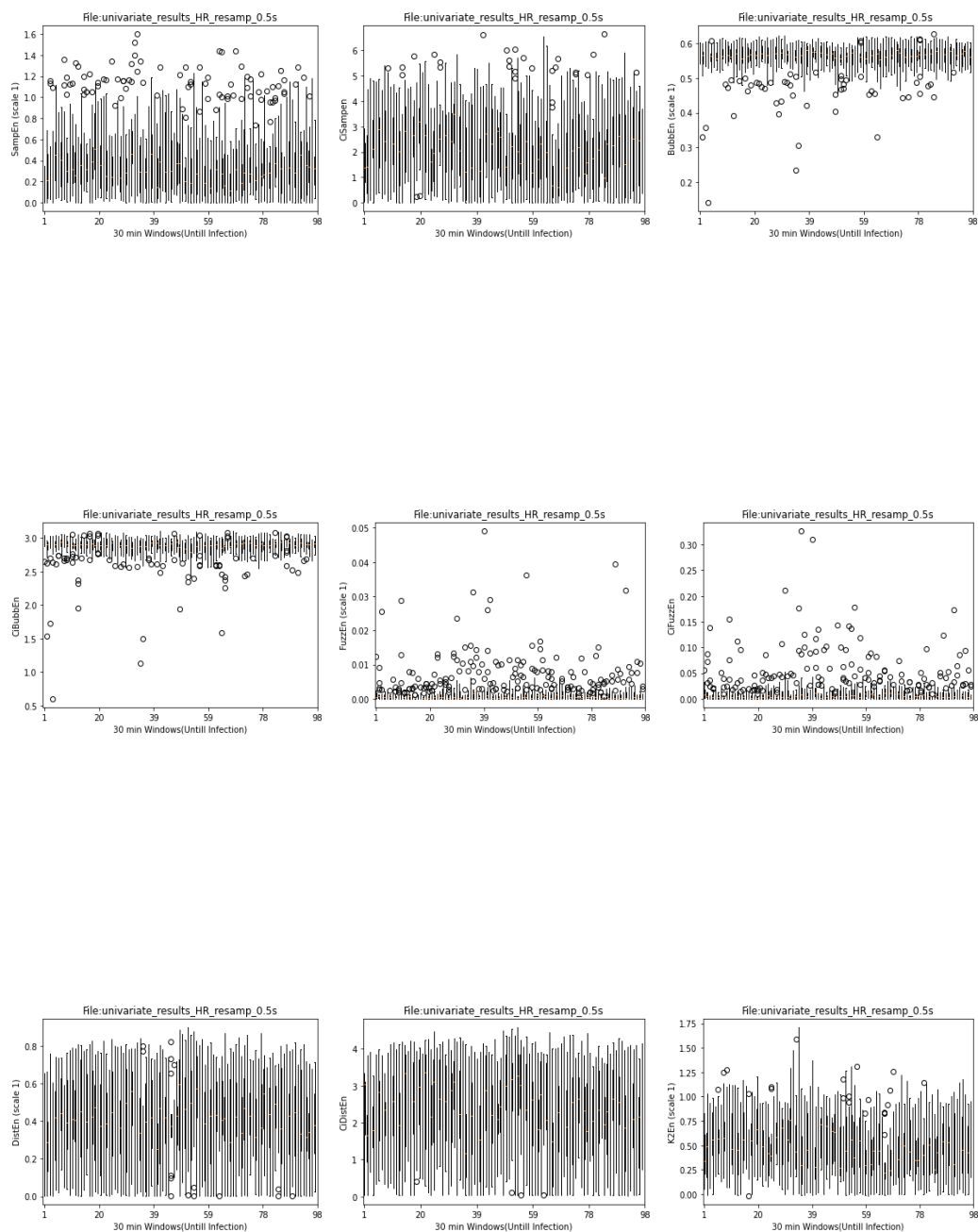


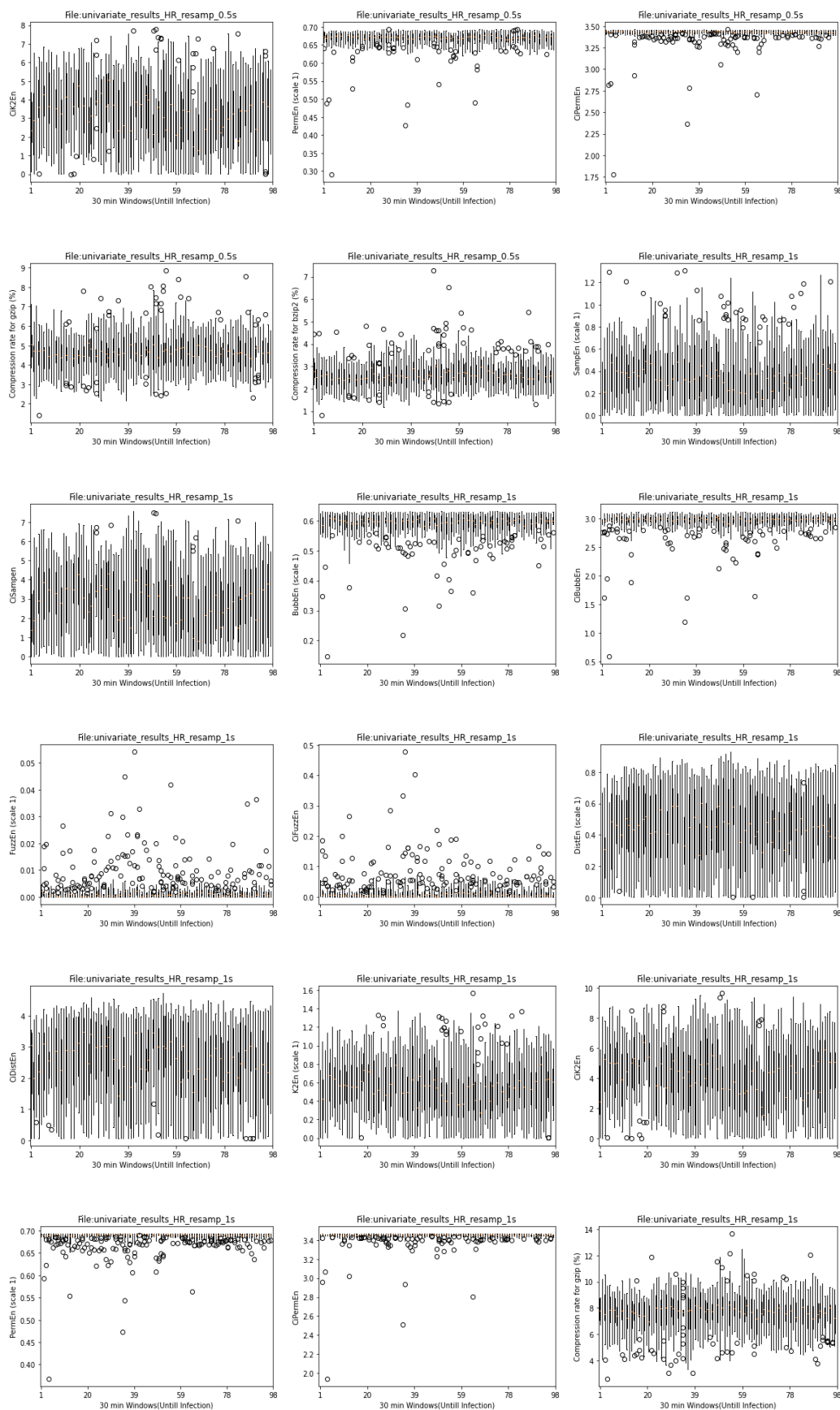


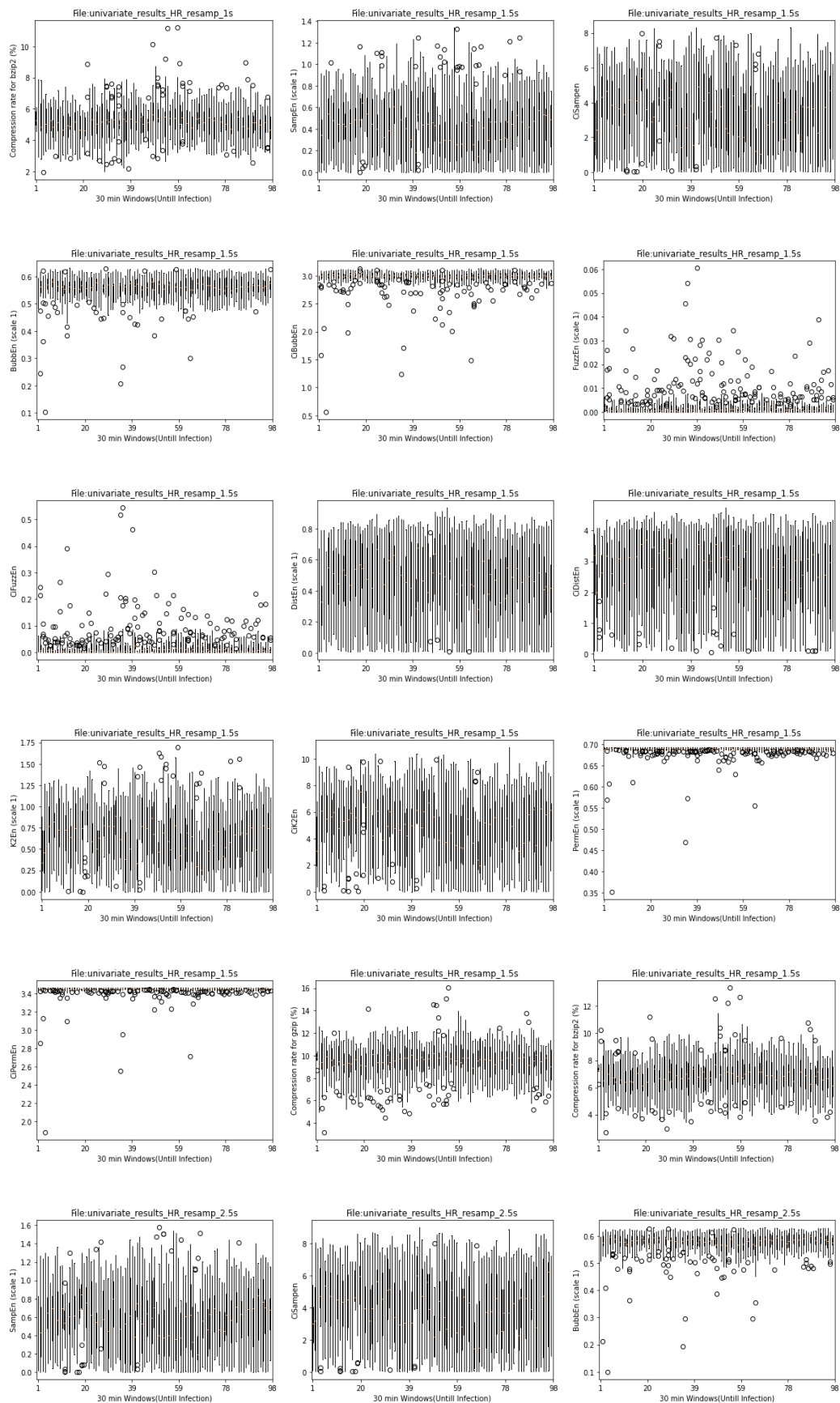


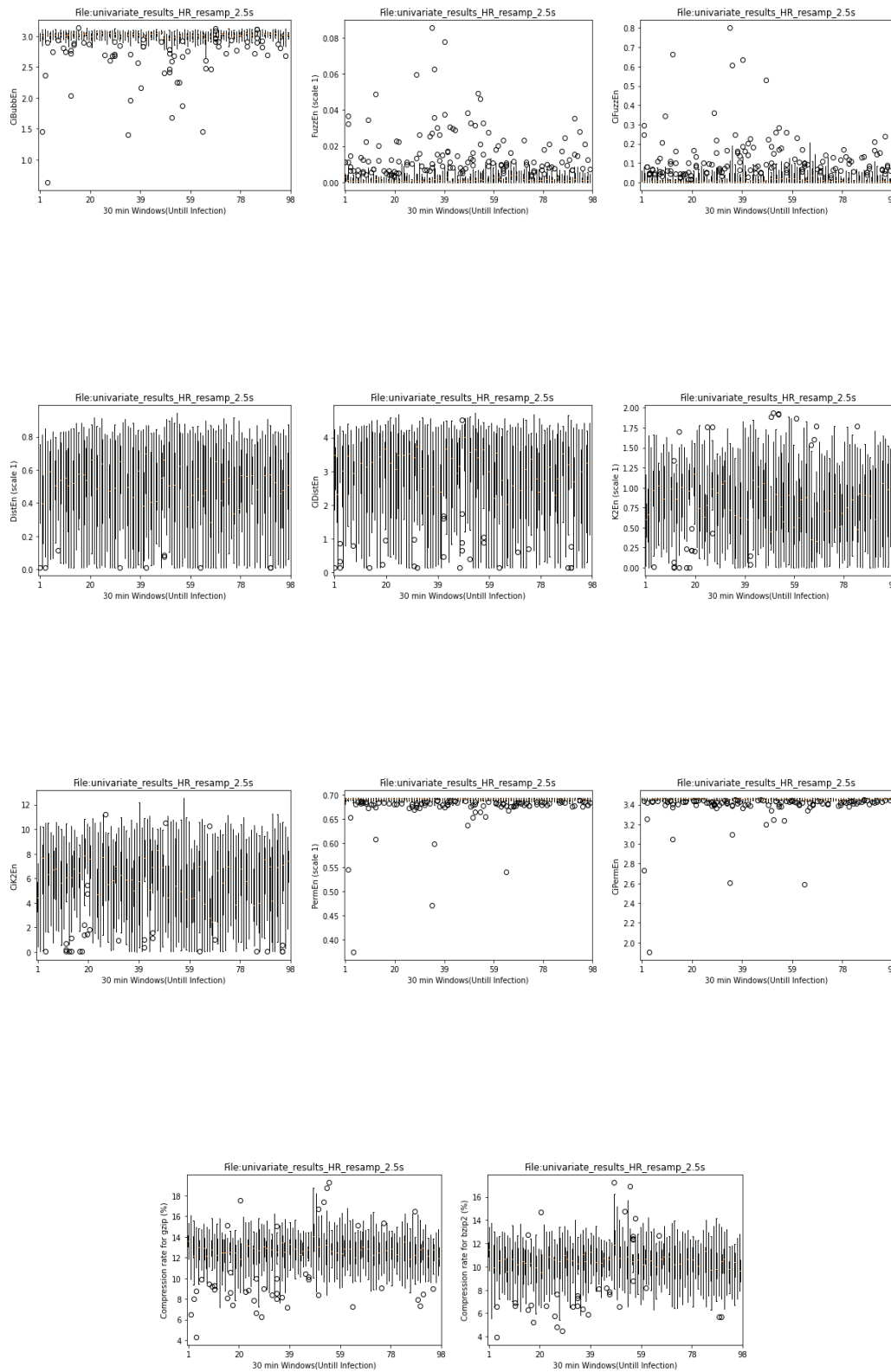


A.2 Boxplots for the Univariate measures for Heart signal

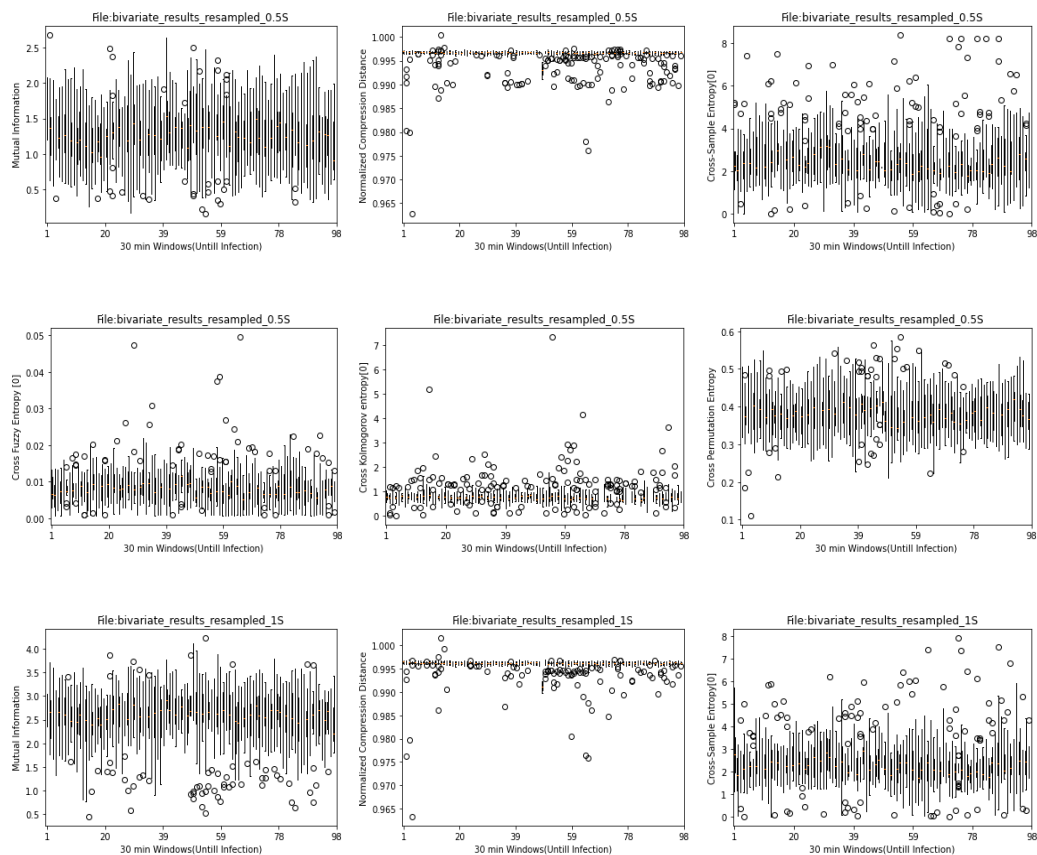


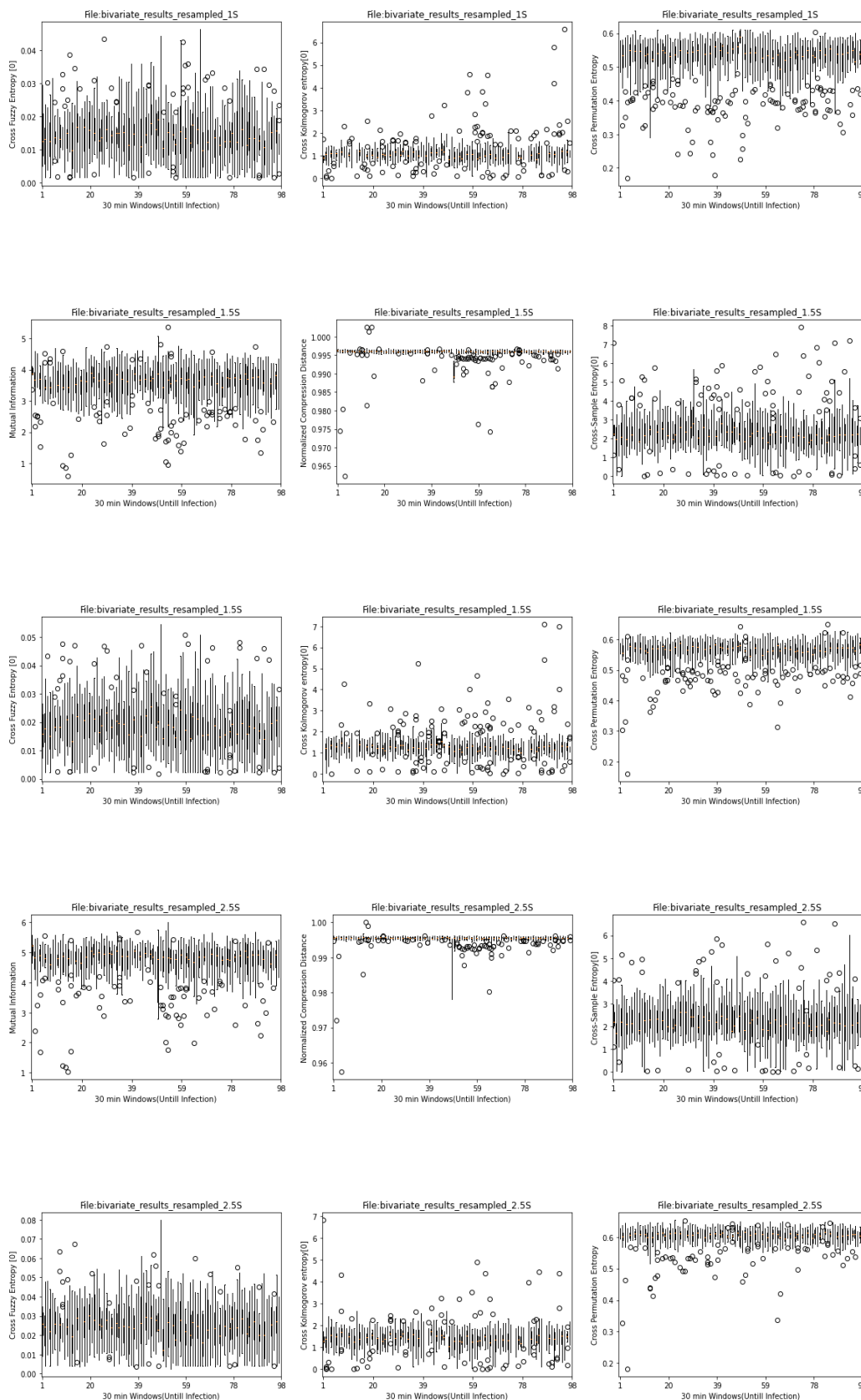






A.3 Boxplots for the Bivariate measures

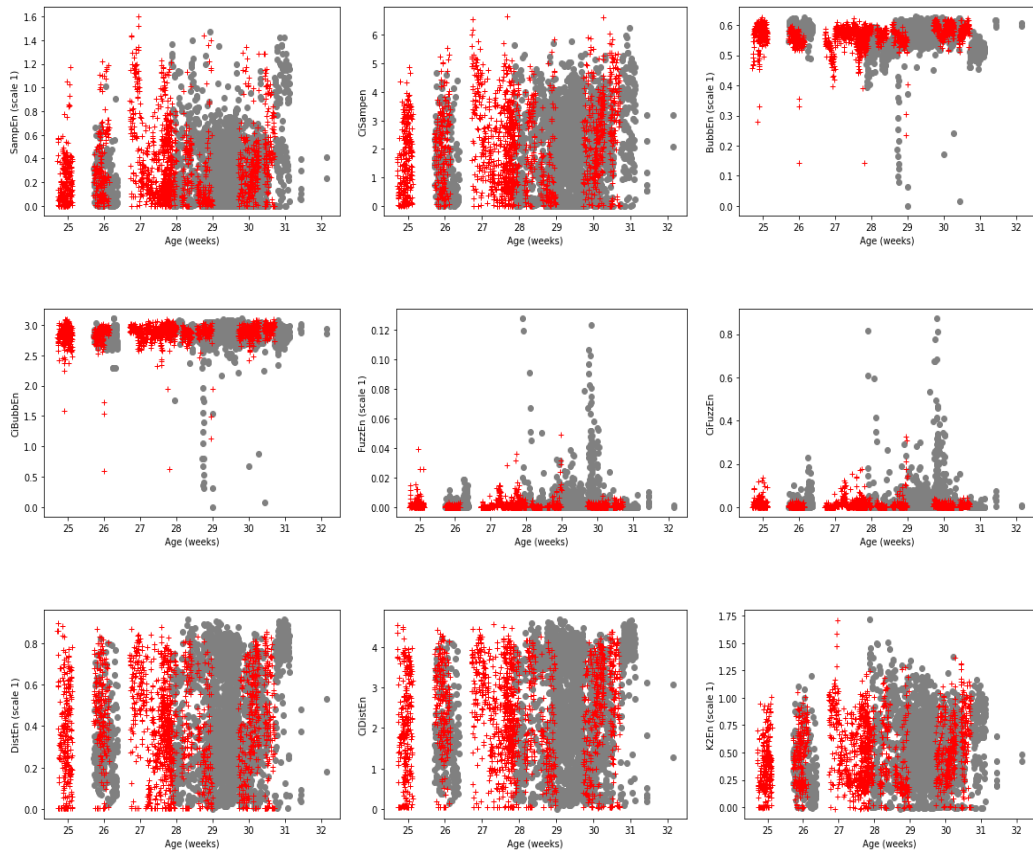


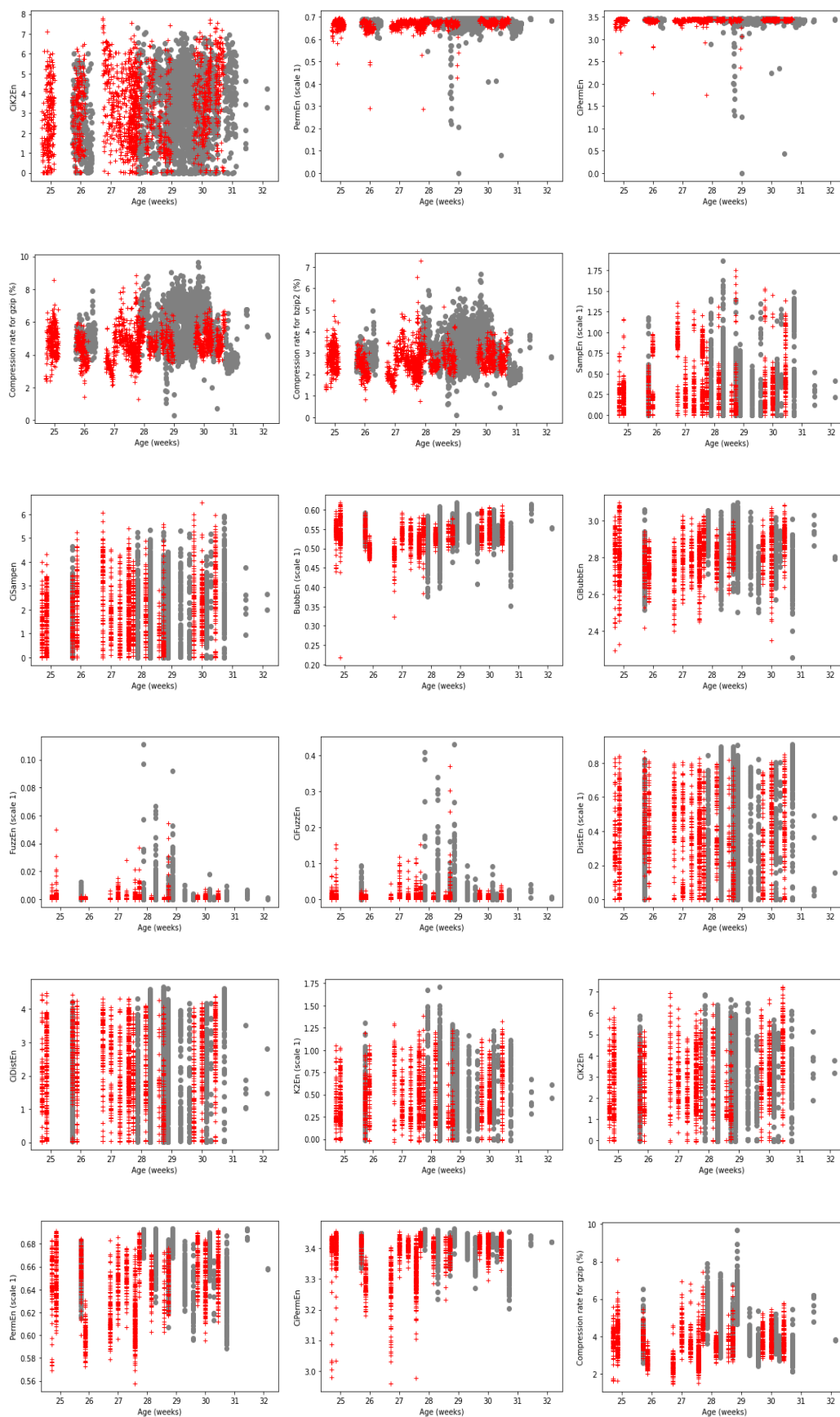


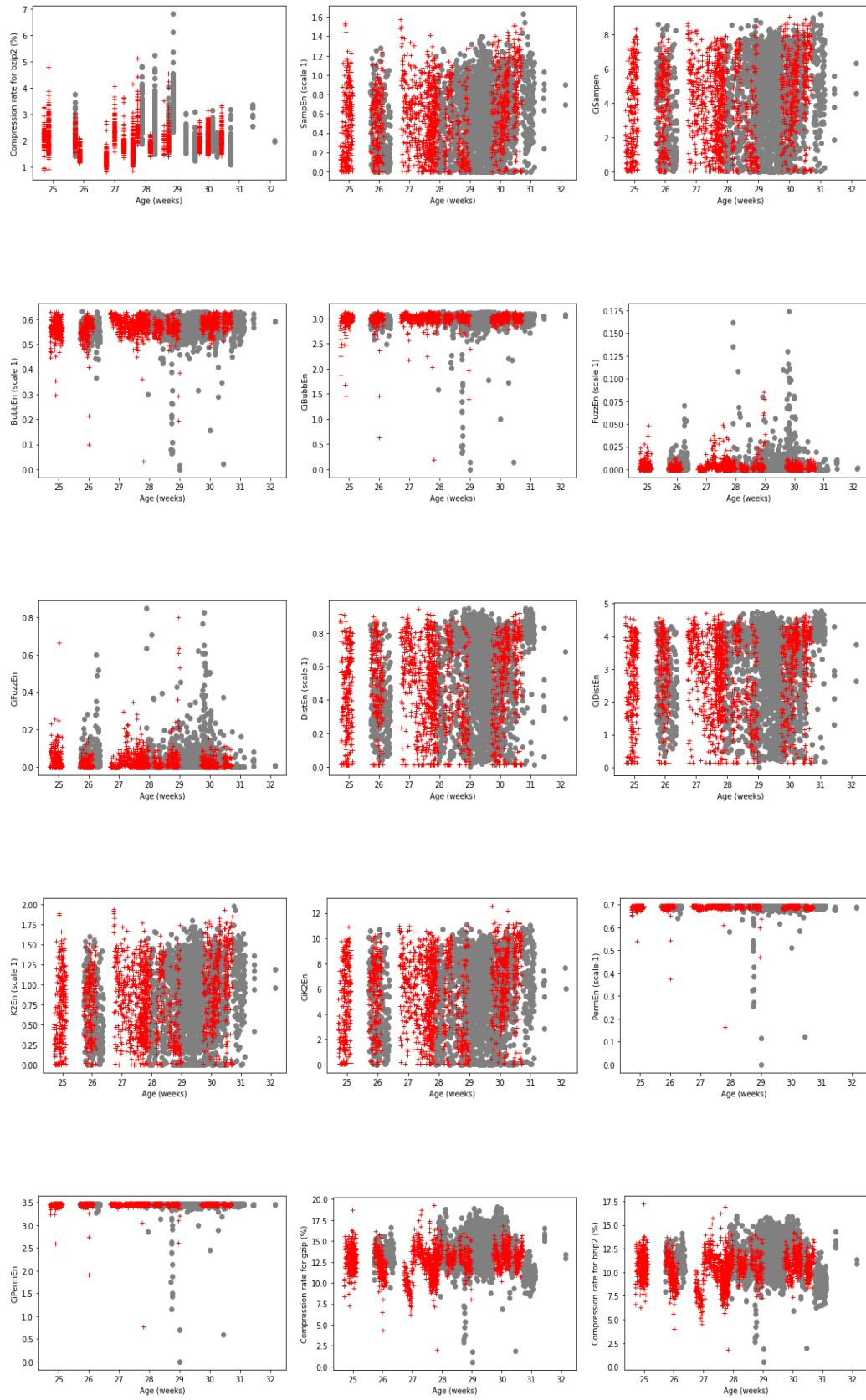
Appendix B

Control and Infected Subjects

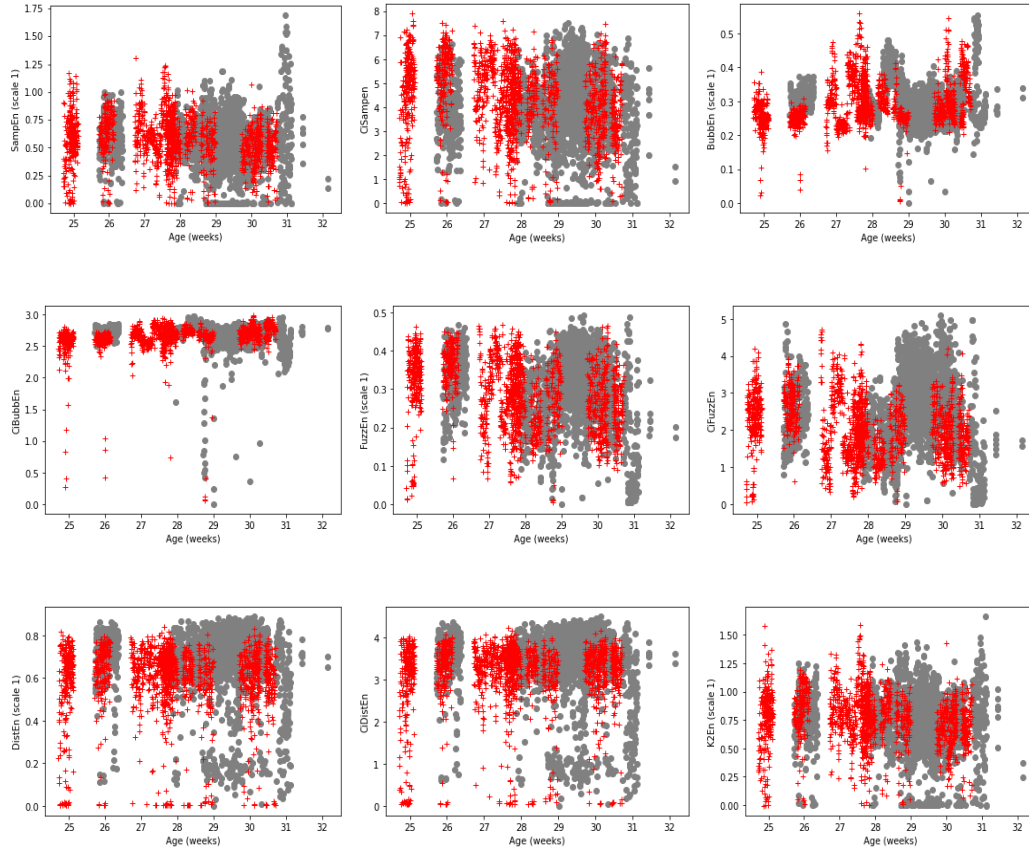
B.1 Control and Infected Subjects for the Heart Rate signal

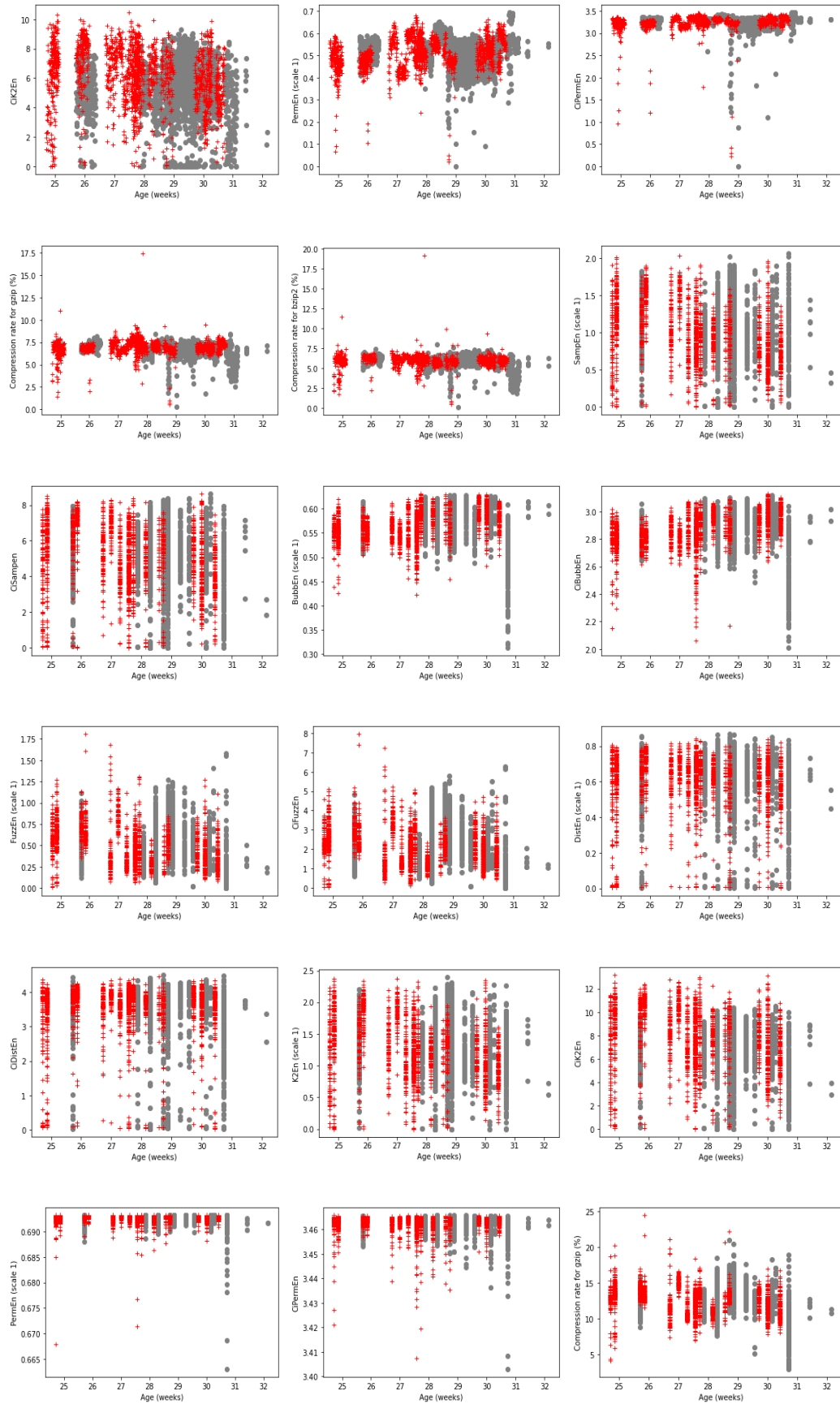


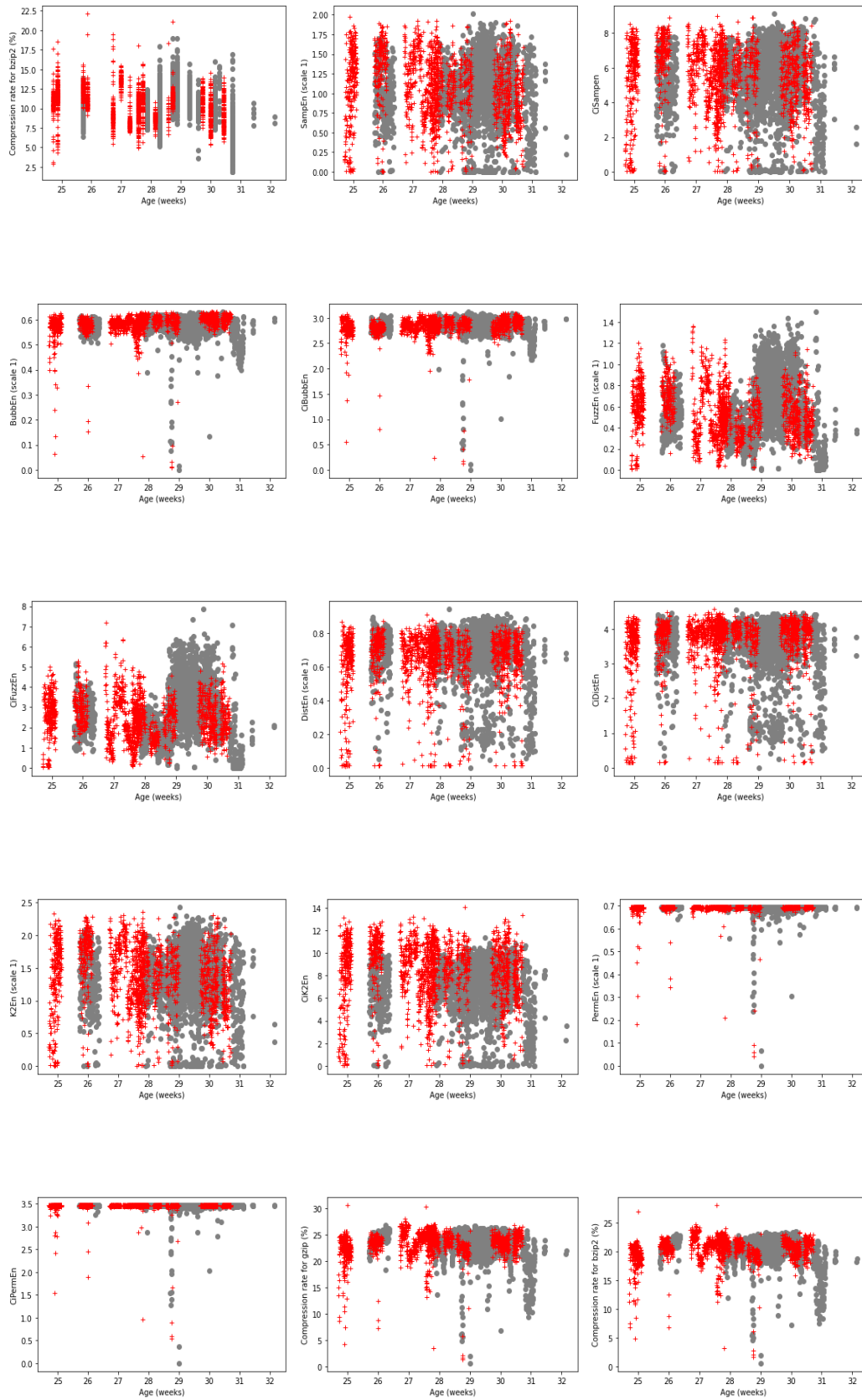




B.2 Control and Infected Subjects for the Respiratory Rate signal







Appendix C

Mann-Whitney test p-values

Table C.1: P-values for the measures of the original Heart Rate signal.

Measure	P-value
SampEn (scale 1)	0.6189130794296036
SampEn (scale 2)	0.6189130794296036
SampEn (scale 3)	0.6988598802286865
SampEn (scale 4)	0.9559267802807303
SampEn (scale 5)	0.6717818635319417
CiSampEn	0.6451232285407815
BubbEn (scale 1)	0.926611028101784
BubbEn (scale 2)	0.8683166060584981
BubbEn (scale 3)	0.5679460944657957
BubbEn (scale 4)	0.8394156924542152
BubbEn (scale 5)	0.8683166060584981
CiBubbEn	0.6717818635319417
FuzzEn (scale 1)	1.0
FuzzEn (scale 2)	0.5679460944657957
FuzzEn (scale 3)	0.5190784427314943
FuzzEn (scale 4)	0.5432386742490749
FuzzEn (scale 5)	0.5931788243337852
CiFuzzEn	0.5432386742490749
DistEn (scale 1)	0.9853022792160973
DistEn (scale 2)	0.8107297271774061
DistEn (scale 3)	0.8107297271774061
DistEn (scale 4)	0.8683166060584981
DistEn (scale 5)	0.8683166060584981
CiDistEn	0.8107297271774061
K2En (scale 1)	0.3666993351601333
K2En (scale 2)	0.6451232285407815
K2En (scale 3)	0.6988598802286865
K2En (scale 4)	0.9559267802807303
K2En (scale 5)	0.6717818635319417
CiK2En	0.34746533694432824
PermEn (scale 1)	0.11738098195165184
PermEn (scale 2)	0.9853022792160973
PermEn (scale 3)	0.10899967223244765
PermEn (scale 4)	0.02581055942790341
PermEn (scale 5)	0.02581055942790341
CiPermEn	0.5190784427314943
Compression rate for gzip (%)	0.12625999116252323
Compression rate for bzip2 (%)	0.34746533694432824

Table C.2: P-values for the measures of the Heart Rate signal, resampled for 2Hz.

Measure	P-value
SampEn (scale 1)	0.16708142496595368
SampEn (scale 2)	0.2611250000774915
SampEn (scale 3)	0.6717818635319417
SampEn (scale 4)	0.6451232285407815
SampEn (scale 5)	0.7541497213237057
CiSampEn	0.31096240865000846
BubbEn (scale 1)	0.5931788243337852
BubbEn (scale 2)	0.9853022792160973
BubbEn (scale 3)	0.6988598802286865
BubbEn (scale 4)	0.8683166060584981
BubbEn (scale 5)	0.5931788243337852
CiBubbEn	0.40711253189153784
FuzzEn (scale 1)	0.6451232285407815
FuzzEn (scale 2)	0.5679460944657957
FuzzEn (scale 3)	0.6189130794296036
FuzzEn (scale 4)	0.6988598802286865
FuzzEn (scale 5)	0.6717818635319417
CiFuzzEn	0.5432386742490749
DistEn (scale 1)	0.8683166060584981
DistEn (scale 2)	0.926611028101784
DistEn (scale 3)	0.7541497213237057
DistEn (scale 4)	0.8973945839305384
DistEn (scale 5)	0.7541497213237057
CiDistEn	0.8394156924542152
K2En (scale 1)	0.3666993351601333
K2En (scale 2)	0.32888566671710195
K2En (scale 3)	0.31096240865000846
K2En (scale 4)	0.5931788243337852
K2En (scale 5)	0.6189130794296036
CiK2En	0.4282779787264367
PermEn (scale 1)	0.4282779787264367
PermEn (scale 2)	0.3865836974872757
PermEn (scale 3)	0.09366120252602836
PermEn (scale 4)	0.12625999116252323
PermEn (scale 5)	0.1786901758210303
CiPermEn	0.6717818635319417
Compression rate for gzip (%)	0.06279895809361855
Compression rate for bzip2 (%)	0.15604876481480082

Table C.3: P-values for the measures of the Heart Rate signal, resampled for 0.4Hz.

Measure	P-value
SampEn (scale 1)	0.5432386742490749
SampEn (scale 2)	0.4282779787264367
SampEn (scale 3)	0.4954853264442981
SampEn (scale 4)	0.5190784427314943
SampEn (scale 5)	0.47247726377686594
CiSampEn	0.3865836974872757
BubbEn (scale 1)	1.0
BubbEn (scale 2)	0.34746533694432824
BubbEn (scale 3)	0.5190784427314943
BubbEn (scale 4)	0.9853022792160973
BubbEn (scale 5)	0.926611028101784
CiBubbEn	0.4282779787264367
FuzzEn (scale 1)	0.5931788243337852
FuzzEn (scale 2)	0.6988598802286865
FuzzEn (scale 3)	0.8973945839305384
FuzzEn (scale 4)	0.7822957315881272
FuzzEn (scale 5)	0.6451232285407815
CiFuzzEn	0.8394156924542152
DistEn (scale 1)	0.9559267802807303
DistEn (scale 2)	0.45007018084347294
DistEn (scale 3)	1.0
DistEn (scale 4)	0.926611028101784
DistEn (scale 5)	0.9853022792160973
CiDistEn	0.9853022792160973
K2En (scale 1)	0.6189130794296036
K2En (scale 2)	0.5679460944657957
K2En (scale 3)	0.47247726377686594
K2En (scale 4)	0.40711253189153784
K2En (scale 5)	0.47247726377686594
CiK2En	0.4282779787264367
PermEn (scale 1)	0.23114144867966557
PermEn (scale 2)	0.11738098195165184
PermEn (scale 3)	0.8107297271774061
PermEn (scale 4)	0.8973945839305384
PermEn (scale 5)	0.7822957315881272
CiPermEn	0.1356534957111977
Compression rate for gzip (%)	0.11738098195165184
Compression rate for bzip2 (%)	0.31096240865000846

Table C.4: P-values for the measures of the original Respiratory Rate signal.

Measure	P-value
SampEn (scale 1)	0.23114144867966557
SampEn (scale 2)	0.10109889531396785
SampEn (scale 3)	0.10899967223244765
SampEn (scale 4)	0.11738098195165184
SampEn (scale 5)	0.019305348578232146
CiSampEn	0.10109889531396785
BubbEn (scale 1)	0.32888566671710195
BubbEn (scale 2)	0.47247726377686594
BubbEn (scale 3)	0.6988598802286865
BubbEn (scale 4)	0.16708142496595368
BubbEn (scale 5)	0.9559267802807303
CiBubbEn	0.45007018084347294
FuzzEn (scale 1)	0.053076273984620066
FuzzEn (scale 2)	0.057767777407812256
FuzzEn (scale 3)	0.0866689484771263
FuzzEn (scale 4)	0.19088853079732504
FuzzEn (scale 5)	0.27708441580501053
CiFuzzEn	0.09366120252602836
DistEn (scale 1)	0.03737216049555432
DistEn (scale 2)	0.09366120252602836
DistEn (scale 3)	0.2611250000774915
DistEn (scale 4)	0.29369582403333283
DistEn (scale 5)	0.9853022792160973
CiDistEn	0.23114144867966557
K2En (scale 1)	0.19088853079732504
K2En (scale 2)	0.014281134106780731
K2En (scale 3)	0.012884099769328262
K2En (scale 4)	0.023457868683019842
K2En (scale 5)	0.0341311491415566
CiK2En	0.015810132601942216
PermEn (scale 1)	0.27708441580501053
PermEn (scale 2)	0.9853022792160973
PermEn (scale 3)	0.19088853079732504
PermEn (scale 4)	0.1356534957111977
PermEn (scale 5)	0.5679460944657957
CiPermEn	0.3865836974872757
Compression rate for gzip (%)	0.9853022792160973
Compression rate for bzip2 (%)	0.9853022792160973

Table C.5: P-values for the measures of the Respiratory Rate signal, resampled for 2Hz.

Measure	P-value
SampEn (scale 1)	0.021293594726057574
SampEn (scale 2)	0.03113346524167029
SampEn (scale 3)	0.03737216049555432
SampEn (scale 4)	0.053076273984620066
SampEn (scale 5)	0.06279895809361855
CiSampEn	0.02836460111285712
BubbEn (scale 1)	0.5679460944657957
BubbEn (scale 2)	0.7541497213237057
BubbEn (scale 3)	0.7263265699621655
BubbEn (scale 4)	0.5432386742490749
BubbEn (scale 5)	0.45007018084347294
CiBubbEn	0.29369582403333283
FuzzEn (scale 1)	0.06818709122848478
FuzzEn (scale 2)	0.06279895809361855
FuzzEn (scale 3)	0.1786901758210303
FuzzEn (scale 4)	0.0866689484771263
FuzzEn (scale 5)	0.31096240865000846
CiFuzzEn	0.14557782852935333
DistEn (scale 1)	0.03737216049555432
DistEn (scale 2)	0.048707448597224035
DistEn (scale 3)	0.12625999116252323
DistEn (scale 4)	0.19088853079732504
DistEn (scale 5)	0.1356534957111977
CiDistEn	0.08010437235277022
K2En (scale 1)	0.02581055942790341
K2En (scale 2)	0.021293594726057574
K2En (scale 3)	0.02581055942790341
K2En (scale 4)	0.02836460111285712
K2En (scale 5)	0.02836460111285712
CiK2En	0.019305348578232146
PermEn (scale 1)	0.45007018084347294
PermEn (scale 2)	0.5931788243337852
PermEn (scale 3)	0.926611028101784
PermEn (scale 4)	0.2611250000774915
PermEn (scale 5)	0.0341311491415566
CiPermEn	0.5931788243337852
Compression rate for gzip (%)	0.34746533694432824
Compression rate for bzip2 (%)	0.6451232285407815

Table C.6: P-values for the measures of the Respiratory Rate signal, resampled for 0.4Hz.

Measure	P-value
SampEn (scale 1)	0.012884099769328262
SampEn (scale 2)	0.10109889531396785
SampEn (scale 3)	0.014281134106780731
SampEn (scale 4)	0.04087149771644004
SampEn (scale 5)	0.0341311491415566
CiSampEn	0.017481290792834652
BubbEn (scale 1)	0.32888566671710195
BubbEn (scale 2)	0.45007018084347294
BubbEn (scale 3)	0.5931788243337852
BubbEn (scale 4)	0.8973945839305384
BubbEn (scale 5)	0.6717818635319417
CiBubbEn	0.9853022792160973
FuzzEn (scale 1)	0.1356534957111977
FuzzEn (scale 2)	0.6451232285407815
FuzzEn (scale 3)	0.9853022792160973
FuzzEn (scale 4)	0.9559267802807303
FuzzEn (scale 5)	0.7822957315881272
CiFuzzEn	0.8683166060584981
DistEn (scale 1)	0.006768599153540975
DistEn (scale 2)	0.053076273984620066
DistEn (scale 3)	0.31096240865000846
DistEn (scale 4)	0.5679460944657957
DistEn (scale 5)	0.15604876481480082
CiDistEn	0.16708142496595368
K2En (scale 1)	0.06279895809361855
K2En (scale 2)	0.08010437235277022
K2En (scale 3)	0.019305348578232146
K2En (scale 4)	0.15604876481480082
K2En (scale 5)	0.06818709122848478
CiK2En	0.10899967223244765
PermEn (scale 1)	0.11738098195165184
PermEn (scale 2)	0.6189130794296036
PermEn (scale 3)	0.7263265699621655
PermEn (scale 4)	0.9559267802807303
PermEn (scale 5)	0.5432386742490749
CiPermEn	0.8683166060584981
Compression rate for gzip (%)	0.21710323794272546
Compression rate for bzip2 (%)	0.15604876481480082

References

- [1] Bzip2. Publisher: WordPress <http://www.bzip.org/docs.html> Accessed on 2023-04-15.
- [2] How gzip compression works – Raul Fraile. <https://2014.jsconf.eu/speakers/raul-fraile-how-gzip-compression-works.html>. Accessed on 2023-04-15.
- [3] Lzma - compression using the lzma algorithm. <https://docs.python.org/3/library/lzma.html#>. Accessed on 2023-04-15.
- [4] Sci-kit learn mutual information. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mutual_info_score.html. Accessed on 2023-04-16.
- [5] Scikit-learn. <https://pypi.org/project/EntropyHub/>. Accessed on 2023-04-16.
- [6] What is lzma? <https://www.winzip.com/en/learn/tips/what-is-lzma/>. Accessed on 2023-04-15.
- [7] What is gzip compression: Enabling gzip commands: Cdn guide: Imperva. <https://www.imperva.com/learn/performance/gzip/>, Sep 2019. Accessed on 2023-04-15.
- [8] Gangaram Akangire, Elizabeth Simpson, Julie Weiner, Janelle Noel-MacDonnell, Joshua Petrikin, and Michael Sheehan. Implementation of the neonatal sepsis calculator in early-onset sepsis and maternal chorioamnionitis. *Advances in Neonatal Care*, 2020.
- [9] Farah Q AL-Khalidi, Reza Saatchi, Derek Burke, H Elphick, and Stephen Tan. Respiration rate monitoring methods: A review. *Pediatric pulmonology*, 46(6):523–529, 2011.
- [10] S. Al-sharhan, F. Karray, W. Gueaieb, and O. Basir. Fuzzy entropy: a brief survey. In *10th IEEE International Conference on Fuzzy Systems. (Cat. No.01CH37297)*, volume 3, pages 1135–1139 vol.2, 2001.
- [11] Diogo Azevedo, Ana Maria Rodrigues, Helena Canhão, Alexandra M. Carvalho, and André Souto. Zgli: A pipeline for clustering by compression with application to patient stratification in spondyloarthritis. *Sensors*, 23(3), 2023.
- [12] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
- [13] PA Bromiley, NA Thacker, and E Bouhova-Thacker. Shannon entropy, renyi entropy, and information. *Statistics and Inf. Series (2004-004)*, 9:10–42, 2004.
- [14] Sumika Chauhan, Manmohan Singh, and Ashwani Kumar Aggarwal. An effective health indicator for bearing using corrected conditional entropy through

- diversity-driven multi-parent evolutionary algorithm. *Structural Health Monitoring*, 20(5):2525–2539, 2021.
- [15] Madalena Costa, Ary L Goldberger, and C-K Peng. Multiscale entropy analysis of biological signals. *Physical review E*, 71(2):021906, 2005.
- [16] Michelle A Cretikos, Rinaldo Bellomo, Ken Hillman, Jack Chen, Simon Finfer, and Arthas Flabouris. Respiratory rate: the neglected vital sign. *Medical Journal of Australia*, 188(11):657–659, 2008.
- [17] Maj Marla J De Jong and David C Randall. Heart rate variability analysis in the assessment of autonomic function in heart failure. *Journal of Cardiovascular Nursing*, 20(3):186–195, 2005.
- [18] FA de Oliveira Júnior, RA Pereira, AS Silva, JL de Brito Alves, JH Costa-Silva, VA Braga, and CM Balarini. Different acquisition systems for heart rate variability analysis may lead to diverse outcomes. *Brazilian Journal of Medical and Biological Research*, 55, 2022.
- [19] Barry De Ville. Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(6):448–455, 2013.
- [20] Stephan Dreiseitl and Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5-6):352–359, 2002.
- [21] Gabriel J. Escobar, Karen M. Puopolo, Soora Wi, Benjamin J. Turk, Michael W. Kuzniewicz, Eileen M. Walsh, Thomas B. Newman, John Zupancic, Ellice Lieberman, and David Draper. Stratification of Risk of Early-Onset Sepsis in Newborns 34 Weeks’ Gestation. *Pediatrics*, 133(1):30–36, 01 2014.
- [22] Luca Faes and Alberto Porta. *Conditional Entropy-Based Evaluation of Information Dynamics in Physiological Systems*, pages 61–86. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.
- [23] Carolin Fleischmann, Felix Reichert, Alessandro Cassini, Rosa Horner, Thomas Harder, Robby Markwart, Marc Tröndle, Yoanna Savova, Niranjana Kissoon, Peter Schlattmann, et al. Global incidence and mortality of neonatal sepsis: a systematic review and meta-analysis. *Archives of Disease in Childhood*, 106(8):745–752, 2021.
- [24] Matthew W Flood. THE ENTROPYHUB GUIDE A user manual for the EntropyHub toolkit. 2, 2021.
- [25] Matthew W Flood and Bernd Grimm. EntropyHub: An open-source toolkit for entropic time series analysis. 16(11), November 2021.
- [26] Jerome H. Friedman and Peter Hall. On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, 137:669–683, 2007.
- [27] Tsehaynesh Geyesus, Feleke Moges, Setegn Eshetie, Biruk Yeshitela, and Ebba Abate. Bacterial etiologic agents causing neonatal sepsis and associated risk factors in gondar, northwest ethiopia. *BMC pediatrics*, 17:1–10, 2017.
- [28] Dž. Gojak, K. Gvožđar, Z. Hećimović, A. Smajović, E. Bečić, A. Deumić, L. Spahić Bećirović, L. Gurbeta Pokvić, and A. Badnjević. The use of artificial intelligence in the diagnosis of neonatal sepsis. *IFAC-PapersOnLine*, 55(4):62–67, 2022. 17th IFAC Conference on Programmable Devices and Embedded Systems PDES 2022 — Sarajevo, Bosnia and Herzegovina, 17-19 May 2022.

- [29] A L Goldberger, L A Amaral, L Glass, J M Hausdorff, P C Ivanov, R G Mark, J E Mietus, G B Moody, C K Peng, and H E Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–20, June 2000.
- [30] Peter Grassberger and Itamar Procaccia. Estimation of the kolmogorov entropy from a chaotic signal. *Phys. Rev. A*, 28:2591–2593, Oct 1983.
- [31] Sasha Hallett, Fadi Toro, and John V. Ashurst. *Physiology, Tidal Volume*. StatPearls Publishing, Treasure Island (FL), 2022.
- [32] Clint Hansen, Qin Wei, Jiann-Shing Shieh, Paul Fourcade, Brice Isableu, and Lina Majed. Sample entropy, univariate, and multivariate multi-scale entropy in comparison with classical postural sway parameters in young healthy adults. *Frontiers in human neuroscience*, 11:206, 2017.
- [33] R Hayes, J Hartnett, G Semova, C Murray, K Murphy, L Carroll, H Plapp, L Hession, J O’Toole, D McCollum, et al. Infection, inflammation, immunology and immunisation (i4) section of the european society for paediatric research (espr). neonatal sepsis definitions from randomised clinical trials. *Pediatr Res*, 10, 2021.
- [34] David A Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.
- [35] Mohammed Babatunde IBRAHIM and Kazeem Alagbe GBOLAGADE. Performance comparison of huffman coding and lempel-ziv-welch text compression algorithms with chinese remainder theorem. *University of Pitești Scientific Bulletin Series: Electronics and Computer Science*, 19(2):7–12, 2019.
- [36] Kimihito Ito, Thomas Zeugmann, and Yu Zhu. Clustering the normalized compression distance for influenza virus data. In *Algorithms and Applications*, pages 130–146. Springer, 2010.
- [37] Antoine Jamin and Anne Humeau-Heurtier. (multiscale) cross-entropy methods: A review. *Entropy*, 22(1), 2020.
- [38] Holger Kantz and Thomas Schreiber. *Nonlinear time series analysis*, volume 7. Cambridge university press, 2004.
- [39] Selcan Kaplan Berkaya, Alper Kursat Uysal, Efnan Sora Gunal, Semih Ergin, Serkan Gunal, and M. Bilginer Gulmezoglu. A survey on ecg analysis. *Biomedical Signal Processing and Control*, 43, 2018.
- [40] P Kavitha. A survey on lossless and lossy data compression methods. *International Journal of Computer Science & Engineering Technology*, 7(03):110–114, 2016.
- [41] Marleen Kerste, Jellina Corver, Martine C Sonneveld, Monique van Brakel, Paul D van der Linden, Babette A M. Braams-Lisman, and Frans B Plötz. Application of sepsis calculator in newborns with suspected infection. *The Journal of Maternal-Fetal & Neonatal Medicine*, 29(23):3860–3865, 2016.
- [42] Peng Li, Ke Li, Chengyu Liu, Dingchang Zheng, Zong-Ming Li, and Changchun Liu. Detection of coupling in short physiological series by a joint distribution entropy method. *IEEE Transactions on Biomedical Engineering*, 63(11):2231–2242, 2016.
- [43] Peng Li, Chengyu Liu, Ke Li, Dingchang Zheng, Changchun Liu, and Yinglong Hou. Assessing the complexity of short-term heartbeat interval series by distribution entropy. *Medical & biological engineering & computing*, 53:77–87, 2015.

- [44] Thomas W. MacFarland and Jan M. Yates. *Mann–Whitney U Test*, pages 103–132. Springer International Publishing, Cham, 2016.
- [45] Subramani Mani, Asli Ozdas, Constantin Aliferis, Huseyin Atakan Varol, Qingxia Chen, Randy Carnevale, Yukun Chen, Joann Romano-Keeler, Hui Nian, and Jörn-Hendrik Weitkamp. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. *Journal of the American Medical Informatics Association*, 21(2):326–336, 2014.
- [46] George Manis, MD Aktaruzzaman, and Roberto Sassi. Bubble entropy: An entropy almost free of parameters. *IEEE Transactions on Biomedical Engineering*, 64(11):2711–2718, 2017.
- [47] R. J. McEliece. *Entropy and mutual information*, page 17–49. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2004.
- [48] Iva Mihatov Stefanovic. Neonatal sepsis. *Biochemia medica*, 21(3):276–281, 2011.
- [49] Alistair Moffat. Huffman coding. *ACM Computing Surveys (CSUR)*, 52(4):1–35, 2019.
- [50] James N Morgan and John A Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, 58, 1963.
- [51] Fahad B Mostafa and Md Easin Hasan. Machine learning approaches for binary classification to discover liver diseases using clinical data. 2021.
- [52] Catherine Muthoni Njeru, J Mark Ansermino, William M Macharia, and Dustin T Dunsmuir. Variability of respiratory rate measurements in neonates-every minute counts. *BMC pediatrics*, 22(1):1–6, 2022.
- [53] Ilkay Özmeral Odabaşı, Ali Bülbul, et al. Neonatal sepsis. *The Medical Bulletin of Sisli Etfal Hospital*, 2020.
- [54] Aakash Parmar, Rakesh Katariya, and Vatsal Patel. A review on random forest: An ensemble classifier. In Jude Hemanth, Xavier Fernando, Pavel Lafata, and Zubair Baig, editors, *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, pages 758–763, Cham, 2019. Springer International Publishing.
- [55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [56] Steve Pincus. Approximate entropy (apen) as a complexity measure. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 5(1):110–117, 1995.
- [57] Romain Pirracchio. Mortality prediction in the icu based on mimic-ii results from the super icu learner algorithm (sacula) project. *Secondary Analysis of Electronic Health Records*, pages 295–313, 2016.
- [58] Vili Podgorelec, Peter Kokol, Bruno Stiglic, and Ivan Rozman. Decision trees: an overview and their use in medicine. *Journal of medical systems*, 26(5):445–463, 2002.
- [59] Alberto Porta, Stefano Guzzetti, Nicola Montano, Raffaello Furlan, Massimo Pagani, Alberto Malliani, and Sergio Cerutti. Entropy, entropy rate, and pattern classification as tools to typify complexity in short heart period variability series. *IEEE Transactions on Biomedical Engineering*, 48(11):1282–1291, 2001.

- [60] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.
- [61] U Rajendra Acharya, K Paul Joseph, Natarajan Kannathal, Choo Min Lim, and Jasjit S Suri. Heart rate variability: a review. *Medical and biological engineering and computing*, 44(12):1031–1051, 2006.
- [62] Keerthi G Reddy, PA Vijaya, and S Suhasini. Ecg signal characterization and correlation to heart abnormalities. *International Research Journal of Engineering and Technology*, 4(5):1212–1216, 2017.
- [63] Ye Ren, Le Zhang, and P.N. Suganthan. Ensemble classification and regression-recent developments, applications and future directions [review article]. *IEEE Computational Intelligence Magazine*, 11(1):41–53, 2016.
- [64] S Renugadevi and PS Nithya Darisini. Huffman and lempel-ziv based data compression algorithms for wireless sensor networks. In *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*, pages 461–463. IEEE, 2013.
- [65] Maria Ribeiro, Luísa Castro, Guy Carrault, Patrick Pladys, Cristina Costa-Santos, and Teresa Henriques. Evolution of heart rate complexity indices in the early detection of neonatal sepsis. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 367–372. IEEE, 2022.
- [66] Maria Ribeiro, Luísa Castro, Guy Carrault, Patrick Pladys, Cristina Costa-Santos, and Teresa Henriques. Evolution of heart rate complexity indices in the early detection of neonatal sepsis. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 367–372, 2022.
- [67] Maria Ribeiro, Teresa Henriques, Luísa Castro, André Souto, Luís Antunes, Cristina Costa Santos, and Andreia Teixeira. The entropy universe. *Entropy*, 23:222, 02 2021.
- [68] Joshua S Richman and J Randall Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278(6):H2039–H2049, 2000.
- [69] Amand F. Schmidt and Chris Finan. Linear regression and the normality assumption. *Journal of Clinical Epidemiology*, 98:146–151, 2018.
- [70] A Shane, P Sánchez, and B Stoll. Sepsis neonatal. *Lancet*, 390(10104):1770–1780, 2017.
- [71] M Singer, MM Clifford S Deutschman, et al. Improving the prevention, diagnosis and clinical management of sepsis. *WHO [Internet]*, page 2017, 2017.
- [72] Muhammad Aliyu Sulaiman and Jane Labadin. Feature selection based on mutual information. In *2015 9th International Conference on IT in Asia (CITA)*, pages 1–6. IEEE, 2015.
- [73] Brynne A Sullivan and Karen D Fairchild. Vital signs as physiomarkers of neonatal sepsis. *Pediatric research*, 91, 2022.
- [74] Brynne A Sullivan, Sherry L Kausch, and Karen D Fairchild. Artificial and human intelligence for early identification of neonatal sepsis. *Pediatric Research*, pages 1–7, 2022.
- [75] Martin J Tobin. Respiratory monitoring in the intensive care unit¹². *Amer. Rev. Respiratory Disease*, 138(6):1625–1642, 1988.
- [76] MJ Tobin. Breathing pattern analysis. *Intensive care medicine*, 18(4):193–201, 1992.

- [77] Anton M. Unakafov and Karsten Keller. Conditional entropy of ordinal patterns. *Physica D: Nonlinear Phenomena*, 269:94–102, 2014.
- [78] Guido Van Rossum and Fred L Drake Jr. *Python tutorial*, volume 620. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.
- [79] Alpo Värrä, Antti Kallonen, Elina Helander, Andres Ledesma, and Patrick Pladys. The digi-newb project for preterm infant sepsis risk and maturity analysis. *Finnish Journal of eHealth and eWelfare*, 10(2-3):330–333, 2018.
- [80] Mohan Venkatesh, Angela Flores, Ruth Ann Luna, and James Versalovic. Molecular microbiological methods in the diagnosis of neonatal sepsis. *Expert review of anti-infective therapy*, 8(9):1037–1048, 2010.
- [81] P Vitanyi and R Cilibrasi. Clustering by compression. *IEEE Transactions on Information Theory*. v51 i4, pages 1523–1545, 2008.
- [82] Alpo Värrä, Antti Kallonen, Elina Helander, Andres Ledesma, and Patrick Pladys. The digi-newb project for preterm infant sepsis risk and maturity analysis. *Finnish Journal of eHealth and eWelfare*, 10:330–333, May 2018.
- [83] R Scott Watson, Joseph A Carcillo, Walter T Linde-Zwirble, Gilles Clermont, Jeffrey Lidicker, and Derek C Angus. The epidemiology of severe sepsis in children in the united states. *American journal of respiratory and critical care medicine*, 167(5):695–701, 2003.
- [84] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010.
- [85] Jiawei Yang, Gulraiz Iqbal Choudhary, Susanto Rahardja, and Pasi Franti. Classification of interbeat interval time-series using attention entropy. *IEEE Transactions on Affective Computing*, 2020.
- [86] George Yuan, Nicole A Drost, and R Andrew McIvor. Respiratory rate and breathing pattern. *McMaster Univ. Med. J*, 10(1):23–25, 2013.
- [87] Xiaohua Douglas Zhang. *Entropy for the Complexity of Physiological Signal Dynamics*, pages 39–53. Springer Singapore, Singapore, 2017.
- [88] Xin Zhang, Hong Xu, Lijun Zhou, Qi Cao, Qian Shen, Li Sun, Xiaoyan Fang, Wei Guo, Yihui Zhai, Jia Rao, et al. Accuracy of early dmsa scan for vur in young children with febrile uti. *Pediatrics*, 133(1):e30–e38, 2014.
- [89] Lulu Zhao, Licai Yang, Zhonghua Su, and Chengyu Liu. Cardiorespiratory coupling analysis based on entropy and cross-entropy in distinguishing different depression stages. *Frontiers in physiology*, 10:359, 2019.
- [90] H.-J. Zimmermann. Fuzzy set theory. *WIREs Computational Statistics*, 2(3):317–332, 2010.
- [91] Xiaonan Zou, Yong Hu, Zhewen Tian, and Kaiyuan Shen. Logistic regression model optimization and case analysis. In *2019 IEEE 7th international conference on computer science and network technology (ICCSNT)*, pages 135–139. IEEE, 2019.