Ecole Polytechnique Federale de Lausanne
School of Life Sciences

Master's project in Life Sciences and Technology

Evaluation and extension of an automated system to diagnose the health of Premature Babies

Carried out in the Institute for Adaptive and Neural Computation
At Edinburgh University's School of Informatics
Under the Supervision of Professor Chris Williams

Camille Joggi

Under the direction of
Professor Wulfram Gerstner
In the laboratory of computational neuroscience

External expert: Professor Chris. Williams

Lausanne, EPFL

19th August 2011

Abstract

Premature babies are not fully equipped to deal with the outside world. Their immature bodies make them highly at risk to a large number of physiological complications. Therefore, while they are still immature, premature babies are taken care of in neonatology intensive care units, where they are kept in a controlled and protective environment. In these units, the babies' vital parameters, such as their heart rate and temperature, are tightly monitored. The devices used to record this data are fitted with alarms that are triggered automatically when something wrong is detected [1]. However, the alarm systems currently used are often unable to discriminate between real physiological problems and artefacts that should be ignored (for example, when a measurement device drops-out). As a result, these systems produce a high rate of false alarms, where the alarm is triggered when nothing is clinically wrong. This can have unwanted consequences: first, it creates a noisy environment, which may disturb the babies; second, it can lead to the alarm being ignored when there is actually something wrong. In order to address this problem, my host laboratory has constructed a system that can automatically infer the state of health of premature babies in neonatology units, based on their recorded physiological data [2]. The new system is designed to be able to deal with artefactual changes in the measurements, hopefully resulting in a reduced number of false alarms.

In this thesis, I describe my contribution to this project, which is twofold. The first part deals with the problem of evaluating the performance of the system. To do this, I developed an online feedback process, where clinicians could provide information about the true clinical interpretation of the physiological data, which could be compared to the output of the system. The feedbacks provided by the clinicians were then analysed to investigate how the system could be improved. The second part of my thesis was motivated by early feedbacks from the clinicians, which indicated that the system was unable to deal with changes in the humidity measurements that occurred when the incubator door was opened and closed. To address this, I extended the current system, so that it was able to account for these changes, thereby increasing its capability to detect 'true' physiological problems during the period after the incubator door is closed.

Acknowledgement

I would like to thank my supervisor Chris Williams for mentoring me during this year, Ioan Stanculescu, for his help all along my project, John Quinn, Yvonne Freer, the transport fellows, and Neil McIntosh for their useful collaboration, and Matthew Chalk for proofreading this thesis.

Contents

Chapter 1

Introduction

Premature birth is an important issue, as it leads often to long term complications, such as neurological disabilities and cognitive dysfunctions [3]. Moreover, this affects a large proportion of the population (12-13% of births in the United States of America, and 5-9 % of births in many other industrialised countries [4]).

Before birth, the uterine environment plays a crucial role in protecting the foetus against temperature changes and physical contacts, as well as an essential role of assistance, as the mother provides him with food and oxygen. When the birth occurs too early, the baby, too immature to function independently, is taken care of in a neonatology unit, where the equipments used permit to reproduce the roles of protection and assistance normally played by the mother [5].

In neonatology units, the physiological data of the premature babies, such as heart rate, respiratory rate, blood pressure and temperature are recorded and displayed on monitors in order for the clinicians to keep control over their state of health. While normal states are characterised by haemostasis and stability in these physiological measurements, each particular condition or pathology produces a specific sequence of measurements which form a very characteristic pattern.

The recording devices currently used in neonatology units are fitted with detection systems which activate an alarm every time a problem is detected in the measurements. Generally, these systems are very basic, and thus often unable to distinguish between physiological problems and artefacts (caused for example by a probe drop-out). This can have harmful consequences: a large proportion of alarms are triggered when nothing is clinically wrong, meaning that real problems could easily be ignored by the clinical staff. Moreover, a neonatology unit must constitute a protected environment, in which the amount of noise should be kept to a minimum, because pre-term babies may be affected by noise generated by the high rate of alarms [6, 7].

In order to address this problem and decrease the rate of false alarms, my host laboratory, in the context of the 'Condition Monitoring in Premature Ba-

bies' project, developed a system permitting automatic inference of the baby's state of health in real time, based on the physiological traces recorded from the baby [8].

This system consists of a model that describes how the state of health of the baby (e.g. healthy / unhealthy) generates the recorded physiological measurements. This model is probabilistic, which permits to deal with noisy data and imperfect knowledge of the world [9]. The parameters of the model are learned directly from the recorded data, which ensures that the model predicts such physiological data as closely as possible. Finally, once the model is learned, it can be used to infer the real state of the baby based on the recordings - and trigger an alarm when there is a problem.

The probabilistic model used here is a Factorial Switching Linear Dynamical System (FSLDS) [10]. This model describes complex time-series which are not well described globally by a single type of dynamics, but are better described by 'jumps' from one regime to another [9].

Physiological data monitored in neonatology units follow this kind of pattern, with data following one particular type of dynamics when the baby is healthy, and instantaneously switching to a different type of dynamics when a physiological problem occurs. Specifically, for the FSLDS model described here, in addition to the 'normal' regime, some 'factors' are explicitly included, which relate to artefact events - e.g. probe drop-out - or physiological events - e.g. bradycardia - affecting the measurements. The FSLDS model is used to infer which set of factors are responsible for the values displayed currently on the monitor. An additional factor (the 'X-factor') accounts for any novel conditions which cannot be explained by the normal regime or by the factors included in the system.

The addition of this X-factor has two benefits. Firstly, because it is impossible to model all the possible regimes encountered in pre-term babies, the X-factor ensures that an alarm will be raised whenever an abnormal type of dynamics is occurring, even though the system can not attribute this new type of dynamics to one of the factors included in the system. Secondly, if a data segment triggers the X-factor, this means that the current model is not able to explain this particular sequence of data, and must be improved.

The FSLDS system developed by my host laboratory has been successful in inferring the different artefact and physiological conditions included as factors in the system. Moreover, the X-factor has shown efficiency in inferring novel conditions [10]. However, the factors currently part of the system have been evaluated only in a limited amount of physiological data (15 babies). An assessment of the performance of the model on new monitoring data could help to give a better evaluation of the system. Moreover, the number of factors currently part of the system is very limited. Further analysis of the segments of physiological data that have triggered the X-factor could show up additional non-normal regimes in the data that could be added to the system as new factors, and therefore used to improve the current model. Indeed, initial analysis of the X-factor, conducted previously, has already highlighted imperfections of the system, indicating that there was a need for re-modelling the effect of opening

the incubator.

In my project, I will address the issues mentioned above by undertaking an analysis of intervals of new physiological data that trigger the X-factor, in order to evaluate the existing factors, and discover clinically significant non-normal regimes. The results will then be analysed and an eventual addition of new factors in the model will be discussed. Finally, I will re-model the factor that has been discovered not to be good enough to explain the data.

1.1 Outline

Chapter 2 is dedicated to presenting the medical background of the project. Specifically, I will give an overview of the main characteristics of premature babies' physiology, some common clinical manifestations following from these characteristics, and condition monitoring in neonatology units.

In Chapter 3, I describe the mathematical background of the project. The first section is devoted to giving a general introduction on time-series probabilistic models. In the second section, I will describe the main aspects of the FSLDS model developed by my host laboratory in the framework of the 'Condition Monitoring in Premature Babies' project.

Chapter 4 is dedicated to presenting the first part of my contribution to this project, which consists of an evaluation of the system developed by my host laboratory. This evaluation is accomplished through an analysis of feedbacks of clinicians on intervals highlighted by the X-factor. I will describe the web-form applications developed in order to gather these feedbacks, as well as the results obtained from the analysis of this collection of feedbacks.

Chapter 5 describes the process followed for re-modelling the incubator open (or handling) factor, that has been discovered not to be good enough to explain the data. Notably, I show how I found the parameters, evaluated the relevance of the model, integrated the model in the overall system and assessed the performance of this re-modelled factor.

The last chapter concludes by presenting the most important findings of this project, and discussing some directions for future work.

Chapter 2

Premature babies: clinical overview

Premature babies are typically highly immature, and unprepared for the real world. Before birth, the uterus protects the baby from temperature changes and physical contact, while the mother provides him with food and oxygen. In contrast, after they are born, the baby must breathe alone, is subjected to changes in the external temperature, and must use their digestive system to get nutrients from food. That is, they are suddenly placed into an environment which requires their organs to function in an independent way, which they are still incapable of doing. Because of the immaturity of their organs and the instability of their homoeostatic control systems [11], premature babies have often difficulties to breathe, to suck and to control their body temperature [12].

To help them deal with this, premature babies are taken care of in neonatal units, where the equipment used permits achievement of the following goals. First, pre-term babies are placed in incubators, a device that tend to recreate the protective environment of the womb, in order to prevent any injury due to the outside world, by controlling tightly many factors like temperature. Secondly, since many of the body's organs function differently in the womb than in the outside world, the neonatal equipments help the baby to make the switch that the organs are not ready to make on their own.

Keeping track of heart beat, breathing rate, blood pressure and temperature is crucial to keep control of the health of premature babies. Indeed, from observing these recordings, it is possible to determine the state of the baby [1]. While normal states appear highly stable, abnormal conditions or pathologies are accompanied by specific recognised changes in the physiological recordings (patterns). [10].

The monitors currently used in neonatal units are fitted with simple detection systems, which activate an alarm every time a problem is detected in the measurements. Generally, these detection systems are very basic, with the alarm being triggered when a particular signal goes above or below preset thresholds

[1]. However, the observations on the monitor do not depend only on the physiological state of the baby. They can also be influenced by noise, due to inaccuracy of the probes, or artefacts, such as when the probe drops out [10]. Therefore, the alarms often sound when nothing is clinically wrong, which can have unwanted consequences. Firstly, since the alarms sound all the time, this could prevent them from being taken seriously by the clinicians, meaning that a real problem is more likely to be ignored, or only attended to after a dangerously long response time [13]. Moreover, in the neonatal unit, sound must be kept to a minimum, since extra noise (as well as additional disturbances caused when the clinicians respond to an alarm), could have adverse effects on the baby's health, for example, by inducing cardiorespiratory instability [6].

In order to address this problem and reduce the rate of false alarms, my host laboratory, in the context of the 'Condition Monitoring in Premature Babies' project [8], designed a detection system, which would be able to trigger an alarm when the health of the baby is in danger, while preventing the alarm from being triggered when changes in the measurements are unrelated to the real state of health of the baby, but are caused by an artefact. In order to do that, this system must be able to recognize, from the vital parameters recorded from the baby (e.g. heart rate, oxygen saturation, blood pressure), if the baby is in a healthy state, in an unhealthy state (e.g. bradycardia, oxygen desaturation), or if an artefact is happening to one or more of the probes (e.g. probe drop-out). To understand the design of this system, and thus, understand how the clinical conditions influence the readings on the monitor, it is essential to be acquainted with the most common clinical conditions encountered in premature babies. It is also crucial to understand the principles underlying the function of the probes used to record the vital parameters from the baby, to be able to comprehend how these readings can be altered by artefacts. Therefore, a basic introduction on the premature babies ' physiology, main clinical conditions and monitoring in neonatology is given in this chapter.

Section 2.1 describes the physiological characteristics of premature babies, highlighting the high immaturity of their organs in comparison with adults. Section 2.2 discusses the common clinical conditions that follow from this characteristic immaturity. Finally, I explain in section 2.3 how pre-term babies are monitored in neonatology units, and give some examples of artefacts that can alter the measurements.

2.1 Physiology

Physiology of premature babies differs from adult physiology, in that most of their organ systems are still undeveloped. This section focusses on the particular characteristics of neonatal and premature physiology that leads to the most common clinical manifestations in neonatology units.

Respiratory considerations One of the main characteristic of pre-term babies is the high immaturity of their respiration control system [14]. This im-

maturity can lead to several diseases or clinical manifestations. Because some of these clinical conditions are discussed in this project, a basic introduction of the pre-term respiratory system is given here.

The first goal of respiration is to convey oxygen to the tissues. Ventilation is the mechanical phenomenon that brings air (and therefore oxygen) into contact with blood in the pulmonary alveoli. This is achieved by an alternation between inspiration (where ambient air is added to an already present volume in the lungs) and expiration (where a part of the air in the alveoli is expelled) [15].

During respiratory stress, ventilation is regulated by varying the air flow entering and leaving the lungs, in order to ensure that the concentrations of O_2 and CO_2 in the blood remain as steady as possible. While healthy adults can vary their respiratory flow both by modifying their tidal volume (the amount of air inspired and expired during normal breathing; see Figure 2.1) and their respiratory rate [16], neonates can only vary the latter [17]. This is because the anatomic shape and the poorly developed muscles of the thorax reduce respiratory mechanical advantages [17], causing their vital capacity (the maximum amount of air they can expel from the lung after a maximal inspiration; see Figure 2.1) to be particularly small in relation to their body weight [11].

The functional residual capacity (FRC; the amount of air that remains in the lungs after a normal expiration; see Figure 2.1) is particularly small in neonates and pre terms - less than one half of an adults in relation to their body weight[11]. Because of their small FRC, neonates can only store a low amount of O_2 [18]. In healthy adults, O_2 stored in the lungs plays an important role in smoothing out the blood gas variations when not enough oxygen is available for ventilation, for instance if the respiration rate slows down. The small capacity of O_2 storage in pre-terms can lead to episodes of dramatic reduction of the amount of oxygen reaching the tissues (hypoxia) during periods of slowed respiratory rate or apnoea induced by the immaturity of the respiration control in premature babies. In extreme cases, the combination of small capacity of O_2 storage and high immaturity of respiratory control system can lead to a syndrome very common in pre-term babies named periodic breathing. This condition is characterised by periods of breathing and apnoea succeeding each other, and can become clinically significant if associated with more prolonged periods of apnoea, episodes of bradycardia, desaturations or hypoxaemia (abnormally low partial pressure of oxygen in the blood) [11, 18].

In addition, premature babies have much smaller alveoli than adults (many of them have alveoli with radii less then one quarter of that of an adult) [11]. Since surface tension is inversely dependent to the radius of the alveolus, the alveolar pressure due to the surface tension is much greater in pre-terms than in adults. Moreover, adult lungs secrete surfactant, a substance that greatly reduces alveolar pressure, preventing them from collapsing. Unfortunately, this substance does not start to be secreted before six or seven month of gestation, sometimes even later. Thus, most premature babies produce little or no surfactant in their alveoli. As a consequence, their lungs have a very strong tendency to collapse [15], tendency enhanced by the small FRC. To compensate for this, newborns keep the volume of air in their lung higher than the FRC, by actively

maintaining their respiratory rate very high (see table 2.1) [19]. The collapse of the lungs (happening e.g. when the respiration cease) is called respiratory distress syndrome of the new born, and is fatal if not treated, notably by the application of positive air pressure in the lungs [11].

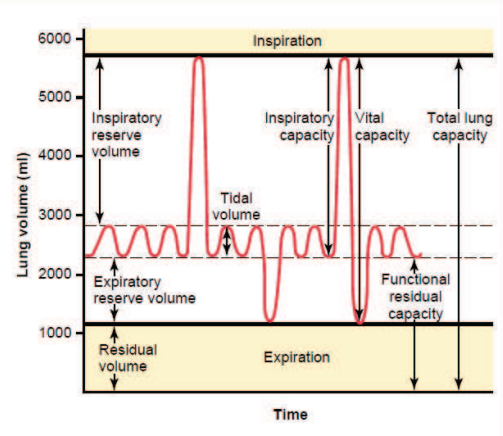


Figure 2.1: Illustration of respiratory volume changes during normal breathing and during maximal inspiration and maximal expiration. Tidal volume: volume of air inspired or expired during normal breathing. Inspiratory reserve volume: extra volume of air that can be inspired above the tidal volume during maximal inspiration.. Expiratory reserve volume: extra volume of air that can be expired after the end of a normal tidal expiration during maximal expiration. Residual volume: Volume of air remaining in the lungs after maximal expiration. Inspiratory capacity: tidal volume + inspiratory reserve volume. Functional residual capacity: expiratory reserve volume + residual volume. Vital capacity: inspiratory reserve volume + tidal volume + expiratory reserve volume. Total lung capacity: vital capacity + residual volume. Source: [15].

Other considerations The first role of the **heart** is to pump blood through the cardiovascular system. This organ can be considered as a combination of coupled muscles which the simultaneous contraction result in the mechanical action of pumping [20].

Cardiac muscles in newborns are much smaller in relation to their body weight than adults, which results in a much weaker pumping power. The baby compensates for this lack of pump power by a faster cardiac pulse [17]. Thus, like respiration, heart rate in newborn is twice as fast in relation to their body weight as in the adult (see table 2.1) [11]. The blood pressure is also much smaller in newborn than in adults (see table 2.1).

The underdevelopment of organ systems in pre-terms leads to great instability in the **homoeostatic mechanisms** of their body. In particular, premature

babies have difficulties to maintain a normal body temperature, which has a tendency to move towards ambient temperature. This is an important issue, because body temperature below 35.5°C can be fatal. To overcome this, premature babies have to be placed in incubators, where the external temperature is heavily regulated [11].

Parameter	Neonate	Adult
Respiratory parameters:		
Tidal volume (ml/kg)	7	7–10
Dead space (ml/kg)	2.2	2.2
VD:VT ratio	0.3	0.3
Respiratory rate	30–40	15
Compliance (ml/cmH2O)	5	100
Resistance (cmH2O/l/s)	25	5
Oxygen consumption (ml/kg/min)	7	3
Cardiovascular parameters:		
Heart rate (bpm)	80-200	60-110
Mean systolic blood pressure (mmHg)	50-90	110-130
Mean diastolic blood pressure (mmHg)	25-60	65-80

Table 2.1: Comparison between newborn and adult vital parameters. The parameters that show large contrasts between neonates and adults are highlighted in red. VD:VT ratio corresponds to the ratio between 'dead space' and 'tidal volume. The dead space is the air a person breathes that is not used for gas exchanges, as it never reaches the gas exchange areas, but only fills respiratory passages where gas exchange does not occur, such as nose, the pharynx, and trachea. The lung compliance corresponds to the extent to which the lungs will expand for each unit increase in transpulmonary pressure. [15]Adapted from [17].

2.2 Clinical issues

Apnoea, bradycardia and oxygen desaturations are the clinical events most commonly encountered in the neonatal unit [21]. The following section is dedicated to presenting the main characteristics of these conditions.

Apnoea Apnoea corresponds to periods with no breathing. Apnoea is defined to be clinically significant when it is longer then 20 seconds, or otherwise, when it is accompanied by bradycardia, cyanosis (bluish discolouration of the skin resulting from poor circulation or inadequate oxygenation of the blood), or pallor [22]. In most neonatology units, apnoea alarms are set to be triggered by periods of apnoea longer than 15-20 seconds.

There are three categories of apnoea. Central, obstructive or mixed. A central apnoea occurs when no respiratory effort is present. In contrast, when

respiratory effort is present, it is called an obstructive apnoea. Finally, a mixed apnoea occurs when elements of both central and obstructive apnoea are present [21].

Apnoea of prematurity is the most common type of apnoea in pre-terms. It consists of the cessation of breathing due to the immaturity of the brainstem centres that regulate respiration. This cessation of breathing leads particularly quickly to severe hypoxaemia (abnormally low partial pressure of oxygen in the blood), which in turn can cause oxygen desaturations and bradycardia episodes [23].

These early and recurrent occurrences of hypoxaemia during apnoea episodes are related to the low lung volume of pre-terms, in particular to their low functional residual capacity (FRC) - that is, low oxygen storage capacity (see section 2.1). It has been observed that the lower the FRC after the apnoea, the quicker the desaturation occurs [24]. This makes sense, since, as seen in section 2.1, the role of this volume is to stabilise oxygenation during brief periods of apnoea.

Clinical monitoring of apnoea is not reliable. Therefore, the presence of bradycardia and oxygen desaturation episodes indicate clinically significant apnoea episodes [21].

Bradycardia A bradycardia is defined as a decrease in heart rate below a predetermined threshold, which is usually fixed at 100 beats per minute, or at 25-30% of decline from baseline. To say that a bradycardia is clinically significant, it has to last more than 5 seconds [21]. Since heart rate monitoring is highly reliable and artefacts on ECG easy to pick out, bradycardia is used by clinicians as a marker for clinically significant episodes of apnoea [21].

Oxygen desaturation Oxygen desaturations are seen on the monitors during hypoxaemia episodes - decreased partial pressure of oxygen in the blood. One possible cause of a hypoxaemia episode is the reduction in alveolar ventilation during periods of apnoea. Therefore decrease of oxygen saturation observed on the monitor is often used as a tool to diagnose apnoea. In most neonatology units, low saturation alarms are set at 85 - 90 % for 3-5 seconds [21].

Temporal relationship between apnoea, bradycardia and oxygen desaturation The time of occurrence of these physiological events - that is, apnoea, bradycardia and oxygen desaturation - constitute important clues to help diagnose the condition of newborns in neonatology units. For example, apnoea occurring during sleep often indicates that it was caused by apnoea of prematurity. Some hypoxaemia episodes induced by apnoea are responsible for bradycardia and desaturations events called "feed-related", because they are triggered by oral feeding. They occur very commonly and almost exclusively in pre-terms during oral feeding, and are due to the brainstem immaturity in coordinating the acts of sucking, swallowing, and breathing [21]. Some hypoxia episodes provoke bradycardia and desaturation events that can be seen recurrently after feeding. These episodes are usually caused by diaphragmatic fatigue

[25]. When bradycardia and oxygen desaturations occur in mechanically ventilated infants, accompanied with a decrease in lung volume and ventilation, these episodes are often triggered by body movements [26]. It seems that hypoxaemia episodes can be triggered by excessive handling [27].

The order in which these clinical events happen is important as well, as it allows understanding of their cause and their physiology. Figure 2.2 gives the most common sequences of these events. In the first and most common case, an episode of apnoea leads to hypo-ventilation, which provokes an oxygen desaturation. This oxygen desaturation triggers a reflex bradycardia [28]. Examples of causes that generate this sequences are: apnoea of prematurity - cessation of breathing caused by the immaturity of the brainstem centres that regulate breathing, sepsis - presence in tissues of harmful bacteria and their toxins, typically through infection of a wound, and Central Nervous System depression [28]. Another possible sequence of events implies inhibitory reflexes, which trigger the development of an apnoea and a bradycardia almost simultaneously. In this case, oxygen desaturation occurs much later in the sequence of events. Consequently, if a sudden bradycardia occurs without any previous desaturation, we know that inhibitory reflexes are somehow involved [21].

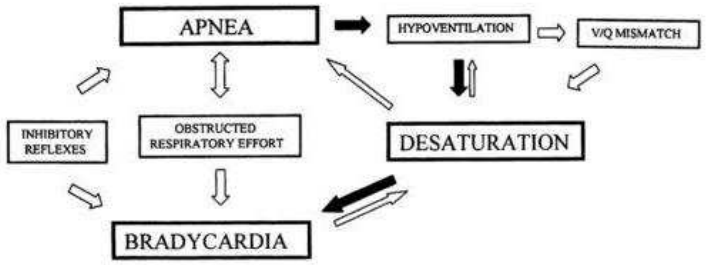


Figure 2.2: Illustration of the temporal relationship between apnoea, bradycardia and oxygen desaturations. The most common sequence of event, indicated by the black arrows, is triggered by an episode of apnoea that leads to hypo-ventilation, which in turn provokes an oxygen desaturation. A less common, sequence of event is indicated by the white arrows. This sequence is triggered by inhibitory reflexes, which provoke the development of an apnoea and a bradycardia almost simultaneously. The bradycardia leads in turn to an oxygen desaturation. Source: [28].

2.3 Condition monitoring

The state of health of the baby cannot be observed directly, but through physiological measurements collected by probes (see Figure 2.3). However, these physiological measurements do not reflect faithfully the true state of health of

the baby. Noise due to probe inaccuracy or artefacts (e.g. probe drop-out) can alter the measurements. Indeed, most of the alterations of the measurements that are unrelated to the babies' state of health are associated with the particular probes used to record the vital parameters from the baby. An understanding of the basic design assumptions of the probes used in neonatology units is therefore essential, in order to be able to distinguish between the physiological data that reflect the true state of health of the baby and erroneous data due to the particular probe used.

This section aims to describe the basic scientific principles of the functionalities of the most commonly used probes in neonatology units.

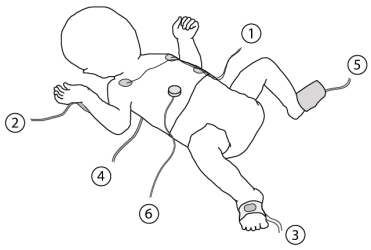


Figure 2.3: Probes used to collect vital parameters from an infant in intensive care. 1) ECG, 2) arterial line (connected to blood pressure transducer), 3) pulse oximeter, 4) core temperature probe (underneath shoulder blades), 5) peripheral temperature probe, 6) transcutaneous probe. Source: [10].

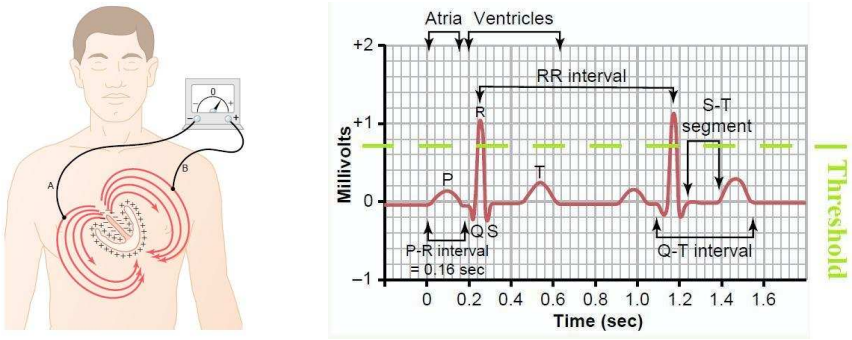
Electrocardiography (ECG) and Heart Rate Monitoring The heart rate is obtained from the electro-cardiogram (ECG) unit. An electro-cardiogram (ECG) consists of two electrodes which are placed on the baby's chest (see Figures 2.3 and 2.4b) and that records potential variations induced by the electrical depolarisation of the cardiac muscle. The resulting electrocardiogram is displayed in figure 2.4a.

The normal electrocardiogram is composed of a P wave, a QRS complex and a T wave. Each different wave is caused by electrical potentials generated by different part of the heart muscles during during the contraction process (see Figure2.4a) [29].

In NICU, only the heart rate is monitored, because the complete ECG can not be determined with enough precision without extra costs. The heart rate is computed from the time difference between successive peaks in the measured electrical signal from the heart (heart rate $= 1/RR_{interval}$ in Figure 2.4a).

Different kinds of artefacts can occur in the heart rate signal. They can be due to poor sensor contact, motion, aberrant detection of the 'T' wave (that would shift the signal up) or missing of a 'R wave' (that would shift the signal

down)[1, 29]. However, these artefacts are usually easy to detect, which makes heart rate a reliable channel [21].



(a) ECG signal.

(b) ECG electrodes.

Figure 2.4: (a) Illustration of ECG electrodes. Two electrodes are placed on the patient's chest. These electrodes record the potential variations induced by the electrical depolarisation of the cardiac muscle. (b) Illustration of the normal electrocardiogram resulting from the ECG recording. The ECG is composed of a P wave, a QRS complex, and a T wave. caused respectively by potentials generated when the atria depolarize before atrial contraction begins (P wave), by potentials generated as the depolarization wave spreads through the ventricles (QRS complex) and potentials generated as the ventricles recover from the state of depolarization. Source: [29].

Pulse oximeter (SpO_2) A pulse oximeter, bound to the baby's foot (see figure 2.3), measures the proportion of haemoglobin molecules in the arterial blood which are loaded with oxygen. The oximeter is composed by an emitting light diode (LED) and a photo-detector positioned opposite to the LED. The oximeter relies on the fact that deoxygenated haemoglobin absorbs more light in the red band (at 600–750nm), whereas oxygenated haemoglobin absorbs more light in the infra-red band (850–1000nm; see Figure 2.5). The ratio of the absorbency of red and infra-red light sent through a tissue relates to the ratio of oxygenated relative to deoxygenated haemoglobin [1].

In principle, the delay between a fall in the oxygen level and its display on the monitor should be approximately 4 seconds - the time it takes for the deoxygenated blood to travel from the lung to the sensor site, here the toe. In practise however, pulse oximeters average their values over a short period of time, varying from 2 to 15 seconds, in order to smooth the output. A downside of this is that it limits the temporal resolution, so that fast changes in oxygen levels aren't observed. Moreover, it makes it difficult to distinguish real desaturations from artefactual drops in SPO_2 readings during periods of body movements (handling, feeding, etc.) [1].

A certain number of artefacts can occur with the pulse oximeter. If the photo-detector is not placed exactly opposite the LED, it will not be sufficiently protected from the ambient light, so that light circumventing the tissue will cause erroneous high or low values. Alternatively, if the oximeter is applied with too much pressure, it can reduce the signal-to-noise ratio and therefore weaken the precision of the SPO_2 measurements [1].

If we are faced with an episode of intermittent body movement (e.g. during feeding, handling, etc.), then this can result in artefactual measurements, as pulse oximeters are very sensitive to sudden changes in background signal. Because SPO_2 readings are easily subject to artefacts, it is very important to be able to distinguish between artefactual and true measurements. In NICU, this is often achieved by comparing the pulse rate from the oximeter with the heart rate from an ECG monitor, which should be identical if the oximeter is not subject to artefacts [30]. As artefacts due to movements or loose leads are easy to detect for the heart rate measurements, they are in general more reliable than oxygen saturation measurements, and if the two channels differ, the SPO_2 values will in general be the wrong ones [21].

The oximeter is highly used in NICU, because it allows efficient detection of hypoxaemia episodes (defined as $SPO_2 < 80\%$). However, oximeters are responsible for a very high proportion of false alarms in neonatology units. Reducing false alarms is a challenging problem, as such reduction should not be accompanied by reduction of true positives. Table 2.2 displays the most commonly used NICU alarm limits.

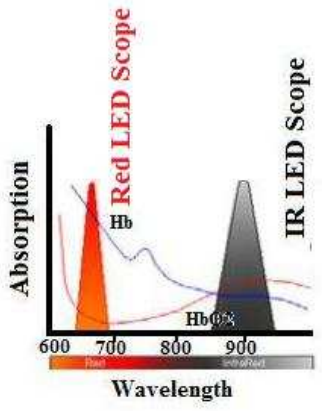


Figure 2.5: Absorbency of oxygenated and deoxygenated haemoglobin. Maximal light absorption of deoxygenated haemoglobin occurs between 600 and 750 nm (red band), whereas maximal absorption of oxygenated haemoglobin occurs between 850 an 1000 nm (infra-red band). The ratio of oxygenated relative to deoxygenated haemoglobin can be determined from the ratio of the absorbency of red and infra-red light sent through a tissue. Adapted from [31].

Other probes The **blood pressure sensor** consists of a catheter with a pressure sensor introduced in the artery to measure systolic and diastolic blood pressure (respectively maximal and minimal blood pressure) continuously [32]. The blood pressure probe can also be used to measure the heart rate (instead of using the ECG) [10]. When a blood sample is taken from the baby, a saline pump acts against the sensor all along the operation, leading to an artefactual ramp in the blood pressure readings. Moreover, during this event, the blood from the arterial line containing the pressure sensor will be deviated. Therefore, if the heart rate is measured using the blood pressure probe, no measurements can be observed on the heart rate channel during a blood sampling (see section 3.2.2 below) [10].

A **transcutaneous probe**, attached to the baby's skin on the chest, monitors the partial pressures of oxygen ('$PTcO_2$') and carbon dioxide ('$PTcCO_2$') in the bloodstream.

The $PTcO_2$ sensor consist of a fuel cell. Normally, a fuel cell used as a source of electricity provides as much electrical current as possible as long as the amount of fuel and oxygen available is sufficient. However, a limitation in the amount of available oxygen reduces the generated current. In the fuel cell used here, oxygen is the limiting factor, so that the electrical current produced by this fuel cell is proportional to the oxygen available. Electrodes in the fuel cell must be heated to improve oxygen diffusion, which takes 10-15 minutes, and have to be recalibrated every 4-8 hours [1].

The $PTcCO_2$ sensor consists of a pH-sensing electrode and a reference electrode, covered by a hydrophobic CO2-permeable membrane from which they are separated by a electrolyte solution. The CO_2 diffusion across the membrane causes a change in the pH of the electrolyte solution, which is sensed by the pH-sensing electrode [1]. As for $PTcO_2$probes, they have to be recalibrated regularly. [1]

The core temperature and peripheral temperature are measured by two **temperature probes**, one placed under the baby's back (or under the chest if the baby is prone) and the other attached to a foot [10].

Environmental measurements (ambient temperature and humidity) are collected directly from the incubator [10].

	Lower limit	**Upper limit**
Heart rate (1/min)	80	220
Respiratory rate - apnoea duration	disabled	disabled
SPO_2 **in pre-term neonates receiving oxygen (%)**	85	95
SPO_2 **in pre-term not receiving** O_2**, or in term infants (%)**	85	disabled
PtO_2 **in pre-term neonates receiving oxygen (kPa)**	6.0	10.7[a]
$PtCO_2$ **(kPa)**	6.0	7.3

Table 2.2: Alarm Limit settings for most NICU Monitors. This table shows that the detection systems used in NICU are very basic, with the alarm being triggered when a particular signal goes above or below preset thresholds. The respiratory rate alarm is often disabled, as an apnoea is not considered as dangerous for the baby as long as the blood gases nor the heart rate are affected. Adapted from [1].

Chapter 3

Probabilistic model: background

The goal of the system built by my host laboratory in the context of the 'Condition Monitoring in Premature Babies' project is to infer, from the vital parameters recorded from the babies (i.e. heart rate), the underlying cause of the observations - that is, the combination of the current condition of the baby (e.g. healthy, bradycardia) and the artefacts occurring to the probes (e.g. probe drop-outs) [8]. To do this, each condition and artefact can be thought of as a discrete cause - or 'factor' - that gives rise to a particular dynamics in the recorded channels such as heart rate, blood pressure and oxygen saturation. In the first step, a generative model is built, that describes how each potential combination of factors generates the data (see Figure 3.1a for an illustration). A probabilistic model is chosen, as it has the advantages of allowing us to deal with noise and imperfect knowledge of the world [9]. In the context of condition monitoring, a probabilistic model is particularly useful, as noise due to probe inaccuracy often alters the measurements, and the readings of the vital parameters recorded from the probes reflect only imperfectly the true physiological state of the pre-term babies. Specifically, a time-series model is used here, as the data - i.e. the vital parameters of the babies - consists of variables that evolve in time. After having learnt the model parameters from the data, the model is used to infer the combination of physiological and artefactual factors responsible for the observations at each timestep (see Figure 3.1b for an illustration). The final goal is to be able to fit this system with an alarm which can be triggered whenever the system detects a clinically significant physiological problem occurring to the baby.

This chapter aims at describing the system built by my host laboratory in the context of the 'Condition Monitoring in Premature Babies' project. Section 3.1 gives an overview of probabilistic time-series modelling. Section 3.2 is dedicated to describing the system itself.

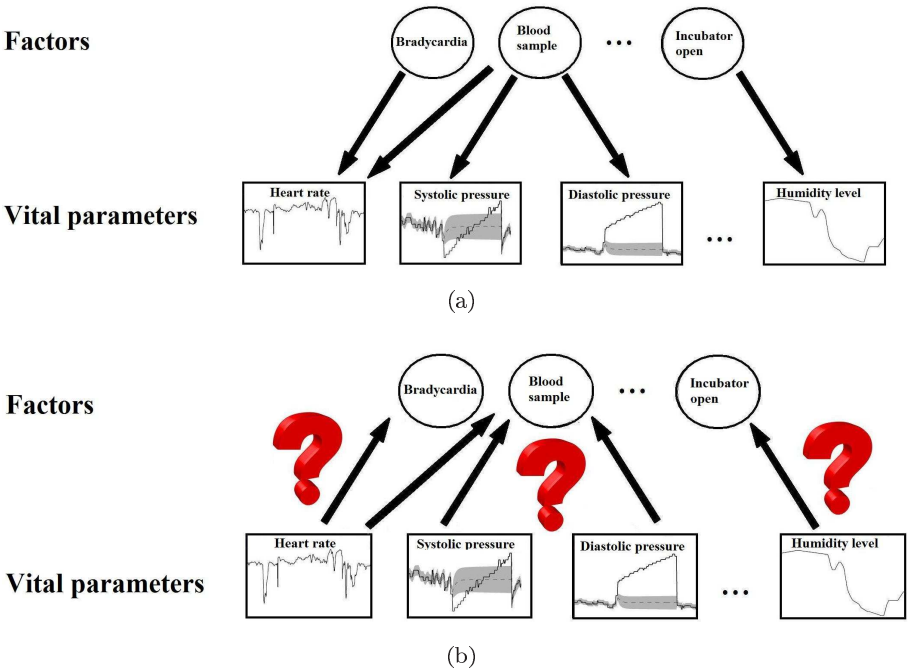


Figure 3.1: Simplified illustration of the FSLDS model. (a) Diagram of the construction of the model. A certain number of discrete factors - corresponding to physiological (e.g. bradycardia) or artefactual (e.g blood sample) events - give rise to a specific kind of dynamic in the measurements recorded from the baby and seen on the monitor. (b) Illustration of the process of inferring the underlying factors - artefactual or physiological - that are responsible for the sequence of measurements observed on the monitor.

3.1 Overview on Time Series Modelling

The data used in this project describes the state of the baby as it evolves in time. In the language of machine learning, it is 'sequential'. Time series data - where we observe the evolution of a variable through time - represent a common form of real-world sequential data. Other examples of time-series data include rainfall or temperature measurements on successive days, or even recorded speech. Specific kinds of datasets need specific kinds of models . This section is dedicated to introducing probabilistic models for sequential data [33, 34].

3.1.1 Auto-regressive model

Auto-regressive (AR) process constitute a useful method to model time series data [35]. An $AR(p)$ process assumes that the current observation (x_{t+1}) corresponds to a linear function of the observations at p previous time steps with additive Gaussian noise. Such a process is described by the following equation

recurrence relation:

$$x_{t+1} = \sum_{i=1}^{p} a_i x_{t-1} + w_t \,, \tag{3.1}$$

where $a < 1$ and $w_t \sim N(0, \sigma^2)$ is a Gaussian random variable with 0 mean and variance σ^2. In general, we can find the steady state solutions to an $AR(p)$ process using the Yule-Walker equations (see section 5.1.2 below).

An $AR(p)$ process can be rewritten in a vector form, replacing the sum in the recurrence relation with a matrix multiplication. For example the AR(2) recurrence would be written as:

$$\begin{pmatrix} x_t \\ x_{t-1} \end{pmatrix} = \begin{pmatrix} a_1 & a_2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ x_{t-2} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} w_t \\ w_{t-1} \end{pmatrix}$$

In general, an $AR(p)$ process can be written in a vector $AR(1)$ process with a p-dimensional state vector.

Finally, the state is not limited to being a scalar value. It can be a vector. In this case, the original $AR(p)$ recurrence relation can be rewritten as:

$$\mathbf{x}_{t+1} = \sum_{i=1}^{p} \mathrm{A}_i(\mathbf{x}_{t-i}) + G\mathbf{w}_t \,, \tag{3.2}$$

where A and B are square matrices. For more details about AR process, see [34, 36].

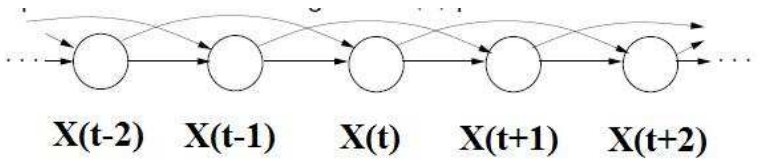


Figure 3.2: Graphical illustration of an $AR(2)$ process. The value of the variable at each timestep, x_t, is a linear combination of the values of the variable at the two previous timesteps, x_{t-1} and x_{t-2}. Source: [36].

3.1.2 Linear Dynamical System

In many datasets, the observed variables do not correspond directly to the variables we are interested in. For instance, if our dataset corresponds to a heart rate recorded using an electro-cardiogram (ECG), the values observed on the monitor can be different to the real heart rate, because they can have been altered by noise (for instance due to the probe inaccuracy), or by artefacts (for instance due to the patient movements). Therefore, many models are composed

by 'hidden' or 'latent' variables, which are essential for the model description but never observed (in the ECG example above, the hidden variable would be the true heart rate) and by 'observed' variables generated by the 'hidden' variables (in the ECG example above, the observed variable would be the heart rate read on the monitor) [9].

In a linear dynamical system (LDS), each latent variable $\mathbf{z}_n$ is a linear function of the latent variable in the previous state $\mathbf{z}_{n-1}$. Specifically, the $\mathbf{z}$ variables are governed by an $AR(1)$ process (equation 3.3), with each latent variable generating an observed variable (equation 3.4).

$$\mathbf{z}_{n+1} = A\mathbf{z}_n + \mathbf{w}_n \,, \tag{3.3}$$

$$\mathbf{x}_n = C\mathbf{z}_n + \mathbf{v}_n \,, \tag{3.4}$$

$$\mathbf{z}_1 = \boldsymbol{\mu}_0 + \mathbf{u} \,, \tag{3.5}$$

where the noise terms are described by the following Gaussian distributions:

$$\mathbf{w} \sim \mathcal{N}(\mathbf{w}; \mathbf{0}, \boldsymbol{\Gamma}) \,, \tag{3.6}$$

$$\mathbf{v} \sim \mathcal{N}(\mathbf{v}; \mathbf{0}, \boldsymbol{\Sigma}) \,, \tag{3.7}$$

$$\mathbf{u} \sim \mathcal{N}(\mathbf{u}; \mathbf{0}, \boldsymbol{P}_0) \,. \tag{3.8}$$

The graphical representation of this process is illustrated in Figure 3.3 [33].

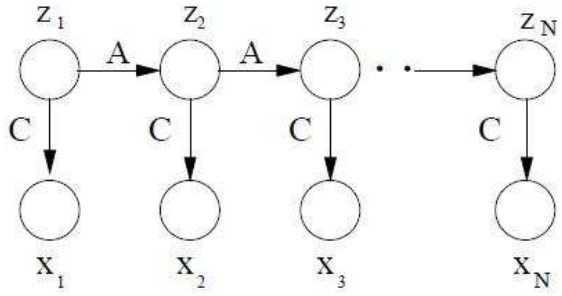


Figure 3.3: Graphical illustration of a state-space model for time series. Each latent variable $\mathbf{z}_n$ is conditioned on the value of the latent variable in the previous state $\mathbf{z}_{n-1}$, via the transition matrix A (equation 3.3). At each timestep, the observed variable $\mathbf{x}_n$ is dependent on the latent state at that time $\mathbf{z}_n$, via the emission matrix C (equation 3.4). Source: [36].

3.2 The Factorial Switching Linear Dynamical System

In some cases, time-series data do not follow a single type of regime, but jump abruptly from one regime to another. In this case, while a single Linear Dynamical System (LDS) may not provide a good description of the data, a combination of different LDS models, namely a 'switching linear dynamical system' (SLDS), can do better [9]. In an SLDS model, a discrete latent variable s_t indicates which of the LDS sub-models is most appropriate to describe the observations at any given time. Consequently, for a given switch setting (i.e. conditioned on s_t), the SLDS becomes equivalent to the corresponding LDS for that switch setting [10]. Switching linear dynamical systems, (SLDS) have been used previously in various applications such as modelling human motion [37] or creatinine levels in patients with kidney transplants [38].

The data used in this project are well suited to an FSLDS model. Different physiological conditions of the baby (e.g. bradycardia), and artefacts that can alter the recordings (e.g. when the incubator door is opened) are represented by discrete factors. Collectively, these factors determine which switch setting is active at any given time. In turn, the switch setting determines the time-series dynamics of the observed data. This model, with multiple discrete factors determining the switch setting, represents a particular case of the SLDS model described above: a factorial SLDS (FSLDS). Factorial SLDS have been used previously in applications such as speech recognition [39] and musical transcription [40].

3.2.1 Model description

In an FSLDS model, different discrete factors determine together the dynamic by selecting between different LDS (see Figure 3.4 for schematic). In this type of model :

- $f_t^{(m)}$ is a discrete 'factor' variable corresponding to a physiological or artefactual event happening to the baby at time t. For example: 'bradycardia', 'blood sample' or 'probe drop-out'.
- s_t is the switch variable, which is determined by the combination of factors that are active at time t . For example: 'Blood sample AND Transcutaneous probe recalibration' or 'Temperature probe disconnection AND Transcutaneous probe recalibration'. Each switch settings corresponds to a particular LDS.
- x_t is the hidden continuous state at time t. This contains estimates of the true state of health of the baby and on the levels of artefactual processes.
- $y_{1:t}$ correspond to the observations, that is, the monitored data.

In the non factorial SLDS, the switch variable (s_t) selects the dynamics for a particular combination of factor settings (equations 3.9 & 3.10), while the hidden

state (x_t) and switch settings (s_t) determine the dynamics of the observed data (y_t), as follows:

$$\mathbf{x}_t \sim \mathcal{N}(\mathbf{A}^{(s_t)}\mathbf{x}_{t-1} + \mathbf{d}^{(s_t)}, \mathbf{Q}^{(s_t)}), \tag{3.9}$$

$$\mathbf{y}_t \sim \mathcal{N}(\mathbf{C}^{(s_t)}\mathbf{x}_t, \mathbf{R}^{(s_t)}), \tag{3.10}$$

In these equations, x_t and y_t are sampled from normal distributions with mean and covariance dependent on the switch setting s_t. For each setting of s_t, the predictive distribution of x_t can be obtained using standard Kalman filter equations [10].

The factorial case (FSLDS model) is an extension of the SLDS, where a set of M discrete 'factor' variables, $f_t^{(1)}...f_t^{(M)}$, determine the state of the switch variable at a given time (s_t). The m^{th} factor, $f^{(m)}$ can take on $L^{(m)}$ different values. The state space for the switch variable is dependent on the combination of factor variables, according to:

$$s_t = f_t^{(1)} \otimes .. \otimes f_t^{(M)}$$

As such, the switch variable s_t can take $K = \prod_{m=1}^{M} L^{(m)}$ different values. The factors are independent of each other, and depend only on their value at the previous time step, so that:

$$p(s_t|s_{t-1}) = \prod_{m=1}^{M} p\left(f_t^{(m)}|f_{t-1}^{(m)}\right)$$

Factor transition probabilities can be estimated from labelled training data according to the relation:

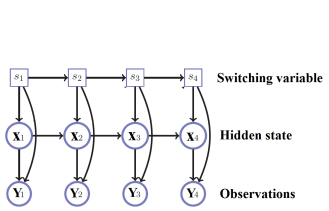
$$P\left(f_t^{(m)} = j|f_{t-1}^{(m)} = i\right) = \frac{n_{ij} + \zeta}{\sum_{k=1}^{M} n_{ik} + \zeta}$$

where:

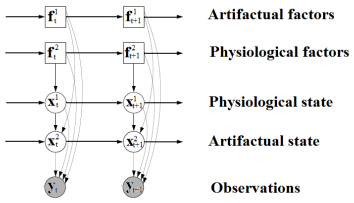
- n_{ij} represents the number of transitions from factor setting i to setting j observed in the training data.
- The constant term ζ is added to prevent the transition probabilities from being too small. (Source: [10]).

3.2.2 Factors

Specific patterns of physiological measurements are associated with different conditions, so that the recognition of such patterns can be used to infer the baby's state of health. However, how described above, the observations rely not only on physiological factors, but also, on artefactual factors. When an



(a) Switching Linear dynamical System



(b) Factorial Switching Linear Dynamical System for Condition Monitoring

Figure 3.4: From non factorial SLDS to factorial SLDS. (a) Graphical illustration of a switching linear dynamical system. The hidden switching variable s_t selects the dynamics at time t. Given a particular switch setting, the model is equivalent to a linear dynamical system (LDS) (See section 3.1.2). The observation variable y_t depends on the value of the hidden state variable x_t and the switching variable s_t at the corresponding time step. The hidden variable x_t is dependent on the value of the hidden variable in the previous timestep x_{t-1} and the value of the corresponding switch setting s_t, The switch variable s_t depends on the value of the switch variable at the previous timestep s_{t-1}. (b) Factorial switching linear dynamical system, illustrating the clinical significance of each of the different model variables. The switching variable s_t is now factorised into two factors $f^1$and f^2, representing respectively an artefactual and a physiological factor. Adapted from: [10].

artefact occurs, this leads to changes in the observed data which are not caused by changes in the baby's physiology. The main physiological and artefactual factors considered in this system are explained in the two following paragraphs and illustrated in Figure 3.5.

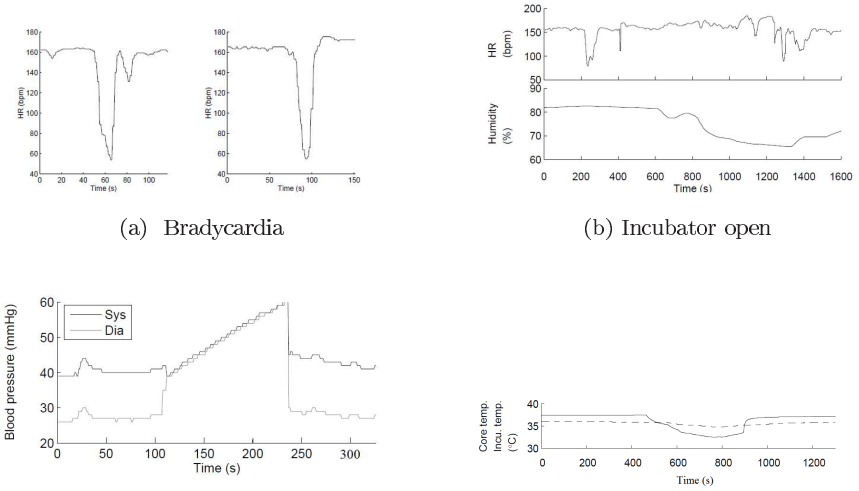
Physiological factors: *Bradycardia:* Bradycardia is a 'slower than normal' heart rate. The corresponding pattern can be seen in figure 3.5a. Brief episodes of bradycardia happen often in premature babies, and can have many causes (see section 2.2).

Artefactual factors:

- *Incubator open/Handling (IO):* The opening of the incubator leads to a drop in incubator humidity. The handling that often accompanies the opening of the incubator causes increased physiological variations. An example of such a pattern can be seen in Figure 3.5b. The drop in humidity at 800 seconds corresponds to the time when the incubator is opened, and the following episode of heart rate disturbance is related to the baby's handling.
- *Blood sample (BS):* Every few hours, the hospital staff take a blood sample (BS) from the baby. This implies that the blood from the arterial line containing the pressure sensor will be deviated. Therefore, no measurements can be observed on the heart rate channel during this event (if the

heart rate was measured using the blood pressure probe, rather than the ECG; see section 2.3). Moreover, a saline pump acts against the sensor all along the operation, leading to an artefactual ramp in the blood pressure measurements (see Figure 3.5c).

- *Temperature probe disconnection (TPD):* This artefact causes the core temperature measurement to decrease to ambient temperature, which usually corresponds to the incubator temperature, but can be lower if the probe is near to the portals, as seen in Figure 3.5d, where the solid line corresponds to the core temperature, and the dashed line corresponds to the incubator temperature.



(a) Bradycardia

(b) Incubator open

(c) Blood sample

(d) Temperature probe disconnection

Figure 3.5: Some physiological and artefactual factors that can affect the measurements. (a) Example heart rate readings during a bradycardia event, labelled by a sharp drop in the heart rate (to <100 bpm). (b) Example humidity level readings during incubator opening. Incubator opening leads to a drop in the humidity level (at $\sim$ 800s). (c) Example blood pressure readings during a blood sampling event. The blood sampling leads to a ramp in the blood pressure measurements (from $\sim$100 - 230s). (d) Example temperature readings during a temperature probe detachment event, leading to a decay of the temperature measurements (from $\sim$500-900s). Source: [10].

3.2.3 Novel condition

The list of factors given above is not exhaustive: a number of other factors, including sepsis, drugs and neurological conditions, likely influence the observed

data. However, modelling all potential dynamical regimes is infeasible. Therefore, an extra factor was added (the 'X-factor'), to indicate when neither the 'normal' regime or any of the known factors could provide a good description of the data (effectively this corresponded to a setting of "none of the above"). That is, the X-factor describes abnormal dynamics that cannot be explained by the model [10].

The benefits of adding the X-factor are twofold. Firstly, it helps us to identify when we have to deal with a new, unknown regime, which could be potentially dangerous for the premature baby. Secondly, by indicating which sections of data are not well described by the model, it can help to inform us how the model might need to be changed, in order to better fit these sections of data. For example, when a regime is classified as "none of the above", it could mean that extra factors should be added.

We now explain, using a static 'toy' model as an example, how the X-factor is added. In this model one setting of the switch variable s corresponds to the "normal" mode where the baby's physiology is stable without artefactual factors being active. Other settings of the switch variable correspond to abnormal dynamics, that may result from specific pathologies or artefactual factors being active. Assuming that $s = 1$ corresponds to the "normal" mode, and $s = 2, 3, ..., K$ to the other known modes, a new mode can be added, indexed by $s = *$, to account for unexpected data points. To do this, we can choose a Gaussian distribution with mean equal to the "normal" mode, but with larger variance:,

$$\Sigma^{(*)} = \xi\Sigma^{(1)}, \qquad \mu^{(*)} = \mu^{(1)} \tag{3.11}$$

Here $\xi > 1$ represents how far outside the normal range new data points have to fall before they are considered as "not normal". The Figure 3.6 (a) shows the likelihood functions for a normal class (solid line) and for the corresponding X-factor (dashed line). We clearly see that the X-factor has the same mean, but a higher variance. This implies that data points which lie far away from the normal regime are more likely to be considered as belonging to the X-factor regime. In other words, the high variance regime wins when the normal model can not well explain the observations.

Figure 3.6 (b) illustrates how X-factor could be used in conjunction with known factors (dashed line). In this example, the high variance model, or X-factor wins when neither the normal mode nor the known modes are able to well explain the observations.

Now we would like to generalize this 'static' example, to the case where the dynamics of the system evolve in time. The conditional distributions for the hidden and observed continuous states were described previously in equations 3.9 and3.10. For the 'normal' regime, with switch setting $s_t = 1$, these distributions are written as,

$$\mathbf{x}_t \sim \mathcal{N}(\mathbf{A}^{(\mathbf{s_t=1})}\mathbf{x_{t-1}} + \mathbf{d}^{(\mathbf{s_t=1})}, \mathbf{Q}^{(\mathbf{s_t=1})}), \tag{3.12}$$

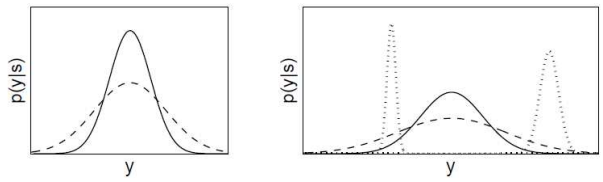


Figure 3.6: (a) Class conditional likelihoods in a static 1D 'toy' model, for the normal mode (solid) and the X-factor (dashed). Data points that are far away from the normal range are more likely under the X-factor regime. (b) Likelihoods of the normal class and X-factor alongside other known, abnormal regimes (dotted). The X-factor has the highest likelihood for regions that are far away from any known modes, as well as far away from the normal range. Source: [10].

$$\mathbf{y}_t \sim \mathcal{N}(\mathbf{C}^{(\mathbf{s_t}=\mathbf{1})}\mathbf{x_t}, \mathbf{R}^{(s_t=1)}) \,. \tag{3.13}$$

Similar to the static case, the new X-factor mode is obtained by inflating the noise covariance of the normal mode dynamics by a factor of $\xi > 1$, as follows:

$$Q^{(*)} = \xi Q^{(1)}, \tag{3.14}$$

Again, as with our static example, all of the other parameters for the X-factor are identical to the normal mode:

$$\left\{A^{(*)}, C^{(*)}, R^{(*)}, d^{(*)}\right\} = \left\{A^{(1)}, C^{(1)}, R^{(1)}, d^{(1)}\right\}. \tag{3.15}$$

3.2.4 LDS model selection

To use the proposed FSLDS model to analyse clinical data, an LDS system must be specified for each setting of the factors (i.e. for each switch setting). First the 'type' of LDS model must be chosen: for example, for the $AR(p)$ process described previously, we must select the 'order' of the regression (the value of 'p'). Secondly the model parameters must be learned from the data.

In general, these tasks are made easier if we have access to labelled data. In this project, obtaining labelled data is facilitated by the fact that the discrete factors have a 'real-world' interpretation: they correspond to known physiological and artefactual states. Therefore, domain knowledge (i.e. medical knowledge), can be used to label which factors are likely associated with different sections of data. This, in turn, allows us to know *which* LDS model to try and fit to each section of physiological data.

For example, in the 'normal' case (i.e. when the baby is stable and no artefactual factors are active) observation channels (heart rate/blood pressure etc.) generally follow a specific set of dynamics. Therefore, after identifying (using

annotations obtained from a clinical expert) which sections of data correspond to the 'normal case' we can then attempt to find an LDS model that best fits these sections of data.

Normal dynamics

We describe how the LDS model was chosen for the 'normal' condition. For concreteness we focus here on the heart-rate measurements, although a similar approach was used to construct the LDS models for all of the observation channels. For more details about the construction of the other models, see [10] and [41].

Initially, the training data for normal dynamics was labelled manually, using the responses of clinical experts, which was very time-consuming and therefore, not efficient. Subsequent work performed by a Masters student in my host laboratory lab permitted automatic labelling of 'normal' data segments [42].

Figure 3.7 shows examples of heart rate measurements in the 'normal' condition. We see that such measurements follow a recognizable pattern, characterized by a slowly drifting baseline around which the measurements fluctuate. Therefore, a model with two hidden components can be used consisting of a signal variable (x_t), which fluctuates around a slowly drifting baseline variable (b_t). The dynamics governing the data are fitted using an $AR(p = 1)$ process for the signal variable, and an $AR(p = 2)$ process for the baseline [10]. This is described mathematically as follows,

$$x_t - b_t \sim \mathcal{N}(\sum_{\mathrm{k=1}}^{\mathrm{p_1}} \mathrm{a_k}(\mathrm{x_{t-k}} - \mathrm{b_{t-k}}), \eta_1), \qquad \mathrm{b_t} \sim \mathcal{N}(\sum_{\mathrm{k=1}}^{\mathrm{p_2}} \beta_\mathrm{k} \mathrm{b_{t-k}}, \eta_2) \tag{3.16}$$

where η_1 and η_2 are noise variances. These dynamics can be represented in state-space form as:

$$x_t = \begin{bmatrix} x_t \\ x_{t-1} \\ b_t \\ b_{t_{-1}} \end{bmatrix}, \quad A = \begin{bmatrix} \alpha_1 & \alpha_2 & \beta_1 - \alpha_1 & \beta_2 - \alpha_2 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & \beta_1 & \beta_2 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \tag{3.17}$$

$$Q = \begin{bmatrix} \eta_1 + \eta_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \eta_2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{3.18}$$

Abnormal dynamics

In the following, we describe LDS systems that were developed to describe 'abnormal' dynamics. We focus here on factors related to artefactual changes. For more details about the construction of the other models, see [10] and [41].

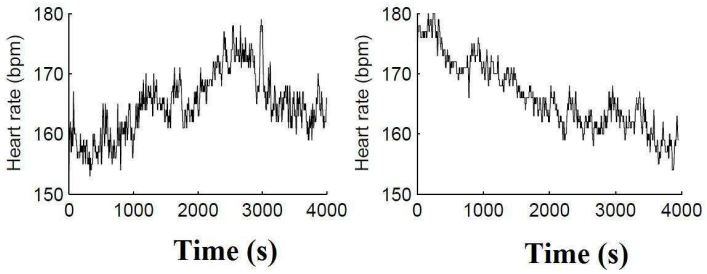


Figure 3.7: Example of heart rate measurements when the baby is in a healthy state. These measurements are characterized by a slowly drifting baseline around which the measurements fluctuate. Source: [10].

Blood sampling: When a blood sample is taken from the baby, a saline pump acting against the sensor produces an artefactual ramp in the blood pressure measurements (see Figure 3.8). The slope of the ramp depends on the rate at which saline is pumped, which is not the same every time.

To describe the pattern shown in figure 3.8, we construct an LDS model in which artefactual measurements a_t evolve according to a gradient which is subject to a random walk [10]:

$$a_t \sim \mathcal{N}(\mathrm{a_{t-1}} + \mathrm{d_a} + \mathrm{c_{t-1}}, \sigma_{\mathrm{a}}^2), \qquad \mathrm{c_t} \sim \mathcal{N}(\mathrm{c_{t-1}}, \sigma_{\mathrm{c}}^2) \tag{3.19}$$

In these expressions, every term is scalar and:

- d_a is a positive constant specifying the average drift.
- c_t corresponds to the gradient of the random walk, which modifies the average drift d_a.
- σ_{a}^2 is the variance of the Gaussian noise on a_t. This accounts for differences in slope of blood samples taken at different times.
- σ_{c}^2 is the variance of the Gaussian noise on c_t. This accounts for differences in slope within a single blood sample operations.

In a state-space representation, these dynamics can be written as:

$$x_t = \begin{bmatrix} a_t \\ c_t \end{bmatrix}, \quad d_{BS} = \begin{bmatrix} d_a \\ 0 \end{bmatrix}, \quad A_{BS} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$

$$Q_{BS} = \begin{bmatrix} \sigma_{\mathrm{a}}^2 & 0 \\ 0 & \sigma_{\mathrm{c}}^2 \end{bmatrix},$$

$$C_{BS} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \quad R_{BS} = \begin{bmatrix} r_{SysBP} & 0 \\ 0 & r_{DiaBP} \end{bmatrix}.$$

A simple way to check if the model is able to explain the data is to sample from it and check by eye whether its dynamics is similar to a sequence of data which is known to follow the same regime. This is done in Figure 3.8 (b), where a sequence of a blood sample measurements has been generated given the model, and can be visually compared to the real data sequence following the same regime shown in Figure 3.8 (a). We see that the real sequence and the sequence drawn from the models show similar behaviour.

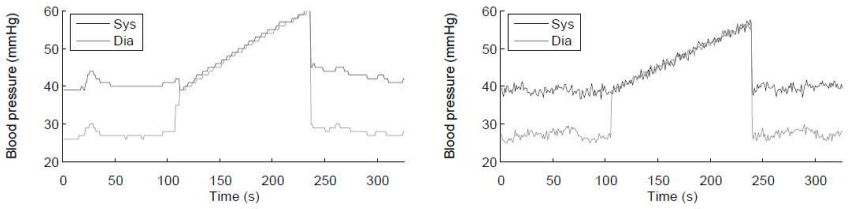


Figure 3.8: Comparison between (a) the blood pressure dynamic generated by a real blood sample episode and (b) the dynamic generated by a sample drawn from the model, with same initial value at $t = 1$ and the switch variable set at 'blood sample' being active. We see that the sampled sequence of blood sampling shows the same characteristics of dynamic than the real sequence of blood sampling. This confirms that the model has been adequately fitted. Source: [10].

Drop-outs: A probe drop-out causes the observations of the concerned channel(s) to go to zero, as shown in figure 3.9. The resulting model is the same as the "normal" model, except that the appropriate entry of the matrix C in equation 3.13 are set to zero.

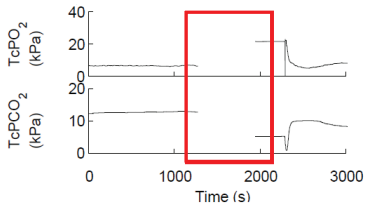


Figure 3.9: Example of probe drop-out. A probe drop-out occurs during the time sequence highlighted by the red square. As a result, no more measurements can be read in the channels $TcPCO_2$ and $TcPO_2$. Adapted from [10].

Temperature probe disconnection: When a temperature probe becomes disconnected, we observe artefactual measurements which indicate that the core

temperature measurement decreases to the ambient temperature. The slope of the decay should be the same for each disconnection. As the same type of probe is used for each baby, they all share the same thermal inertia (i.e. they all change temperature at the same rate). This constancy in the observed temperature changes can be used to differentiate between the dynamics caused by this artefact and the dynamics due to the baby getting colder. Looking at examples of such a disconnection (e.g. Figure 3.5d above), permits us to find a good model. In this case, an exponential decay, equivalent to an AR(1) process is used and learnt from the data., using the Yule-Walker equations.

Opening of the incubator: Temperature and humidity in incubators are closely regulated, which means that their variance is kept very low. The opening of the incubator leads to a significant drop in incubator humidity, as shown in figure 3.5b . An AR(1) decay constitutes a good model for these drops in humidity. The fact that the data cannot drop below the ambient humidity and temperature of the room must also be take into account in the model.

When a incubator is open, this usually means that the baby is being manipulated, which leads to an increase of the heart rate variance, (Figure 3.5b above), as well as a faint decrease in peripheral temperature, due to the influx of room air in the incubator. The resultant dynamic is the same as during normal dynamics, but with larger variance.

3.2.5 Factor interaction

We need to consider how different factors combine to condition the switch settings. If there are a large number of factors, learning the model parameters associated with every single combination of factors could require a large quantity of data. However, with multiple measurements channels, it is possible for some factors to 'overwrite' others. In other words, when two factors are active at the same time, the effects of both factors on a particular channel may often be the same as if only one of the factors were active. An example of such a case is shown in figure 3.10, where 'Bradycardia' and 'Blood sample' occur at the same time. In this case, the resulting observation on the monitor is the same as if only the blood sample were occurring, because no measurements of the 'heart rate' can be taken when a blood sample occurs (see sections 2.3 and 3.2.2).

Because of this 'overwriting' effect, examples of every combination of factors do not need to be found in order to train the full factorial model. The *Condition Monitoring in Premature Babies* project takes advantage of this fact, by training the model with individual factors separately, which are then combined together using simple reasoning about which channels should be overwritten for each combination of factors [10].

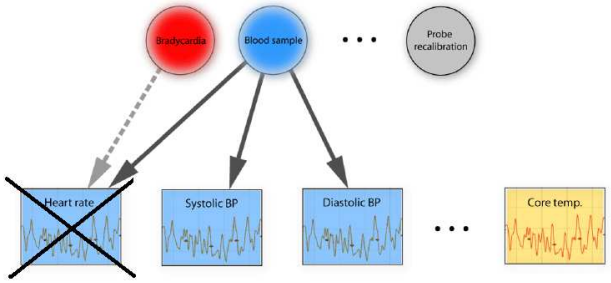


Figure 3.10: Illustration of factor interactions. In the case where both the bradycardia and the blood sample factors are being active at the same time, the resulting observations are similar to the case where only the blood sample factor is being active. This is because when the blood sample factor is being active, as a result, no measurements can be observed on the heart rate channel. Source: [43].

3.2.6 Learning

Given the factors associated with a particular data section, the parameter values for the corresponding LDS model had to be learned from the data. The learning algorithms used for the different LDS models differed, depending on their structure. For example, for heart rate dynamics in the normal condition, parameters were learned using EM updates, to maximize the log-likelihood. Because the EM algorithm is not guaranteed to find a global maximum of the log-likelihood, parameters were initialized using a heuristic procedure, so that they were initially close to their 'optimum' values that were required to fit the data. For more details of the learning algorithms used to train each of the LDS models see [10].

In common with the parameters of the LDS models for known factors, the novelty threshold for the X-factor (ξ) was directly learned from the data. However, unlike the other LDS models, the sections of the data that corresponded to the X-factor could not be labelled *a priori*. Therefore, this parameter had to be learned in a semi-supervised fashion, with only parts of the data labelled [10, 41].

3.2.7 Inference

The purpose of the system was to infer the baby's state of health at each time step. For our model this corresponds to inferring the hidden state of the baby at a time t $(s_t, \boldsymbol{x}_t)$, given the data that has been observed up to this point ($\mathrm{y}_{1:t}$): $p(s_t, \mathbf{x_t} \mid \mathbf{y_{1:t}})$. Unfortunately inference in an FSLDS is formally intractable, because it scales exponentially with time [9, 10]. Therefore, approximate methods had to be used to perform inference in this model.

Why inference is intractable At timestep 1, the posterior distribution, $p(s_1, \mathbf{x_1} \mid \mathbf{y_1}) = p(\mathbf{x_1} \mid \mathbf{y_1}, s_1)p(s_1|\mathbf{y_1})$ is given by an indexed set of Gaussians [9]. The posterior distribution at the next timestep, $p(s_t, \mathbf{x_t} \mid \mathbf{y_{1:t}})$ is obtained by taking a weighted sum over the switching states s_{t-1}, giving the recurrence relation:

$$p(s_{t+1}, \mathbf{x_{t+1}} \mid \mathbf{y_{1:t+1}}) = \sum_{s_t} \int_{\mathbf{x_t}} p\left(s_{t+1}, \mathbf{x_{t+1}} | s_t, \mathbf{x_t}, \mathbf{y_{1:t+1}}\right) p\left(s_t, \mathbf{x_t} \mid \mathbf{y_{1:t}}\right) \tag{3.20}$$

It follows that if S is the number of switch settings, then the $p(s_t, \mathbf{x_t} \mid \mathbf{y_{1:t}})$ is described by a mixture of S^{t-1} Gaussian at time t. Clearly, as the number of Gaussians required to parametrize the posterior distribution grows exponentially with time, exact inference is intractable [9].

Gaussian Sum approximation In the project *Condition Monitoring in Premature Babies*, in order to address this problem, a method called Gaussian Sum approximation is used. The main steps of this methods are presented as follows.

First, $p(s_t, \mathbf{x_t} \mid \mathbf{y_{1:t}})$ can be broken into a continuous and a discrete part, respectively (equation 3.21) [9]:

$$p(s_t, \mathbf{x_t} \mid \mathbf{y_{1:t}}) = p(\mathbf{x_t} \mid \mathbf{y_{1:t}}, s_t)p(s_t|\mathbf{y_{1:t}}) \tag{3.21}$$

At each time step, an approximation of $p(\mathbf{x_t} \mid \mathbf{y}_{1:t}, \mathrm{s_t})$ is maintained as a Gaussian mixture of I components, where $I < S^{t-1}$. The computation of the Kalman updates for the next time step will give $p(\mathbf{x_{t+1}} \mid \mathbf{y_{1:t+1}}, s_{t+1})$ as a Gaussian mixture of $S \times I$ components, due to the summation over all of the switch settings s_t. These $S \times I$ components can be collapsed back into I components, for example by matching means and variances of the distributions for each setting of s_t. An illustration of the Gaussian sum approximation is given in Figure 3.11.

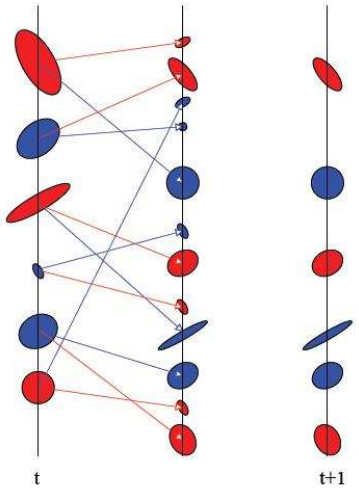


Figure 3.11: Gaussian Sum Approximation. The leftmost column represents the previous Gaussian mixture approximation of $p(\mathbf{x_t} \mid \mathbf{y_{1:t}}, s_t)$ for two states S = 2 (red and blue) and three mixture components I = 3. The mixture weight is depicted by the area of each oval. Each component of the mixture generates two new components through each of the S = 2 dynamic systems, with the colour of the arrow indicating which dynamic system is used. Therefore, at the next timestep, the joint approximation of $p(\mathbf{x_{t+1}} \mid \mathbf{y_{1:t+1}}, s_{t+1})$ contains $S \times I$ components (middle column). To keep the representation computationally tractable, the mixture of Gaussians for each state s_{t+1} is collapsed back to I components. This means that each coloured state needs to be approximated by a smaller I component mixture of Gaussians. One of the many ways to do it is to simply ignore the lowest weight components (see right column). Source: [9].

Chapter 4

System evaluation

When a segment of physiological data is classified as belonging to the X-factor, it means that this particular segment is neither classified by the normal regime, nor by any of the known abnormal regimes (see section 3.2.3). As it is only possible to model a limited number of factors, this has the advantage of ensuring that any novel regime is picked up by the system, which is particularly useful in a clinical context, for example in order to raise an alarm [10, 41].

Another benefit of the X-factor, is that it highlights segments of physiological data that cannot be explained by the factors explicitly included in the system, highlighting the existence of structure in the data that is lacking in the model. Further analysis of the segments classified by the X-factor would therefore provide additional knowledge about the different regimes that it claims. This would give clues of how to ameliorate the system, notably by flagging up existing factor models that could be improved or new factors that could be added to the system. My contribution to this project consists in conducting this investigation.

Experiments undertaken on the X-factor within the framework of the baby monitoring project have already shown that a significant number of non-normal regimes in the data have not yet been formally analysed. Notably, it has been observed that deep oxygen desaturations are an abnormal regime that is presently claimed by the X-factor, but constitute a clear and frequent pattern and could be usefully learnt as a new factor in the model. Oxygen desaturation seems to be often followed by a bradycardia, which is explained by the fact that the lack of oxygen causing the bradycardia slows the heart pulse. This suggest that the bradycardia factor and the new desaturation factor could be made mathematically dependent [10].

Finally, it has been seen that bradycardia are frequently linked with a compensatory rise in blood pressure, suggesting that an additional factor, modelling the blood pressure could be added and made dependent on the bradycardia factor. Other common patterns have been observed that could be added as new factors to the system, such as hypotension, hypertension, hypothermia and pyrexia, as well as more serious conditions, such as pneumothorax intraventric-

ular haemorrhage [10].

The investigation of physiological data segments highlighted by the X-factor could have the potential benefit to quantify the recurrence and verify the clinical significance of these different regimes, confirming or invalidating the relevance of the addition of these factors to the system. Thus, if a new pattern was highlighted by the system that appeared regularly, it would suggest that this pattern should be modelled and learnt by the system as a new factor. Alternatively, it could reveal that an existing factor was being regularly missed by the system and claimed instead by the X-factor, suggesting that the current model for this factor was not accurate enough.

Medical knowledge is essential to investigate the clinical significance of physiological data segments. Therefore, the study is based on a collaboration between medical experts and the machine learning lab, where the physiological data segments are selected by the X-factor to be submitted to clinical experts for medical interpretation.

A technical part of my contribution to the project consisted in the development of communication tools between the neonatology unit of the hospital and the machine learning lab, in order to gather clinical interpretations of segments highlighted by the X-factor. I describe this technical contribution in section 1. In section 2, I present the results of the analysis and interpretation of these feedbacks. This results are then discussed in section 3.

4.1 Methods

In this section, I describe the communication tools set up in order to gather clinical interpretations of the physiological data segments flagged up by the X-factor.

4.1.1 Automatic process

In order to obtain clinical feedback on physiological data segments highlighted by the X-factor, an evaluation system must be set up. This system should comprise the following stages. Firstly, physiological data from premature babies in the neonatology unit should be obtained. Secondly, the system should be used to classify the data segments which are due to the X-factor. These data segments should then be submitted to the clinical staff for medical interpretation. Finally, the interpretations of the data given by the clinical staff should be collected for analysis in a way that is as easy to interpret as possible.

The establishment of such an evaluation system poses several practical problems. First of all, the data manipulated concern real people protected by the medical privacy. This implies that the information going out of the hospital must be anonymized, so that it would be impossible for anyone working with the anonymized informations to know which data correspond to which person. Therefore, we work with identification numbers (IDs), and the web-based interface is password protected. Secondly, clinical data if the clinical experts look

at the output segments of physiological data one or more days after it has been recorded, it will likely be impossible for them to remember what happened, and thus, to give precise medical comments. In order to overcome this problem, an automatic procedure has been set up, consisting of the following steps:

1. Data feed specification

   (a) A simple web interface, allowing the medical staff to:

   - Create a list of babies to collect data on. This should permit for the addition and deletion of baby ID entries.
   - Specify when to collect this data.
   - Specify the number of hours of data to be collected.

   (b) Once the changes made by the user have been submitted, a 'settings.xml' file containing baby's ID, duration and time specifications is produced.

2. Data collection

   (a) Based on the updated 'settings.xml' file, physiological monitored data can be recorded for each of the babies, with the period and time specified in the 'settings.xml' file.

3. Data transfer

   (a) Once the recording is finished, the data collected is sent to the University of Edinburgh's School of Informatics.

   (b) The data arrives in the University of Edinburgh's School of Informatics server.

4. Data processing

   (a) The arrival of the data at the server is detected automatically. The system is then run on the data, to infer the probability that the different factors are activated at each time-step. For the X-factor, this provides us with a continuous signal, specifying the probability that the X-factor is activated at any given time.

   (b) This continuous signal goes through three steps of processing:

      i. **Step 1:** The continuous signal corresponding to the probability distribution of the X-factor is then thresholded (with a threshold of 0.15) to produce a binary output, classifying whether each section of the data is due to the X-factor (see Figure 4.1).

      ii. **Step 2:** The thresholded X-factor intervals are combined if they are too close to each other (see Figure 4.1).

iii. **Step 3:** If the same segment is flagged up simultaneously by the X-factor and one of the artefactual known factors (that is, 'incubator open', 'temperature probe detached', or 'blood sample'), the overlapping part is removed from the X-factor segment. Additionally, 'X-factor' intervals occurring less than three minutes following an 'incubator open' event are removed as well. This is because we do not want to consider the X-factor intervals that are due to the recovery of humidity to set level after the incubator have been opened and closed (see chapter 5 for more details). Figure 4.1 shows an example of how the X-factor classification is obtained from the raw output data, illustrating the influence of each processing step.

(c) Finally, the system outputs a list of the physiological data segments which have been classified as belonging to the X-factor. Only 10 of the X-factor intervals that last more than 3 minutes are selected in this list, to be submitted to the clinicians for physiological interpretations. Each these interval is allocated a score from one to a hundred quantifying how well it is explained by the X-factor. This score corresponds to the average posterior probability of the original X-factor interval.

5. Web mediated feedback collection

   (a) A file with the list of the time intervals for the selected data segments, is sent to the server and read by a web interface, which outputs feedback forms to be completed by the clinical experts. Each form displays the baby's ID with the date of the recording followed by the list of data segments selected by the system. Below each time interval there is a list of questions and a free text box so that the clinical experts can add their own additional comments.

   (b) The forms are completed by the transport fellows in charge of watching over the babies, and by the doctor responsible for the neonatology unit.

   (c) The completed forms are recorded in text files indexed by the baby's ID and the date.

My contribution to the setting up of this automatic system consisted in the development of the feed setting web interface (1st step) and the webform application (5th step).

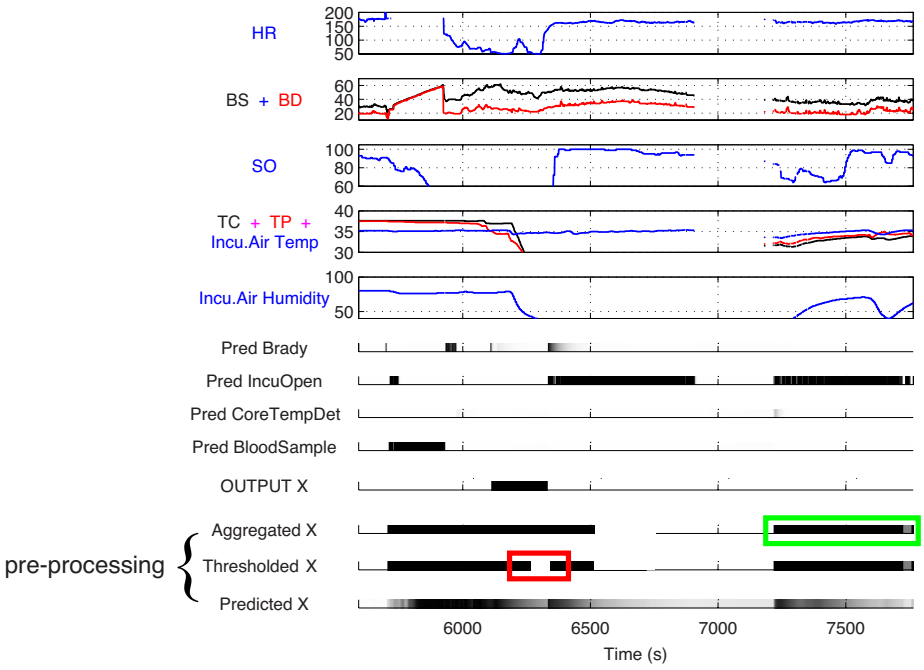


Figure 4.1: Illustration of how the X-factor classifications are obtained from the raw data. First, the continuous 'predicted X' signal obtained from the model is thresholded, to produce a binary output ('thresholded X'). Segments of the 'thresholded X' signal are then combined if they are too close to each other (less than 90 seconds; see red square for an example), producing the 'aggregated X signal'. If the same segment of data is flagged up simultaneously by the X-factor and one of the known artefactual factors (see green square for an example; the intervals for the posterior probability for the artefactual factor is 0.5), the overlapping part is removed from the X-factor segment, to produce the 'Output X'. Also, every 'X-factor' interval occurring during the three minutes following an 'incubator open' event is removed as well (see section 5 below for more details). 'HR' corresponds to heart rate, BS and BD corresponds to systolic and diastolic blood pressure respectively, SO corresponds to oxygen saturation, TC and TP correspond to central and peripheral temperature.

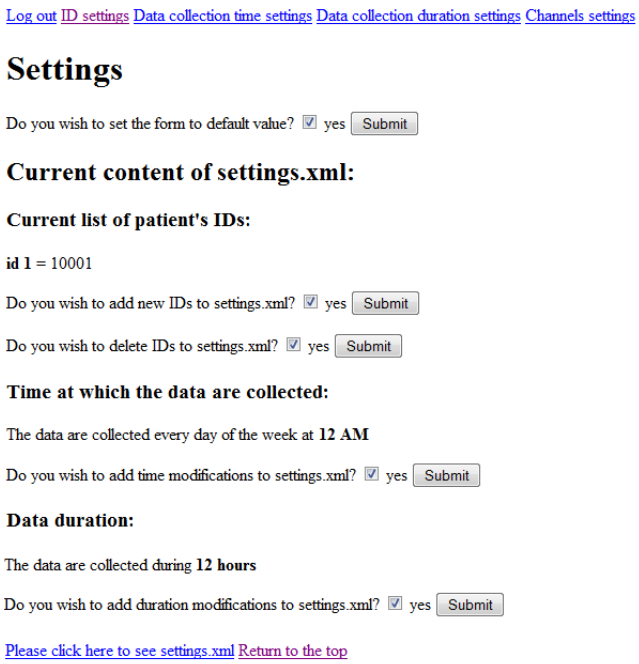
4.1.2 Feed settings

As mentioned above, it is necessary to have control over which baby to collect data from, as well as the duration and starting time of the recording. All these specifications are specified in a xml file called 'settings.xml'. As shown in Figure 4.2a, this kind of file is not user-friendly. Thus, a more friendly web interface was built to modify and update this xml file. A screen shot of this web interface

is displayed in figure 4.2b. The different functionalities are shown in more detail in Figure 4.3.

```
processingtime minute="0" hour="8" dayofmonth="*" dayofweek="1,2,3,4,5,6,7" month="*"/>
dataduration minutes="0" hours="8"/>
channellist>
channel name="Heart rate" code="tr_heartrate"/>
```

(a) Snapshot of an extract of the file 'settings.xml'



(b) Snapshot of the web interface permitting to the user to modify the file settings.xml

Figure 4.2: Setting specifications (a) directly via the original non user-friendly xml file and (b) using the user-friendly web interface I built. The xml file specifies the specifications to apply for the data recordings. More precisely, it specifies the list of patient to record the data on, when the recording should start (what time, which day of the week, etc..), how much time must the recording lasts, and finally, which particular channels should be recorded. The web-interface displayed in (b) allows the user to add and delete entries in the list of patient to record the data on, to specify and modify when the recording should start and how long it should last. The list of channels is by default exhaustive.

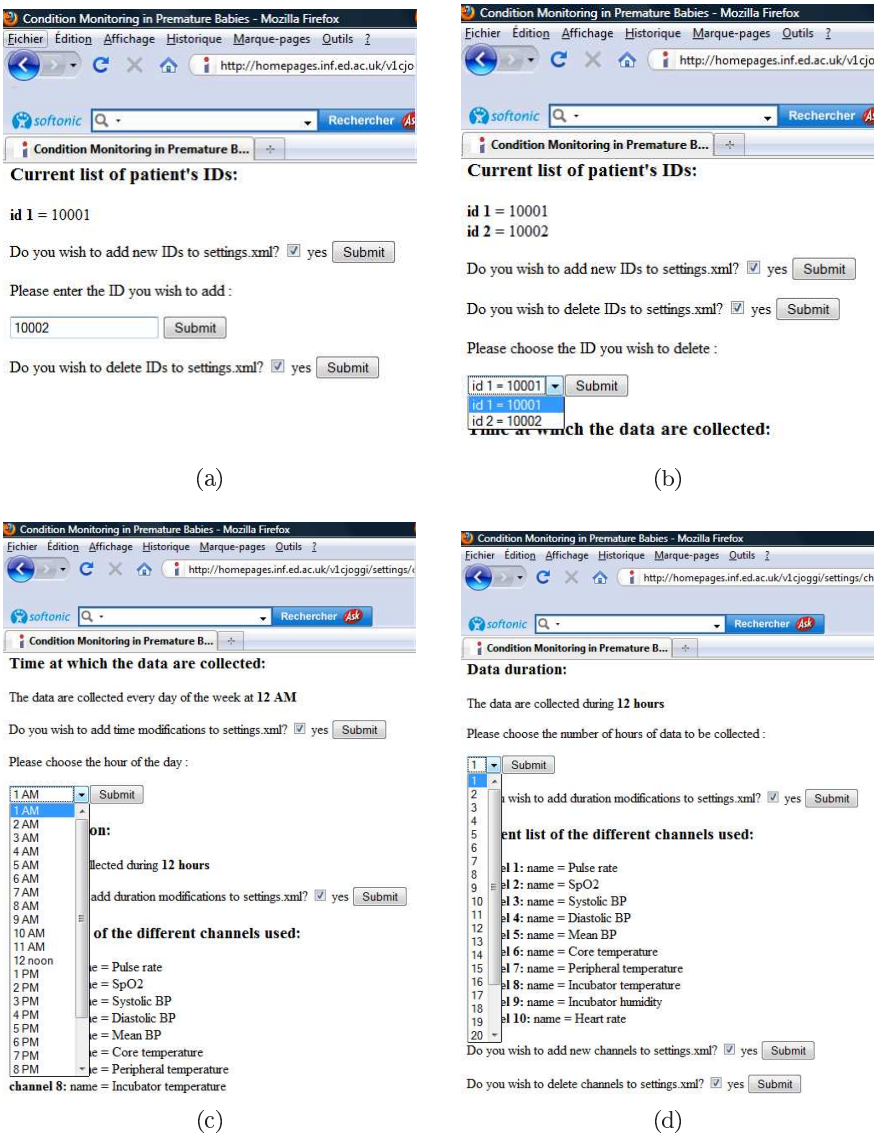


Figure 4.3: Illustrations of the (a) add and (b) delete baby's ID functionalities in the setting web application. (c) and (d) illustrate the time modification functionalities in the setting web application.

4.1.3 Webform application

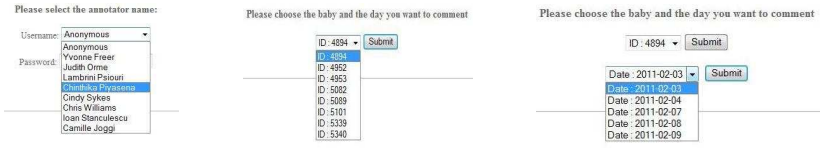
The web interface built to collect feedback from the clinicians is displayed in figure 4.4a. As shown in this Figure, the last version of this webform allows multiple annotators to comment on the same intervals.

Once the username and the password have been entered, the user is presented with a web interface where he can choose the baby's ID he wants to comment on (Figure 4.4b) as well as the date when the data have been recorded (Figure 4.4c).

Subsequently, after he has chosen the baby he wants to add comment to and the date when the data were recorded, the user is presented with the webform interface itself. A screen shot of the last version of this webform is displayed in figure 4.5. This version of the web interface is the most recent, but has evolved through time, in response to the requirements of the clinical practitioners, in order to make it clearer and more user-friendly. As can be seen in this figure there is a 'heading' box at the top of the webpage. The first line of this heading displays the username of the annotator, the recording date and the baby's ID. The second line displays a scroll down menu, which lists all the X-factor intervals for current date and baby's ID. By default, the first interval of the list is selected, although the user is able to choose the interval that he would like to comment on. The third line of the heading displays the details of the current interval that has been selected (i.e the exact time of occurrence and the score attributed to this interval; see section 4.1.1 above).

Further lines within the heading indicate if some bradycardia episodes have been detected by the FSLDS system during the current interval, and display the details concerning these episodes (i.e the exact time of occurrence).

The last line of the header asks the annotator: "In which of the following categories could the event happening during the [current] interval be classified". This question refers to the remainder of the form, where the annotator is required to tick boxes, indicating which of many proposed events are observed to occur for the selected data interval. These events are separated into two different types: factors which are included explicitly as 'known factors' by the FSLDS system ('events known') , and events which are not included as factors in the FSLDS system ('events unknown').



(a) (b) (c)

Figure 4.4: Snapshot of the webpage where, once the users have entered their username and their password their are presented with a web interface where their can choose (a) the baby's ID their want to comment on as well as (b) the date when the data have been recorded.

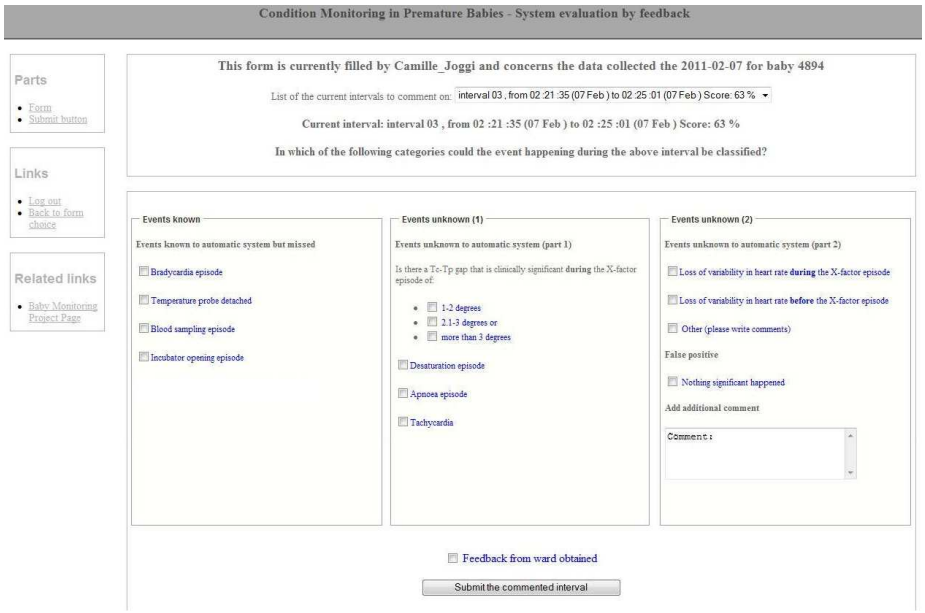


Figure 4.5: Snapshot of the webform application. The heading displays the username of the annotator currently filling the form, the date of recording of the data and the ID of the baby concerned by the data. The second line of the heading displays a scroll down menu, which lists all the X-factor intervals concerning the current date and baby's ID. The third line of the heading displays the details of the current interval selected by the user. The lines following the headings display the details concerning the potential bradycardia episodes detected by the FSLDS system during the current X-factor interval. The body of the webform lists different kind of events the annotator can tick. These events are displayed in two categories. The first category corresponds to the factors known to the FSLDS system, and the second category corresponds to the events unknown to the FSLDS system.

Events known to automatic system but missed: Events that are explicitly included as factors in the FSLDS system are listed in table 4.2. As described previously, only the X-factor intervals that do not overlap with other known artefactual factors are selected for evaluation by the clinicians (see Figure 4.1). Therefore, if the annotator ticks one of the known artefactual factors, then this implies that this factor has been missed by the FSLDS system, and wrongly claimed by the X-factor.

The webform was set up to display bradycardia events that were detected by the FSLDS system. Therefore, the annotators was only required to indicate bradycardia events that were missed by the system; that is, when none was listed

in the form. This would imply that the bradycardia event had been wrongly claimed by the X-factor.

Unfortunately, we only began to display whether bradycardia events were detected by the FSLDS system on the webform after beginning to collect feedback forms from the clinicians. Therefore, for these early feedback forms, a bradycardia event ticked by the clinician could also correspond to episodes that were indeed correctly detected by the FSLDS system. In addition, data from these early recordings, indicating whether bradycardia events had in fact been detected by the system during each of the intervals was not stored for data analysis. Because of this, we removed all of these earlier feedbacks from our analysis.

Events unknown to automatic system The second list of events in the form includes those that are not explicitly modelled by the FSLDS system. Because it is impossible to enumerate all of the possible events that could occur to the baby, a non-exhaustive list is terminated by a tick box labelled as 'other'.

False positive The box labelled as 'false positive' should be be ticked if nothing unusual can be observed in the selected interval. This implies that the 'X-factor' has been wrongly triggered.

Free text box At the end of the webform, a free text box can be filled in by the annotator, so that they can give additional comments about the selected interval. This is particularly important in the case where the tick boxes do not allow them to give the complete picture of what is happening. Specifically, if the label 'other' has been ticked, the annotator is expected to explain in detail what is happening here. If the label 'false positive' has been ticked, the annotator can optionally give an explanation of they belief of what could have wrongly triggered the X-factor.

Events known to automatic system but missed:	Bradycardia episode
	Temperature probe detached
	Blood sampling episode
	Incubator opening episode
Events unknown to automatic system:	Clinically significant gap between core and peripheral temperature
	Oxygen desaturation episode
	Apnea episode
	Tachycardia
	Loss of variability in heart rate during the X-factor episode
	Loss of variability in heart rate before the X-factor episode
	Other
False positive:	Nothing significant happened
Additional comment:	Free comment

Table 4.2: Lists of events displayed in the webform on the form of labelled events (with a tick box).

Feedback collection The feedback from the clinicians were recorded in a text file, as displayed in Figure 4.6. The file was named after the name of the annotator, the ID of the baby and the date of the recording concerned by the form. The file had a fixed format: each labelled event that had been ticked by the annotator was written with the corresponding numbered interval. Follow, on the next line, the term 'on' (to indicate that the box was ticked), or, for the free text box, the full text input of the clinician.

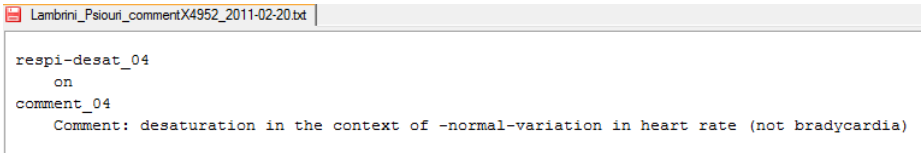


Figure 4.6: Snapshot of the text file in which the online forms filled by the clinicians are stored. The name of the file corresponds to '{Name of the annotator}_commentX{baby's ID}_{date of the recording (in the format YYYY-MM-DD)}.txt', Each event that has been ticked by the clinician is written in the file, linked by an underscore to the number of the corresponding interval. Follow, on the next line, the term 'on' (to indicate that the box was ticked), or, for the free text box, the full text input of the clinician. In this example, the annotator Lambrini Psiouri has commented about the baby 4952, about data recorded on 20th of February 2011. During the fourth X-factor interval, she has detected a desaturation ('respi-desat') and made the following comment about it: "desaturation in the context of -normal- variation in heart rate (not bradycardia)".

Categorisation of free text comments In order to analyse the free text comments provided by the clinicians, I categorised them manually into the following categories:

- Apnoea
- Artefact in heart rate
- Artefact in oxygen saturation
- Baby manipulation
- Bradycardia associated with desaturation
- Drop in heart rate
- Drop in oxygen saturation
- Feed
- Fluctuations in blood pressure
- Fluctuations in oxygen saturation and blood pressure related
- False positive in heart rate
- Normal oxygen saturation variability
- Rise in blood pressure
- Rise in heart rate
- Rise in heart rate variability
- Temperature gap (between central and peripheral temperature).

4.2 Results

We gathered annotations from 103 different intervals. Some intervals were annotated by more than one clinician, so that we collected 166 annotations in total. As stated previously, we removed 39 intervals that were commented before the webform was updated to its current form . This left us with 127 annotations of 103 different intervals, annotated by 5 different annotators.

4.2.1 Known factors performance

If a significant number of X-factor intervals were attributed by the hospital staff as corresponding to a particular 'known factor' (i.e. bradycardia, incubator open, blood sampling or temperature probe detached), then this would imply that such events had been repeatedly missed by the system. Consequently, this would indicate that the LDS model corresponding for this factor was not accurate enough to detect these events, and needed to be improved. Conversely, if only a small number of X-factor intervals were attributed by the clinicians as corresponding to a particular known factor, then this would indicate that the LDS model for this factor was good enough.

Only two 'bradycardia' and two 'incubator open' events were found among the X-factor intervals. Three blood sampling episodes were found by the annotators, but blood pressure data were missing for two of these three episodes, effectively making it impossible for the system to detect this category of event. Finally, no temperature probe detached events were flagged by the annotators. These results, summarized in figure 4.1.3a, show that only a very small number of events corresponding to factors that were explicitly modelled in the FSLDS system were missed, demonstrating the accuracy of the system for detecting these known factors.

4.2.2 Main clinical events behind the X-factor

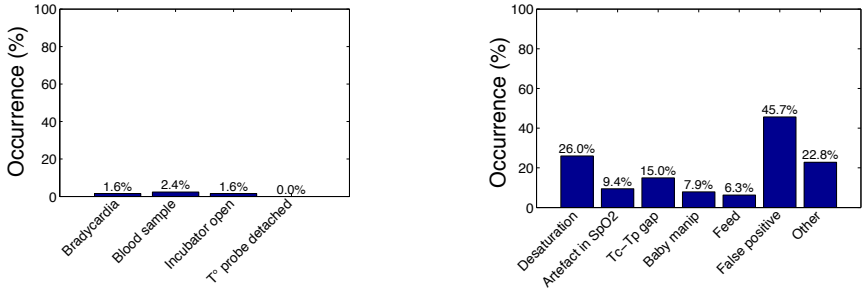
The goal of this evaluation process was to identified recurrent patterns in the data, which could potentially be added as additional factors to improve the model. Indeed, out of the many labelled events that could have been attributed to the X-factor intervals, we found that the annotators responses were dominated by only a few of the many different labelled events. Therefore, we decided to restrict the analysis to only the factors that were appearing in more than 5% of the answers.

Out of the 11 types of event that were labelled in the webform (figure 4.5), and 16 categories of free comment (see section 4.1.3), nearly all X-factor intervals were attributed by the clinicians as corresponding to only seven main types of event. The percentage of intervals classified as corresponding to these seven types of event is shown in figure 4.7b: 26.0% of the intervals included desaturation events, 9.4% included an SpO2 artefact, 15% included a gap between central and peripheral temperature ('Tc-Tp' gap), 7.9% included a baby manipulation, 6.3% included a baby feeding event, 45.7% were false positive,

and 22.8% were classified as due to an unlabelled factor ('other'; note that each interval could correspond to more than one type of event).

Clinically, it is known that bradycardia events are often accompanied by other complications, such as oxygen desaturations (see section 2.2). Our data is consistent with this: a high proportion of X-factor intervals also included bradycardia events that were detected by the FSLDS system (32.3 % of X-factor intervals). This may imply that many of the unknown events that were missed by the system correspond to complications that arise alongside bradycardia events. Taken together, intervals with bradycardia events detected by the FSLDS system, and intervals that were categorized by the annotators as due to one of the 7 main event types (desaturations, SpO2 artefacts, false positive or other) accounted for 97.6% of X-factor intervals.

More than 5% of events included a gap between core and peripheral temperature (15%). However, from discussion with clinicians, we established that different annotators used different definitions from this event (see section 4.3 below for details), meaning that their responses where highly ambiguous. Because of this ambiguity, we decided not to focus on their responses for this category of event in the following sections.



(a) Percentage of occurrence of the events known by the FSLDS system

(b) Main events flagged by the annotators

Figure 4.7: Classification of the X-factor episodes by the clinicians. (a) Events known by the FSLDS system (known factors) but missed. The small proportion of these events that have been flagged up by the annotators confirm the accuracy of the FSLDS system for detecting the known factors. (b) 7 main categories of events flagged by the annotators. 'Artefact in SpO_2', 'Feed' and 'Baby manip' corresponds to categories that I inferred from the free comments. The other categories corresponds to events labelled in the webform (with tick boxes).

False positive False positives are the most common category of X-factor intervals. In order to investigate which clinical events are associated with the episodes flagged as 'false positive' by the annotators, we display in Figure 4.9a

the main categories of events that occurred simultaneously with false positive events.

Approximately half (51.7% of false positive episodes) of the episodes classified as false positive were not commented on in the free text box, making it difficult to explain what could have wrongly triggered the X-factor in these cases (figure4.9a). In the following paragraphs, we describe possible causes of these false positive events.

One possible cause of false positives could have been fluctuations in the recorded heart rate. Indeed, a large proportion (20.7% of false positive intervals) of intervals labelled as false positive by the annotators were accompanied by a bradycardia event detected by the FSLDS system. However, many of these detected bradycardia events (8.6 % of false positive intervals, - i.e. 41.54 % of bradycardia events detected by the FSLDS system during false positive intervals) were described by the annotators as incorrect (in their free text comments). In addition, also in their free text comments, the annotators described a significant proportion of false positive intervals as due to a 'drop in heart rate' (6.9 % of false positive intervals; figure 4.9a). Therefore, it is likely that many of the intervals that were labelled as false positive were triggered by higher variability in the heart rate, including drops in heart rate that were too small to be classified as clinically significant (i.e. as bradycardia). Example of such an event is shown in Figure 4.8a. In this figure, we can see that an event identified as bradycardia by the system is triggered by increased variability in the measured heart rate, including a small and short dip where it drops to 132 bpm - larger than the defined threshold for bradycardia (100 bpm or 33% below baseline) [21]. These results suggest that improving the normality model to account for these periods of minor instability in the heart rate could significantly reduce the number of false positives picked up by the system.

It is also possible that some of the false positives could have been due to variations in oxygen saturation, reflected in the SpO_2 measurements. In 19 % of false positive intervals the clinicians indicated in the free text box, that the SpO_2 variability was 'clinically normal'. However, on visual inspection, these commented intervals appeared to feature variations in SpO_2 that were larger than average. Therefore, one interpretation is that these comments corresponded to intervals where the X-factor was triggered by above average changes in SpO2 levels, but which were too small to be clinically significant. Further work will need to be performed to establish whether this is indeed the case.

As stated previously 'false positives' were defined to correspond to intervals that were incorrectly flagged up by the X-factor, when there were no artefacts or clinically significant physiological changes (i.e. they should have been accounted for by the 'normal' dynamics in the model). However, 13.8 % of the intervals classified as false positives by the annotators were also classified by them as featuring artefactual changes in SpO_2 levels (for example, due to changes in the pressure applied to the probe; see section 2.3). Instead, according to our definition of a false positive, these intervals should have been categorised as artefactual SpO_2 events, but not false positives (as an artefact was taking place - the data should not be modelled by 'normal' dynamics). Thus it appears that

we did not explain clearly enough to the annotators what was meant by a 'false positive'. As a result, they thought that 'false positive' meant 'no clinically significant physiological event happened during this interval', when in fact, we meant 'nothing significant happened'. In the future, we could alter the structure of the webform to overcome this confusion.

Bradycardia Bradycardia is often detected by the system during the X-factor intervals. This could be because bradycardia is often accompanied by other clinical events that alter the recordings from other channels in ways that cannot be accounted for by the model, triggering the X-factor (see section 2.2). The events that were found to occur most often alongside the bradycardia events were oxygen desaturation, baby manipulation and feeding. In addition, some episodes of bradycardia detected by the system were classified by the annotators as false bradycardia and false positive. In Figure 4.9b we plot a histogram, showing the proportion of different types of event that occurred simultaneously to bradycardia. The bradycardia events displayed in this figure include both those that are detected by the system and those that are flagged up by the annotators. In total, they add up to 43 bradycardia events.

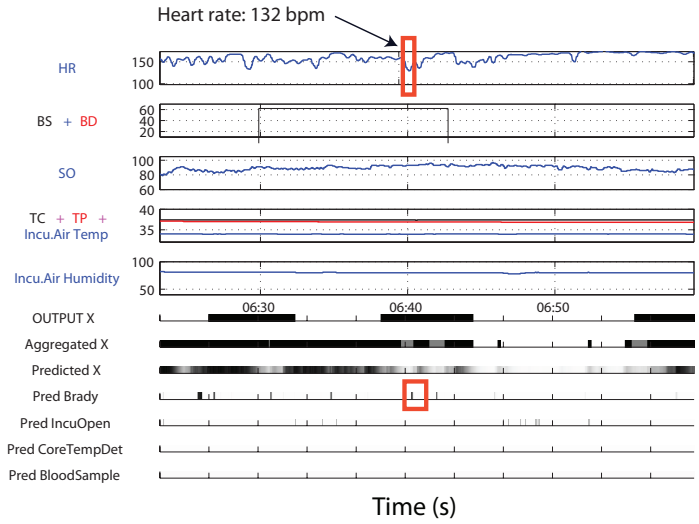
Unsurprisingly, the main category of events associated with bradycardia episodes is oxygen desaturation, which occurs for 44.2% of the bradycardia events. This might be expected from a physiological and a clinical point of view. As stated earlier, bradycardia and oxygen desaturation often occur together (section 2.2). This is because they are both consequences of hypoxaemia (decreased partial pressure of oxygen in the blood) which can be caused by apnoea. Oxygen desaturation is a direct consequence of a hypoxaemia episode, whereas bradycardia is triggered by a reflex that is induced by the hypoxaemia episode, and reinforced by the cessation of lung inflation [21, 23].

As shown in Figure 4.9b, almost 7 % of bradycardia events occur concurrently to body manipulations. Out of these events, around half are associated with oxygen desaturation (50 % of events where bradycardia and body manipulations occur together). Clinically, it is known that hypoxaemia episodes, can be triggered by excessive handling of the baby [27], and that hypoxaemia leads to both desaturations and bradycardia episodes [23] (see section 2.2). This could explain why bradycardia episodes are frequently observed alongside body manipulations.

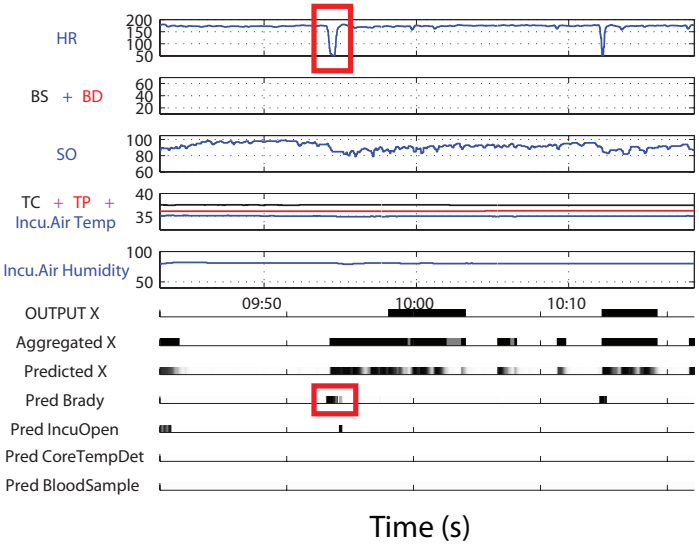
Bradycardia events often occur during feeding (7% of bradycardia events; figure 4.9b). Indeed, it is known that feeding is often accompanied by other physiological complications (section 2.2). This is thought to be due to the immaturity of the premature baby's brainstem, causing problems in coordinating the acts of sucking, swallowing, and breathing [21].

A large number of bradycardia episodes detected by the FSLDS system were flagged by the annotators as false positive (27.9% of bradycardia events; Figures 4.9b), and as false bradycardia (23.3% of bradycardia events; Figures 4.9b). This suggests that some clinically insignificant drops in heart rate are often mistaken by the system as corresponding to real bradycardia episodes (see

previous section). Finally, discussions with the hospital staff also showed that the system was sometimes unable to discriminate between obvious artefacts in the heart rate, such as when the probe failed to detect a heart beat, and a real bradycardia. An example of such an event is shown in Figure 4.8b.



(a) Non clinically significant drop in heart rate



(b) Artefact in heart rate

Figure 4.8: Events wrongly flagged up as bradycardia by the FSLDS system. (a) A non clinically significant drop in heart rate (see top red square) has been picked up by the bradycardia factor (see inferred distribution of bradycardia factor highlighted by bottom red square). (b) Artefact in heart rate (top red square) mistaken with a real bradycardia by the FSLDS system (bottom red square).

Desaturation Oxygen desaturation is very common during the X-factor intervals (26 % of the X-factor interval; Figure 4.7b). Figure 4.9c displays the main categories of events occurring simultaneously to the oxygen desaturation episodes.

Unsurprisingly, given our previous analysis of the bradycardia events, the main categories associated with oxygen desaturation episodes are bradycardia (occurring for 57.6 % of the oxygen desaturation intervals - 51.5 % detected by the system, and 6.1 % annotated by the clinicians) and drops in heart rate (occurring for 3 % of the oxygen desaturation intervals). As before, this can be explained by the close association between bradycardia and oxygen desaturation, which are both consequences of hypoxaemia (low partial pressure of oxygen in the blood) and often associated with apnoea [21, 23] (see section 2.2).

In common with bradycardia, a significant fraction of the desaturation episodes occur simultaneously to body manipulations (12.1 % of desaturation intervals, Figure 4.9c). Out of these events, approximately half are associated with bradycardia episodes as well (50 % of events with both body manipulations and desaturation). This is likely explained by the fact that excessive handling of the baby can trigger hypoxaemia episodes, leading to desaturations and bradycardia episodes [27] (see section 2.2). Also in common with bradycardia, some of the desaturations occurred during feeding [21] (7.3%, or 3/41 desaturation episodes).

A small proportion of desaturation episodes are classified by the clinicians as corresponding to apnoea (9%). We saw in the theoretical section (see section 2.2), apnoea is a very common condition in premature babies, and is most of the time accompanied by both bradycardia and desaturation. Therefore, it is perhaps surprising that so small a number of apnoea episodes occurred during desaturation episodes. However, an apnoea episode is mainly detected by directly observing the baby, as the respiratory rate reading is unreliable because it is riddled with body movement artefacts (see section 2.2). Therefore, the small number of reported apnoea events can be explained by the fact that it is almost impossible, by just looking at the traces, to tell if the desaturations and the bradycardia are due to apnoea or to some other reason.

Finally, only one oxygen desaturation episode was classified as a false positive. This implies that desaturations are almost always clinically significant. This can be clearly understood clinically, since oxygen desaturation is the marker of hypoxaemia episodes (see section 2.2), a reduction of partial pressure of oxygen in the blood, that is potentially dangerous for the baby [19].

Other Figure 4.9d displays the main types of event associated with the 30 episodes categorised as 'other'. These categories were more numerous than for the other main event types; to reflect this, we plot the 8 types of event that occurred most frequently when the annotators selected 'other'.

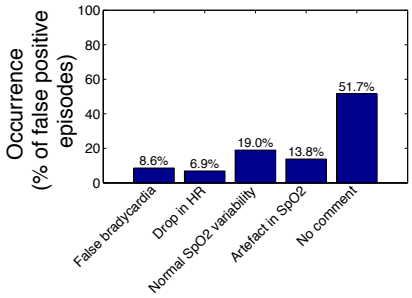
The large number of events associated with this category ('other') comes from the fact that the annotators, when ticking the 'other' category, gave more details in the free text box about the event occurring at the give interval. There-

fore, most categories displayed in Figure 4.9d correspond to categories derived from the free text comments.

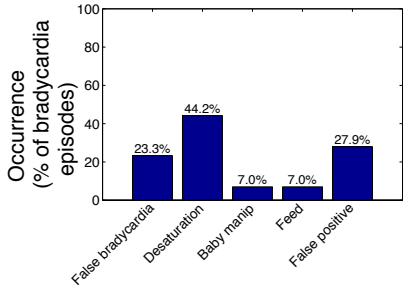
The information extracted from the analysis of this category is useful to determine the events that can be added to the webform in the future.

Baby manipulation In our analysis of the bradycardia and desaturation events, we found that these episodes often occurred simultaneously with periods of baby manipulations (7 % of bradycardia episodes and 12.1 % of desaturation episodes; Figure 4.9b and 4.9c). Figure 4.9e plots the main categories of events accompanying the 14 baby manipulation episodes. In addition to the categories display in Figure 4.9e, 14.3 % of baby manipulation episodes include a bradycardia detected by the system. We see that bradycardia, oxygen desaturation and a drop in SpO_2 (which can be view as a minor desaturation) occur repeatedly during baby manipulation episodes. This could be explained by the hypothesis suggested in [27], that excessive handling can trigger hypoxaemia episodes, leading to desaturations and bradycardia episodes (see section 2.2).

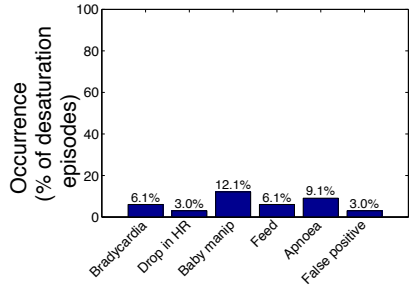
Feeding We found that both bradycardia and desaturation episodes often occurred simultaneously with periods of feeding (7 % of bradycardia episodes, 6.1 % of desaturation episodes; Figures 4.9b and 4.9c). Figure 4.9f plots the main categories of events accompanying the 9 feeding episodes. In addition to the categories display in Figure 4.9e, 33.3 % of feeding episodes include a bradycardia detected by the system. Similar to baby manipulation events, bradycardia, oxygen desaturation and drops in SpO_2 occur repeatedly during feeding episodes. This could be explained by the hypothesis mentioned in section 2.2, that feeding-related episodes of bradycardia and desaturation can be clinically explained by the brainstem immaturity in coordinating the acts of sucking, swallowing, and breathing [21].



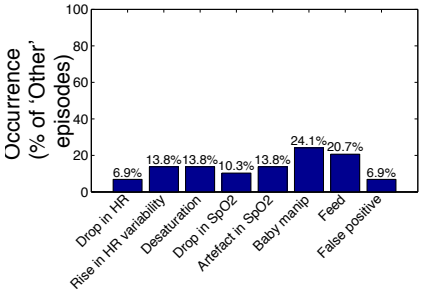
(a) Classification of false positive episodes



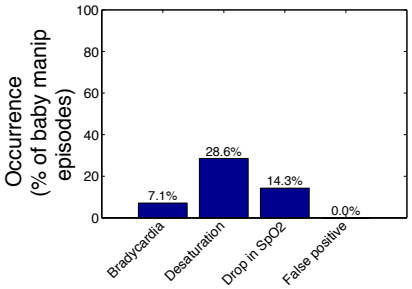
(b) Classification of bradycardia episodes



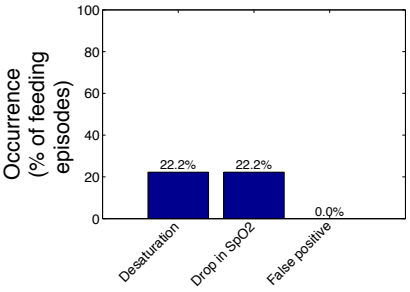
(c) Classification of oxygen desaturation episodes



(d) Classification of "Other" episodes



(e) Classification of Baby manipulation episodes



(f) Classification of Feeding episodes

Figure 4.9: Main categories of events that are happening simultaneously with (a) 'false positive', (b) 'bradycardia', (c) 'oxygen desaturation', (d) 'other', (e) 'baby manipulation' and (f) 'feeding' events. 'T°' holds for 'temperature'. 'FSLDS bradycardia' holds for bradycardia detected by the FSLDS system. 'False bradycardia', 'Drop in HR', 'Normal SpO_2 variability', 'Rise in HR variability', 'Drop in SpO_2', 'Artefact in SpO_2', 'False bradycardia', 'No comment', 'Baby manip' and 'Feed' correspond to categories that I inferred from the free comments.

4.3 Discussion

In previous work, my laboratory constructed a system that can be used to infer the health of premature babies from collected physiological recordings. In order to use this system in a clinical setting, as well as to discover what needs to be improved, it is important to evaluate the performance of the system using real physiological data. However, such an evaluation requires clinical expertise, so that the true clinical interpretation of the physiological data can be established.

The goal of the project described here was to set up a process whereby the system could be evaluated. This involved constructing a webform, which provided an interface for clinical experts to give feedback about sections of the data that were deemed to be of interest (i.e. intervals of data that the system was unable to attribute to either the healthy state, or known physiological or artefactual factors - 'X-factor' intervals). We used the collected clinical feedback to perform an initial evaluation of the system performance.

We found that the system was able to deal well with the artefactual factors that were explicitly included in its construction ('known' factors). The majority of X-factor intervals corresponded to a small number of types of event, that were not explicitly included in the system construction ('unknown' factors). We analyzed these additional unknown factors, using clinical knowledge to explain why, and in what situations, they might come about. We discuss in the following sections how this information might be used to improve future versions of the system.

Improvement of the system

Oxygen saturation The results obtained from our evaluation suggested that a significant proportion of X-factor episodes flagged as false positive by the annotators included great variability on the SpO_2 readings, that was commented by the clinicians as being clinically normal (see section 4.2.2). If this hypothesis is confirmed by further quantitative analysis, it means that the normality model for SpO_2 reading can be improved by allowing more variability. This would hopefully result in a decrease in the proportion of false positives.

Previous studies on the FSLDS model have commented on the fact that oxygen desaturation events constitute an important feature of the clinical data, forming a very characteristic pattern that can be clearly identified (see Figure 4.10) and appears repeatedly. This work also observed that such events occur most often during bradycardia events [10]. The results obtained from our evaluation of the clinical feedback are consistent with these observations. We found that a significant proportion of X-factor episodes were classified by the clinical experts as oxygen desaturation (see Figure 4.7b). Moreover, almost no desaturations were classified as false positive, showing that this type of event is pratically always clinically significant (see Figure 4.9c). Consistent with the previous study, we found that a significant proportion of desaturation episodes occurred concomitantly with bradycardia events (or smaller drops in the heart rate; see Figure 4.9c).

These results suggest that adding a new oxygen desaturation factor to the model would greatly reduce the number of X-factor events. Moreover, to reflect what was reported in the feedbacks, the bradycardia and desaturation factors could be made depend on each other, so that when one of them was active the probability of the other being simultaneously active was increased. Hopefully, this would also help the model to detect other related clinically significant events such as apnoea, for which oxygen desaturation closely followed by bradycardia events are the strongest markers [21]. Thus, by adding only one factor to the model, we would be able to account for more than one additional type of clinically significant event, that could endanger the health of the baby.

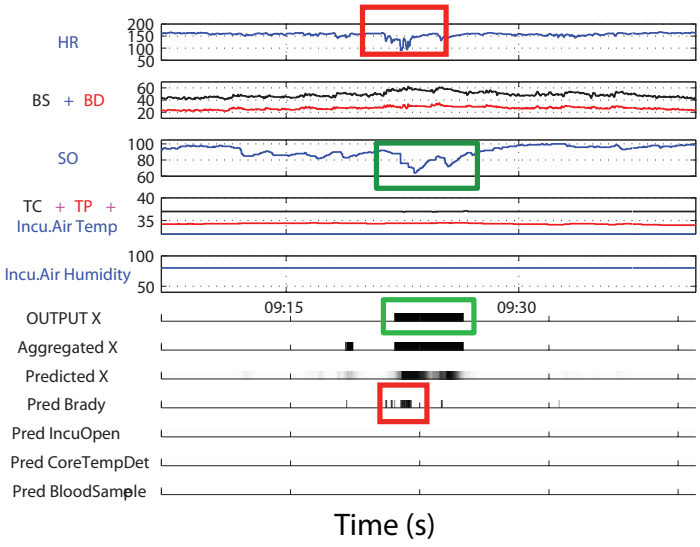


Figure 4.10: Bradycardia episodes accompanied by an oxygen desaturation. In this example, the bradycardia episodes (top red square) have been correctly picked up by the system (see bottom red square highlighting the inferred distribution of the bradycardia factor). However, the drop in oxygen saturation (desaturation; top green square) that accompanies the bradycardia episode does not fit with the normality model describing the normal oxygen saturation dynamic. Therefore, this drop in SpO_2 is picked up instead by the X-factor (see bottom green square highlighting the X-factor interval).

Artefactual and true desaturation Among the feedbacks gathered, a significant proportion of X-factor intervals were categorised as due to artefact in the SpO_2 readings (Figure 4.7b). Indeed, artefacts in SpO_2 readings are often difficult to distinguish from real oxygen desaturations (see section 2.3) [1]. This was also communicated to us verbally by the hospital staff, during discussions

about the feedback forms. An example of an interval for which different annotators gave different interpretations (one reporting a true desaturation, and the other an SpO_2 artefact) is shown in Figure 4.12b.

In order to distinguish between true oxygen desaturations and artefacts, the hospital staff compare the pulse rate obtained with the oximeter with the heart rate obtained with the ECG (see section 2.3). These two readings should be identical, and if they are not, it strongly suggests that the oximeter reading is not reliable - probably as an artefact due to body movements is altering the SpO_2 readings. In fact, this method of discriminating between artefacts and true desaturation constitutes the basis of an existing computer algorithm (Edentec Motion Annotation System) that to is able to efficiently discriminate between artefactual and true SpO_2 measurements [30]. This could be incorporated into our system by including an artefactual SpO_2 factor, which altered the pulse rate measurements from the oximeter, without changing the heart rate readings. Therefore, changes in the oximeter reading that are not accompanied by a corresponding change in heart rate would be attributed to this artefactual SpO_2 factor.

Adding this additional factor should have a large effect on the number of false alarms triggered by the system. A previous study, investigating false alarm in neonatology units found that the differences in false alarm rates between different systems were mostly due to how the different systems dealt with artefactual changes in pulse oximeter and $PTcCO_2$ readings, each of which contributed at approximately $40\,\%$ to the total number of alarms [7].

Heart rate A significant number of bradycardia events detected by the FSLDS system were marked as false by the annotators (Figure 4.9b). This indicates that either an artefact present in the heart rate measurements, or clinically insignificant changes in the heart rate were mistaken by the system as real bradycardia episodes. The large proportion of these episodes that were simultaneously flagged as false positives by the annotators supports the hypothesis that a number of bradycardia episodes detected by the system were due to clinically insignificant drops in the heart rate (which should be included within the 'normal' dynamics), rather than some other artefact. These observations demonstrate that the model could be improved, by either altering the normality model to account for clinically insignificant changes in the heart rate, or by altering the bradycardia model so that it was only triggered by large, clinically significant, changes in the heart rate (defined as $< 30\%$ below baseline).

In addition, sometimes artefactual changes in the heart rate occurred when the probe missed a heart beat. This resulted in a very sudden drop in the measured rate, which could be falsely detected by the system as a bradycardia event (Figure 4.8b). The FSLDS model could be easily improved to deal with this artefact, adding a new factor ('skipped heart beat' factor) that would be triggered by very sharp changes in the heart rate.

Periodic pattern One clinically significant pattern was encountered very rarely, but is worth mentioning here because of its very important clinical significance. This pattern, displayed in Figure 4.11, consists of periodic waves in the SpO_2, blood pressure and heart rate readings, oscillating concomitantly. This dynamic reflects the immaturity of the brainstem in premature babies. Specifically, it can be cause to the immaturity of the respiration control system: causing an episode of periodic breathing, a condition where the pre-term alternates between breathing and apneic episodes (see section 2.1) [18, 44]. Alternatively, this pattern in the sequence of measurements can be caused by the immaturity of cardiac output regulation system. From discussions with the clinicians, it stands out that this pattern always indicates that the life of the baby is in danger, whatever its exact cause. As it is both clinically important, and easily identifiable, this pattern could be usefully added as an additional factor for the model.

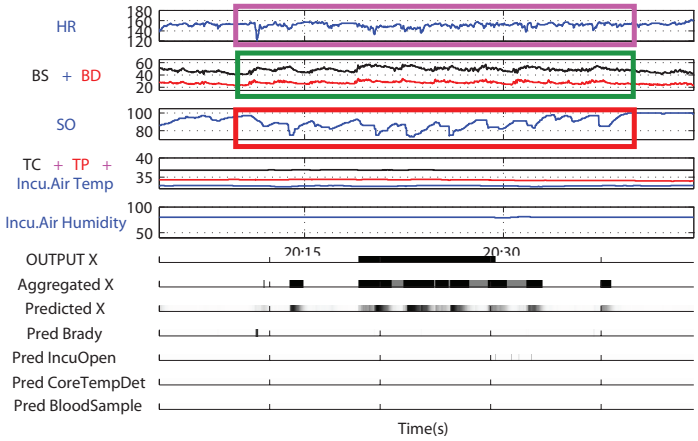


Figure 4.11: Illustration of a periodic pattern described as clinically dangerous for the baby's life. This pattern is characterised by periodic waves in SpO_2 (red square), blood pressure (green square) and heart rate (purple square) readings that oscillate concomitantly.

Multiple annotations considerations

Some of the intervals were annotated by more than one clinician. Comparing annotations obtained from different clinicians for the same interval can provide useful information. First, it can highlight events that are particularly difficult to interpret. Second, it shows which questions in the webform are confusing and should therefore be clarified.

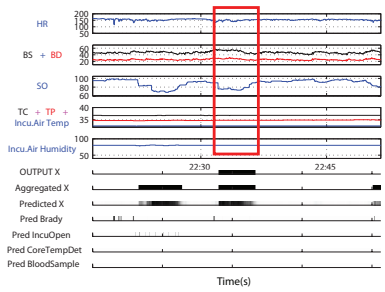
Measuring the variability between the responses from different annotators tells us how reliable their responses are in identifying different types of event.

This could help us to determine how closely we should try and make our model match these individual responses. As an extreme example, it would not make sense to try and aim for a 100% match between the model and the annotator responses, if the annotator responses themselves were only 50% reliable. Unfortunately, from our data there were relatively few intervals annotated by more than one clinician. Therefore, while we can gain some idea about the types of disagreement that occurred for different events, a more quantitative analysis will require further data to be collected.

One type of disagreement between annotators occurred when the physiological data was ambiguous, and could thus have many different interpretations. Figure 4.12b shows a good example of this. Here, a single interval has been interpreted differently by the different clinicians, two of them reporting an oxygen desaturation event and the other reporting an SpO_2 artefact. This is explained by the fact that it is extremely difficult to discriminate between a real oxygen desaturation and an artefactual event (section 2.3)[1].

A second type of disagreement between annotators occurred not because the physiological data itself was ambiguous, but because different clinicians disagreed on the definition of a certain event. For example, this was the case for $Tc - Tp$ gap events (gap between central and peripheral temperature of one or more degrees). Although these events were flagged up for a relatively large proportion of X-factor intervals (figure 4.7b), it was extremely difficult for us to analyse this data, as we were informed by the hospital staff that different clinicians were not using the same definition to identify this event (see figure 4.12d for an illustration). Such types of confusion could be avoided by providing the annotators with clear written guidelines on how each of the events should be defined for the form.

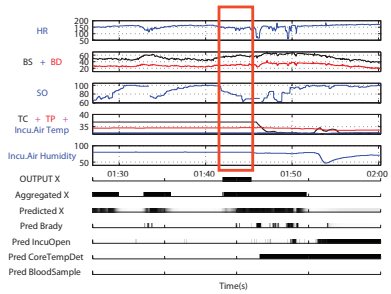
A related issue is the disagreement between different annotators about the definitions of 'false positive' and 'other' in the form. In the example shown in figure 4.12d, we see that one annotator classifies an interval as 'false positive' while another annotator classifies the same interval as 'other', even though they both give the same description of the interval (artefact in SpO_2). This disagreement appears to be partly due to the weak distinction between 'clinically significant' and 'not clinically significant' events in clinical practise. Also it may partly be due to the lack of clarity in the way the question is asked in the webform, leading some annotator to think that 'false positive' means 'no clinically significant physiological event happened during this interval' (which would mean that each artefact is a 'false positive'), when in fact, we meant 'nothing significant happened' (which means that an artefact should not be classified as 'false positive'). Further improvement of the webform, as well as clear written guidelines, could aim at addressing this problem.



(a)

	Anonymous	Chinthika Piyasena	Lambrini Psiouri
SpO_2	artefact in oximeter	Desaturation	artefact in oximeter
	False positive	-	False positive

(b) Illustration of the difficulty to discriminate between artefact in SpO_2 readings and a true desaturation.



(c)

	Anonymous	Judith Orme
Tc-Tp gap	-	Tc-Tp (not significant)
False positive	False positive	-
Free text comment	artefact in SpO_2	artefact in SpO_2

(d) Illustration of the confusion of the annotator about the definition of the events 'Tc-Tp gap', 'False positive' and 'Other'.

Figure 4.12: (a) and (b) Illustration of the difficulty to discriminate between artefact in SpO_2 readings and a true desaturation. In this example, one annotator classified the interval highlighted by the red square in (a) as a true desaturation, while two other annotators classified the same interval as an SpO_2 artefact). (c) and (d) Illustration of the confusion of the annotator about the definition of the events 'Tc-Tp gap', 'false positive' and 'Other'. In this example, the two annotators agree on the fact that an artefact in SpO_2 is occurring during the interval highlighted by the red square in (c). However, one annotator also classify this interval as false positive (rather than 'other'), showing that the definition of 'false positive' is unclear to the annotators. The same lack clarity stand out for the definition of 'Tc-Tp gap' leading one annotator only to tick the 'Tc-Tp gap' box.

Webform quality considerations

This evaluation relies on a collaboration between the machine learning lab and the clinical experts who annotate the "X-factor" intervals. One of the main difficulties faced was to set up a form with questions that were clear enough to be understood by the clinicians, but precise enough so that the answers could be used to evaluate and improve the system. This is particularly difficult, because it implies an interaction between two different specialised fields of knowledge (i.e. medical and mathematical knowledge).

Therefore, it was decided to begin with a very simple webform. We would then modify the webform, based on the answers obtained from the filled online forms and from direct feedback received from the clinical staff. The adaptation and improvement of the webform application required a lot of technical work in order to make it as user-friendly as possible.

The downside of this method is that we started to get results while the webform was still not at its final form. Notably, the bradycardia episodes detected by the system during the X-factor episodes were not indicated in the webform to begin with, and data from early recordings, indicating whether bradycardia events had in fact been detected by the system during each of the intervals was not stored for data analysis (section 4.1.3). Therefore, all the comments about bradycardia received during this earlier part of the project had to be removed from our analysis.

Another problem concerned questions that were considered too confusing for the clinical staff, and which had to be clarified. For example, the first part of the form was entitled "Known events" to start with, and listed all the factors already modelled by the system, such as 'incubator open', 'temperature probe disconnected', 'bradycardia' or 'blood sample'. The idea was that the annotators had to select an event if it was happening during the X-factor interval. But this title was confusing for the clinical staff, because it seemed to mean that all these events were in fact detected by the system in the concerned interval. In order to address this problem, the title has been changed to "Events known by the system but missed".

Another example of a question that had to be modified in order to make the form more clear is related to the temperature of the baby. Originally, the form included a question asking: "*is there a gap between central and core temperature during the X-factor episode [...]?*". This question was confusing to the annotators, because they did not know whether they had to select this event only if the temperature gap was new (it had started during the X-factor interval) and only if it was clinically relevant. This is because the annotators assumed that we were not wanting to collect feedback about anything that could be easily established just by looking at the data; that this information would be gathered by us, and not require medical expertise. Therefore, this question was changed to the following: *"is there a gap between central and core temperature that is clinically significant during the X-factor episode [...]?"*. While hopefully this

should improve the quality of their responses, this change was made at the end of the project, and therefore we have not collected enough data to establish whether this is the case.

One of the weakness of the webform is that it has a very small amount of relevant categories, and thus, a great proportion of the important information was acquired from the free text comments. In order to perform quantitative analysis on these comments, I had to classify each of the free comments manually as corresponding to a set of categories that I determined by myself. As well as being very laborious, this process introduces a degree of subjectivity to the results. To reduce this problem, in the future, the most frequent types of description given in the free comments should be incorporated as labels (with 'yes/no' answers) in future version of the webform. Examples of such additional questions that could be added to the webform are given as follows.

A large number of SpO_2 artefacts were been described in the free comments (figure 4.9d), indicating that this event should be added in the list of 'unknown events' in the webform. This would be particularly useful if it is planned to model the oxygen desaturation factor so that the system is able to discriminate between true desaturations and artefactual SpO_2 events. Our analysis indicated that the system was unable to distinguish between small, clinically insignificant, drops heart rate, and larger drops, associated with bradycardia. To determine whether this was the case, the webform could be adapted to include a main question "drop in heart rate?", followed by two sub-questions, "clinically insignificant" and "clinically significant (bradycardia)". Finally as it is thought that body manipulations and feeding events induce physiological problems, such as desaturation and bradycardia (section 2.2, 4.2.2 and 4.2.2), it would be interesting to add the events "baby manipulation" and "feeding" to the form, to confirm whether this is indeed the case. In addition, the annotator could be asked to give information about whether theses events are related to any concomitant bradycardia and/or desaturation event.

Chapter 5

Incubator open factor re-modelling

In the previous chapter, we described an automatic way to evaluate the performance of the FSLDS system in classifying physiological data from premature babies, based on expert feedback from clinicians. Early feedbacks from this evaluation established that the X-factor was often firing just after handling (i.e. incubator open) episodes, even though the baby was in a normal physiological state. This suggested that the X-factor was picking up the recovery of the humidity readings to set level occurring after the closing of the incubator (see Figure 5.1). The more likely explanation for this behaviour is that the humidity readings during recovery to set level show values far away from the normal range of humidity readings (see top green square in Figure 5.1). Therefore, this recovery phase was not classified by the normality model (describing the normal regime of incubator humidity readings), and was consequently picked up by the X-factor (see section 3.2.3). As the recovery of humidity to set level is an artefactual event (i.e. it does not reflect the physiological state of the baby), the X-factor should not account for it.

This problem was addressed by better modelling of the incubator open factor (which is called as well 'IO' or 'handling' factor, as the incubator is often opened to handle the baby), so that the re-modelled factor was able to account for the recovery of the humidity readings to the set value after the closing of the incubator. Previously, the IO factor was described by two states:

1. Incubator closed - humidity maintained constant (described by the 'normality' model for the humidity channel)
2. Incubator open - humidity declining (described by the 'incubator open' / 'decay' model)

This factor was extended by adding a third state, the model of which describes how the humidity recovers to the set level after the incubator has been opened and closed.

In this chapter, I aim at describing the different steps followed to re-model the IO/handling factor, integrate the new model in the FSLDS system and evaluate the performance of the extended system.

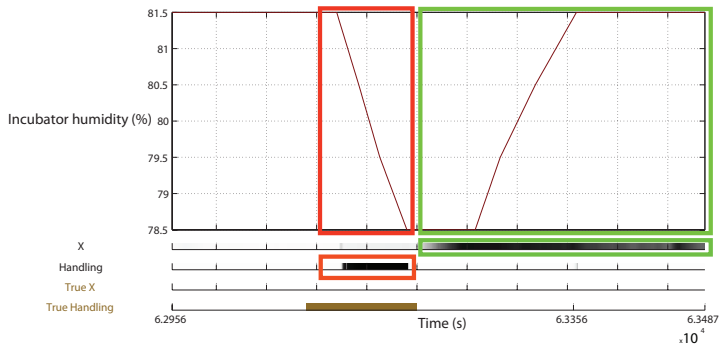


Figure 5.1: Example of incubator humidity readings during the opening and closing of the incubator, with the corresponding probability distribution of the IO/handling factor and the X-factor inferred by the FSLDS system. The gold panels display the segments labelled as 'true' IO/handling or X-factor episodes by clinical experts. The decay of the incubator humidity (see top red square) is picked up by the IO factor (see bottom red square). However, the IO/Handling factor does not account for the recovery of the incubator humidity (see top green square), and nor does the normality model of incubator humidity. Therefore, this recovery is picked up instead by the X-factor (see bottom green square).

5.1 Methods

This section describes the different steps that have been followed to choose the model describing the recovery stage of the new incubator open/handling factor, estimate the parameters, and evaluate how well the chosen model with its estimated parameters fits the data. We explain the modifications that have been done to the system in order to integrate this new model as a new stage of the IO/handling factor. Finally, we present the tool that has been used to evaluate the performance of this extended factor in classifying data: the ROC analysis.

5.1.1 Model construction

Looking at examples of recovery phase segments, such as the one displayed in Figure 5.2, it can be observed that the humidity recovery curve shows a similar behaviour than a capacitor charging (Figure 5.3). This motivates the recovery

phase to be modelled as an inverted exponential decay, where x_t is expressed as:

$$x_t = (x_{max} - x_0)(1 - a^t) + x_0 \tag{5.1}$$

where $a < 1$, x_{max} corresponds to the limit value of the process (i.e. the set humidity value in our case) and x_0 corresponds to the initial value of the process.

Equation 5.2 is expressed in the form of an autoregressive process by following the steps described below:

Method:

Starting from the previous equation (equation 5.2), which describes the time evolution of the humidity :

$$x_t = (x_{max} - x_0)(1 - a^t) + x_0 \tag{5.2}$$

This expression can be rearrange to obtain:

$$(x_t - x_{max}) = -(x_0 - x_{max})a^t \tag{5.3}$$

We can write a similar expression for the next time-step:

$$(x_{t+1} - x_{max}) = -(x_0 - x_{max})a^{t+1} \tag{5.4}$$

which can be rewritten in the form:

$$(x_{t+1} - x_{max}) = -(x_0 - x_{max})a^t a \tag{5.5}$$

Substituting $(x_t - x_{max})$ from equation 5.3 into this expression, gives:

$$(x_{t+1} - x_{max}) = (x_t - x_{max})a \tag{5.6}$$

To simplify this expression we set $x_{max} = 0$, so that the expression can be written in the autoregressive form 5.6:

$$x_{t+1} = ax_t + w_{t+1} \tag{5.7}$$

where $a < 1$, $w_t \sim N(0, s^2)$, with $s = 0$ corresponding to a strict exponential decay, $s > 0$ allowing for noise.

We can check this result by generalising equation 5.7 through equations 5.8, 5.9 and 5.10:

$$x_1 = ax_0 \tag{5.8}$$

$$x_2 = a\,(ax_0) \tag{5.9}$$

$$\vdots$$

$$x_t = x_0 a^t \tag{5.10}$$

By adding the limit value of the process in 5.10, and rearranging the equation, we find again equation 5.2:

$$x_t = (x_0 - x_{max})\, a^t + x_{max} \tag{5.11}$$

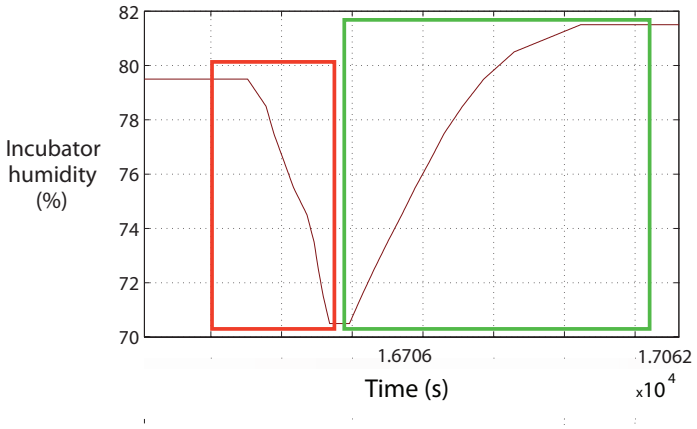
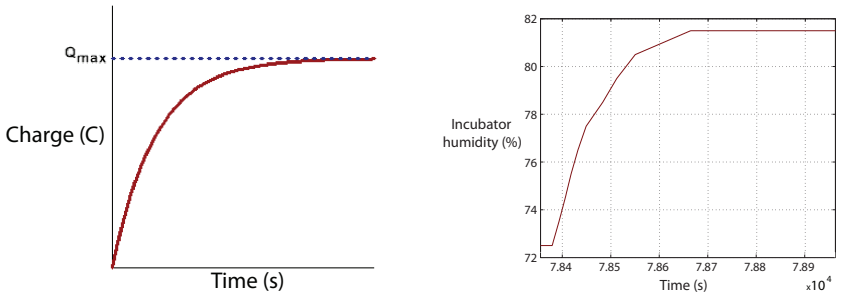


Figure 5.2: Example of incubator humidity recovering to set level after having been closed. The red square highlights the decay of the incubator humidity readings caused by the incubator opening. The green square highlights the recovery of the humidity readings after the incubator has been closed after having been opened. This curve increases exponentially, and stabilizes at set humidity level.



(a) Capacitor charging. Source: [45]. (b) Recovery of humidity readings to set level

Figure 5.3: Comparison between (a) a capacitor charging and (b) the humidity recovering after the incubator has been opened and closed. The two dynamics are similar and consist of an inverted exponential decay.

5.1.2 Parameter estimation

Once the model has been constructed, some samples of humidity recovery episodes are selected manually to train the model, in order to estimate the model parameters. Since the model constitutes an AR(1) process, the parameters can be obtained by solving the Yule Walker equations (see section 3.1.1), as described below.

Yule Walker equations

We display again the equation for the AR(1) process (equation 5.7):

$$x_{t+1} = ax_t + w_{t+1} \tag{5.12}$$

Equation 5.12 is multiplied by x_t:

$$x_t x_{t+1} = ax_t x_t + w_{t+1} x_t \tag{5.13}$$

Then, we take the expectation of equation 5.13:

$$\langle x_t x_{t+1} \rangle = a \langle x_t x_t \rangle + \langle w_{t+1} x_t \rangle \tag{5.14}$$

We use the fact that$\langle w_{t+1} x_t \rangle = 0$, since the noise at the current time is uncorrelated to the previous value of the process, and obtain equation 5.15:

$$\langle x_t x_{t+1} \rangle = a \langle x_t x_t \rangle \tag{5.15}$$

We divide equation 5.15 by $N - 1$:

$$c_1 = ac_0 \tag{5.16}$$

Finally, we divide equation 5.16 by c_0, to obtain the parameter $'a'$:

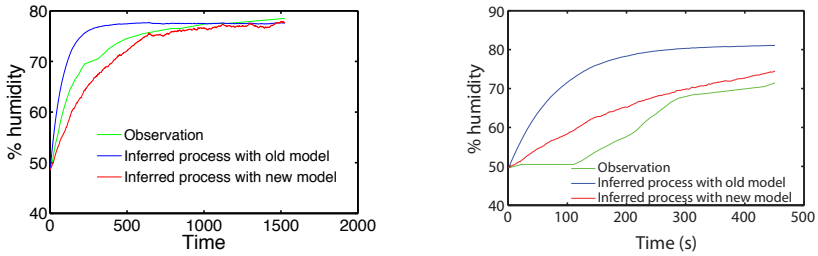
$$\frac{c_1}{c_0} = a \tag{5.17}$$

Then, we obtain the variance of the process by subtracting each value to the value predicted by the model, taking the square, and computing the mean over all the values as follow:.

$$s^2 = \frac{1}{N}\left(\sum_{t=1}^{N}(x_{t+1} - ax_t)^2\right) \tag{5.18}$$

5.1.3 Model evaluation

In order to confirm that our model was good enough to be able to explain the data, the same technique as the one seen in section 3.2.4) was used here. We sampled from our new model (the recovery model) and verified that the dynamic followed by these data generated by our model was similar to the dynamic of real sequences of humidity readings during recovery. In Figure 5.4, we display two examples with real humidity measurements while recovering to set level (green curve), humidity measurements sampled from the old model (i.e. normality model; blue curve) and humidity measurements sampled from the new model (recovery model; red curve). We can see that the dynamic of the real humidity readings (green curve) is similar to the dynamic of the data sampled from the new model (red curve). Moreover, it can be observed that the new model (red curve) fits better the data then the old model (blue curve).



(a) (b)

Figure 5.4: Comparison between real humidity measurements while recovering to set level (green curve) and some humidity measurements sampled from the old (blue curve) and new model (red curve). The new model (red curve) achieves a better qualitative fit to the data than the old model (i.e the normality model; blue curve).

5.1.4 Integration of the model in the system

In this section, I describe the modifications that have been made to the system in order to be able to integrate the new recovery model as a third stage of the IO factor.

Retraining of 'normal' stage parameters

The parameters of the model describing the normal regime of the humidity readings (the normality model) were hard-coded in the system. This was an important issue, as both the normality model and the new recovery model are $AR(1)$ process, and the parameters of the recovery model were found to be very closed to the hard-coded parameters of the normality model, leading to both models accounting for similar dynamics (see Figure 5.5).

We addressed this by retraining the normality model parameters. As for the recovery model, the normality model is an $AR(1)$ process with parameters found by solving the Yule-Walker equations. The hard coded and learned parameters are displayed in table 5.1.

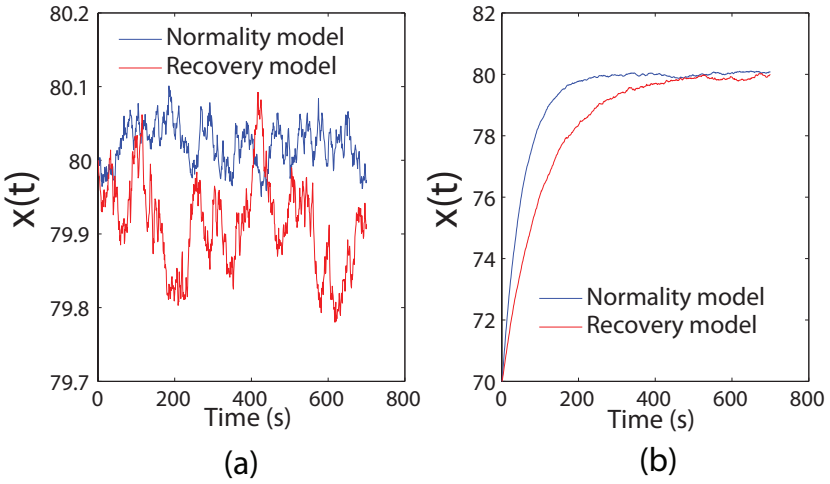


Figure 5.5: (a) Example dynamics of the normal regime of incubator humidity readings, generated by a sample drawn from the normality model (blue) and the recovery model (red) with initial value equal to set value ($x_0 = 80 = x_{max}$). (b) Example dynamics of the recovery of incubator humidity generated by a sample drawn from the normality model (blue) and the recovery model (red) with same initial value ($x_0 = 7$) and set value ($x_{max} = 80$). The plots underline how similar these the dynamics generated by both model are.

Parameters	Non-trained normality model	Trained Normality model
a	0.988	0.998
$\sigma^2 = v$	1e-4	1e-5

Table 5.1: Normality model parameters. Left: previous hard coded parameters. Right: retrained parameters.

The three-stage handling factor

Originally, the handling factor was composed of only two states:

1. Incubator open - humidity declining (described by the'incubator open' / 'decay' model; see top blue squares in Figure5.6)
2. Incubator closed - humidity maintained constant (described by the 'normality' model)

The IO factor is now extended by the addition of the recovery stage as a third stage as follow:

1. Incubator open - humidity declining (described by the 'incubator open' / 'decay' model; see top red square in Figure5.6)
2. Recovery to set level (described by 'recovery' model; see top green square in Figure5.6)
3. Incubator closed - humidity maintained constant (described by the 'normality' model; see top blue squares in Figure5.6)

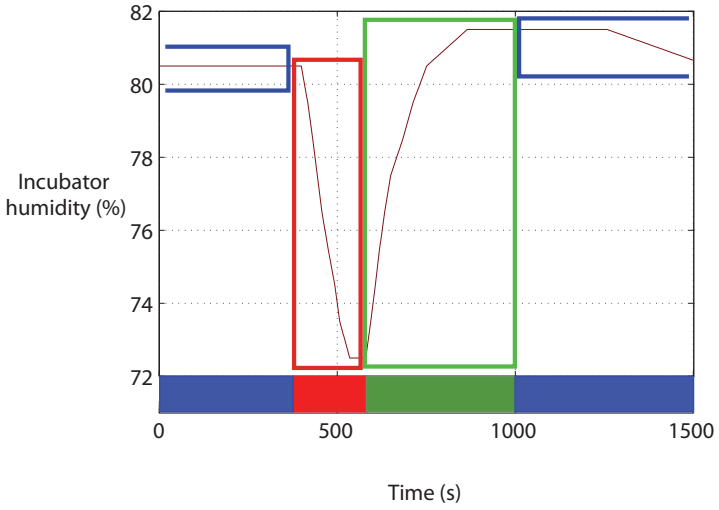


Figure 5.6: The three states of the new handling factor: 1) The incubator open state (causing the decay of the humidity readings) is highlighted by the red square. 2) The incubator closed state (causing the recovery to set level of the humidity readings) is highlighted by the green square. 3) The normal state of the handling factor (causing the humidity reading to be stable around the set value) is highlighted by the blue squares.

The transition probability matrix

Transforming the handling factor into a three-stage factor implies that the transition probability matrix must be changed accordingly. Indeed, the humidity channel was initially described by only two states, the normal state and the incubator opening state, which implied the use of a 2×2 transition probability matrix. After integration of the recovery state, the humidity channel is described by three states, which leads to a 3×3 transition probability matrix, which is computed by following the steps described below:

Method: Given the following definitions:

- $n_{i|j}$ the number of transition from state j to state i, and
- $\theta_{i|j}$ the probability of going to state i, given that we are currently in state j.

The transition probability matrix can be computed using the following formula (see section 3.2.1):

$$\hat{\theta}_{ij} = \frac{n_{i|j}}{\sum_k n_{k|j}} \tag{5.19}$$

In reality, it is only possible to change stages in the order: 'normality' (incubator closed - humidity constant at set level) $\longrightarrow$ 'incubator open' (humidity decrease) $\longrightarrow$ 'recovery' (humidity increase and stabilize at set level). Every other sequence transition (except transition from one state to the same state) is forbidden. These rules are used to compute the different values of the transition probability matrix, and are written in mathematical language as follow:

- $n_{i|i} = T_i$ where T_i corresponds to the total time passed in state i
- $n_{i+1|i} =$ where N_i corresponds to the number of transitions from state i to state $i+1$
- $n_{j|i} = 0$ for all $j \neq i+1$

These rules are presented in diagram form in Figure 5.7.

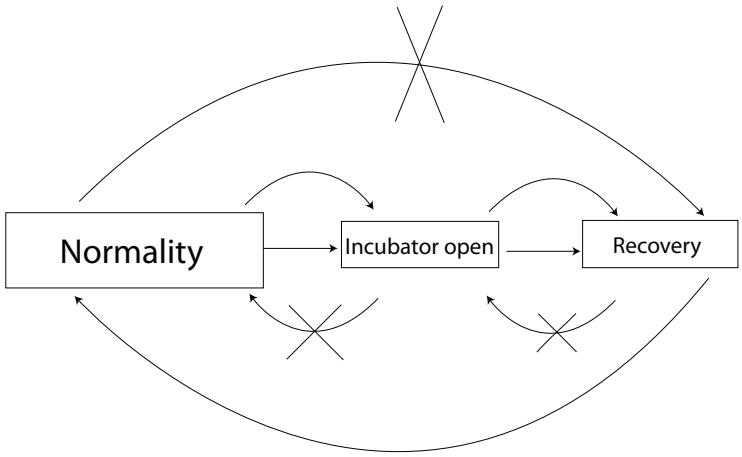


Figure 5.7: Illustration of the state transitions. The state transition reflect the possible sequences of events in reality, where the incubator is closed ('normality' state), than is opened ('incubator opening' state), and than closed ('recovery state'), leading to the recovery of the humidity readings to set level. When the set level is reached, the system return in the 'normality' state. From each state, the system can either stay in the same state, either follow the transition described above. All the other transitions are forbidden, as shown by the crossed-out arrows.

5.1.5 ROC theory

This section describes the theory behind the ROC analysis, tool that has been used in this project to assess the performance of the re-modelled IO factor in classifying handling episodes (including both decay of humidity readings due to

the incubator opening and humidity recovery to set level caused by the closing of the incubator). Receiver operating characteristics (ROC) graphs constitute a useful method to visualise classifiers performance. Table 5.2 displays the confusion matrix and the common performance matrix calculated from it. The numbers along the major diagonal represent the correct decisions made, and the numbers of this diagonal represent the errors—the confusion—between the various classes [46]. Below are display some basic definitions concerning the classifiers, necessary at the understanding of the ROC graphs.

- TP: True Positive, corresponds to positives correctly classified
- TN: True Negative, corresponds to negatives correctly classified
- N: Total number of Negative, corresponds to FP+TN.
- FP: False Positive, corresponds to positives incorrectly classified
- FN: False Negative, corresponds to negatives incorrectly classified
- P: Total number of Positive, corresponds to FN+TP.
- The true positive rate of the classifier is:

$$tp\,rate \approx \frac{TP}{(FN+TP)}$$

- The false positive rate (also called false alarm rate) of the classifier is:

$$fp\,rate \approx \frac{FN}{(FP+TN)}$$

- Specificity:

$$specificity = \frac{TN}{FP+TN} = 1 - fp\,rate$$

- Sensitivity:

$$sensitivity = \frac{TP}{FN+TP} = tp\,rate$$

		Predicted value	
		Positive	Negative
Actual value	Positive	TP	FN
	Negative	FP	TN

Table 5.2: Diagram of a confusion matrix. A confusion matrix is a visualization tool that gives information allowing to assess the performance of a binary classification system. Each column of the matrix contains the predicted outcome obtained from the classifier, and each row contains the real classifications.

In a **ROC curve**, the true positive rate (or sensitivity) is plotted in function of the false positive rate (1-specificity) for different decision threshold values. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. In fact, a ROC graph depicts relative trade offs between benefits (true positives) and costs (false positives). A perfect classifier would have a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the classifier [46].

The **AUC value** corresponds to the area under the ROC curve. In regards of the theoretical properties of ROC curves, the closer the AUC value to one, the better the performance of the classifier [46].

The **EER value** is the error rate corresponding to the threshold for which false acceptance rate and false rejection rate are equal. The value indicates that the proportion of false acceptances is equal to the proportion of false rejections. The lower the equal error rate value, the higher the accuracy of the classifier [46].

Labelling the new handling episodes for ROC analysis

In order to asses the performance of the IO factor in classifying handling episodes, that is, to compute the ROC curve, the AUC and the EER values, we had to obtained the new 'true' handling episodes - that would include, in addition to the humidity readings decreasing due to the incubator opening, the humidity readings recovering to set level due to the closing of the incubator. The labelled data for the humidity readings decreasing due to the incubator opening were already available (provided by the clinical experts). The labelling of the humidity recovering to set value were obtained automatically, by using the existing labelled data for humidity decay as follow: for each interval labelled as humidity decay, the end of the segment was matched to the beginning of the recovery segment, and the end of the recovery segment was matched to the moment where the humidity measurements were reaching their maximum value before the next 'true handling' interval, or, if this maximum value was above the set humidity value, the end of the recovery segment was matched to the moment where the humidity measurements were reaching the set value.

5.2 Results

In the following section, we describe the results obtained from the performance evaluation of the re-modelled IO factor in classifying the new handling episodes (including both decay of humidity readings due to the incubator opening and humidity recovery to set level caused by the closing of the incubator).

Figure 5.8 displays the ROC graph for the IO (handling) factor before and after re-modelling. Some points of the new model ROC curve (red curve) lie below the old model ROC curve (blue curve), indicating that, at least for small

positive rates, the old model performs slightly better. However, most of the ROC curve is greater for the new model than for the old model. This is reflected in the AUC value, a summary statistic relating to the overall classification performance (see section 5.1.5), which is much higher for the new model than for the old model (respectively 0.905 and 0.850 for the new and for the old models), as well as for the EER (respectively 0.184 and 0.209 for the new and for the old models).

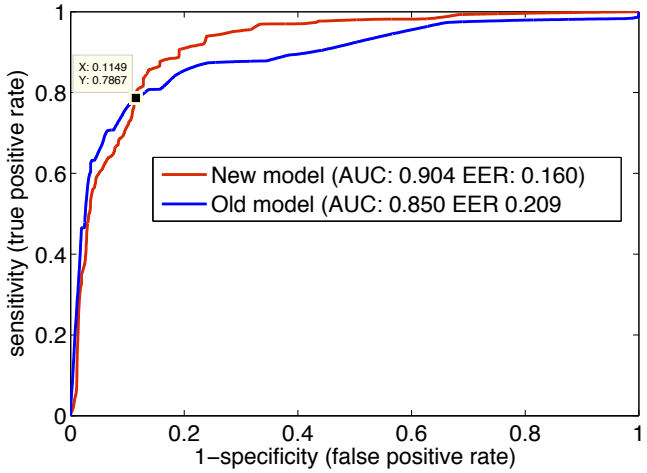
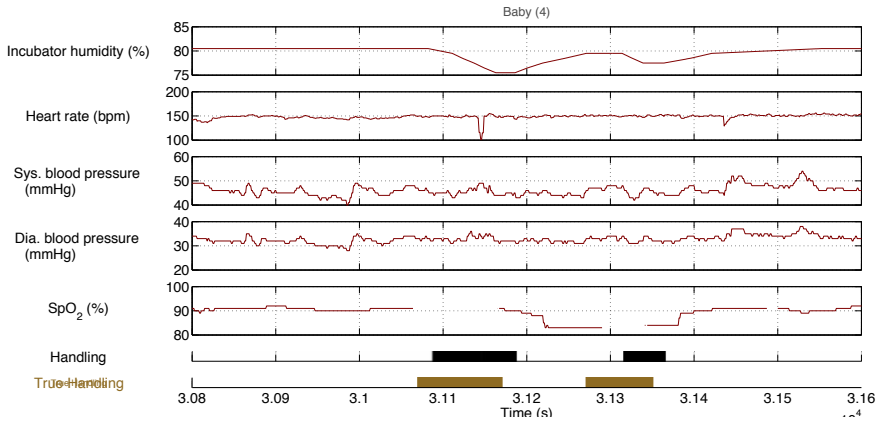
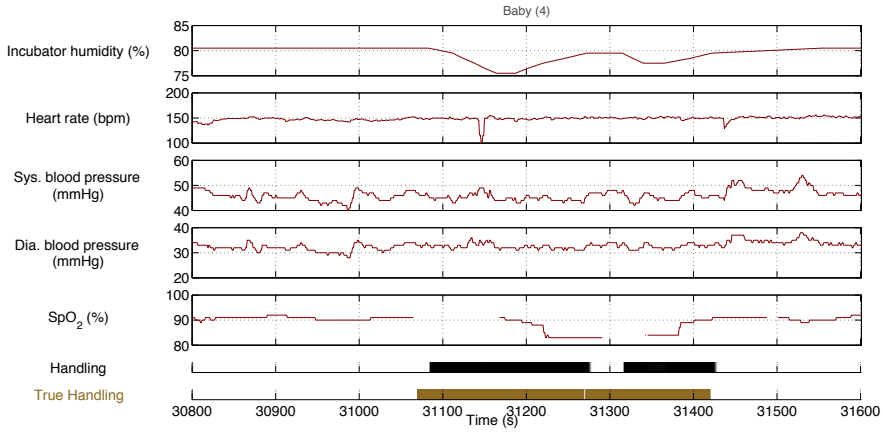


Figure 5.8: Comparison between the ROC curves for (a) the old and (b) the new handling factor. Both ROC curves have been computed by comparing the inferred handling intervals with the new labelled data.

Inspection of inferred distributions of the old and the re-modelled IO factor during handling episodes were undertaken, in order to visually compare the performance of the two factors. Most of these visual inspections confirmed the ability of the new IO factor to account for the humidity recovery to set level. One representative example is displayed in Figure 5.9 where we can see the inferred probability distributions (in grey levels) of the old IO factor (Figure 5.9a) and the new IO factor 5.9b during two subsequent handling episodes. It can be seen that when using the old IO factor (Figure 5.9a), only the decay of the humidity was picked up by the IO (handling) factor, whereas after the re-modelling of the IO (handling) factor, the inference is almost perfect, with the recovery segment now being completely picked up by the recovery model.



(a) Inferred switch settings of the old IO (handling) factor

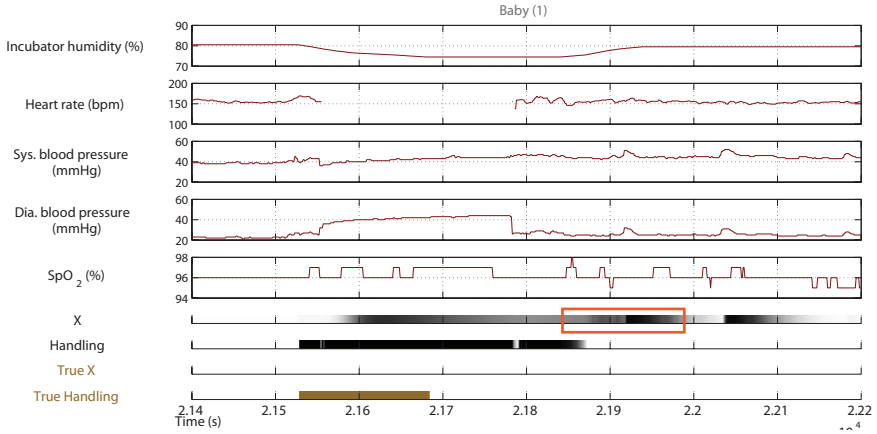


(b) Inferred switch settings of the new IO (handling) factor

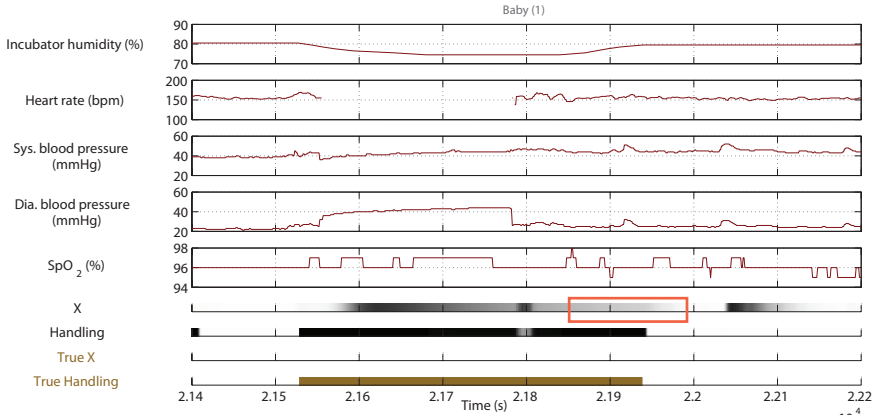
Figure 5.9: Representative example of inferred switch settings for the Handling (Incubator open) factor before the addition of the humidity recovery model (top Figure), and after this addition (bottom Figure). In (a) (old model of the IO factor), we see that the incubator factor was only accounting for the decay of the humidity level. In (b) (new model of the IO factor), we see that the IO factor accounts now for the decay of humidity level, as before, and for the totality of the recovery of the humidity to set level that follows.

Finally, we inspected the inferred probability distributions of the X-factor during humidity recovering episodes, in order to see if the fact that the humidity recovery was now picked up by the new IO factor was stopping the X-factor from firing during such recovery phases. The result of these inspection suggested that the likelihood of the X-factor firing during recovery phases were reduced with the new IO factor, in comparison to the case were the old IO factor was used. In

Figure 5.10, we display a representative example of handling episode with the corresponding probability distributions of the X-factor and the **old** IO factor (in Figure 5.10a) and the corresponding probability distributions of the X-factor and the **new** IO factor (in Figure 5.10a). In this particular example, we see that the addition of the humidity recovery model significantly reduced the probability of the X-factor firing during recovery segments (see red squares).



(a) Old IO (handling) factor



(b) New IO (handling) factor

Figure 5.10: Comparison of inferred switch settings for handling (incubator open) factor and X-factor with the (a) old and (b) new IO (handling) factor model. The re-modelling of the IO (handling) factor reduced the probability of the X-factor firing during the recover of humidity to set level (see red square highlighting the X-factor inferred distribution in (a) and (b) for comparison).

5.3 Discussion

Early feedback from the clinical experts indicated that the system was unable to deal with the recovery of incubator humidity readings to set level, following the opening and closing of the incubator door. This was signalled by the X-factor triggering during these events. As the recovery of humidity reading to set level does not reflect a change that is related to the true state of the baby, the X-factor firing for such type of data was unwanted. To address this, the handling (incubator open) factor was re-modelled, so that it would be able to account for the recovery of the humidity readings to set level. The new model was able to better account for sections of the data where the humidity level was undergoing recovery (quantified via the ROC curves in section 5.2). Moreover, qualitative observations from the data appeared to suggest that the new model helped reduce the likelihood that the X-factor is firing during these recovery periods (see section 5.2).

However, despite the overall good performance of the new incubator open factor, some undesired behaviour was observed (see Figure 5.11). Specifically, we sometimes observed that after the humidity decreases that were due to the opening of the incubator, the humidity stabilised for a while (see top blue square in Figure 5.11) before recovering to set level. When this stabilisation occurred for values of humidity that were close to the set level, it was sometimes not picked up by the new incubator open factor (see bottom blue square in Figure 5.11). This is an important issue, as it could lead to the X-factor firing during these particular intervals.

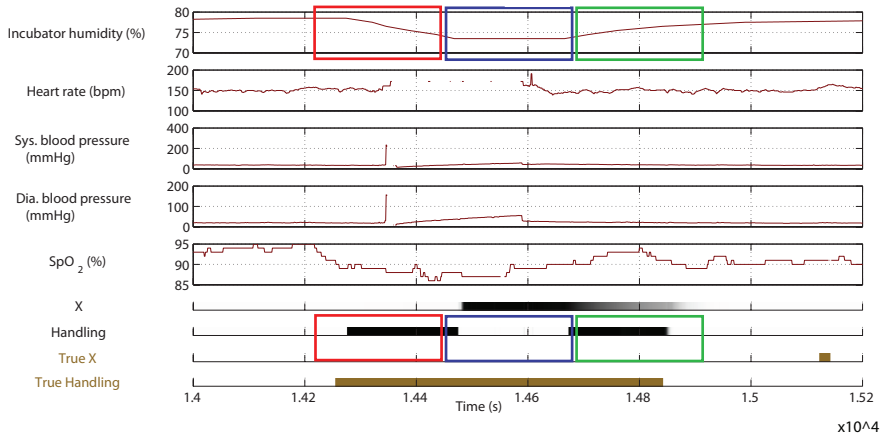


Figure 5.11: Inferred switch settings for handling (Incubator open) factor during the opening of the incubator and its subsequent closing. While the decrease of the humidity level (top red square) and the recovery of the humidity level to set value (top green square) are being correctly picked up by the new IO (handling) factor (bottom red and green square respectively), the intermediate segment (top blue square) is not picked up by the handling factor (bottom blue square).

Further analysis showed that the segment of data during humidity stabilisation (blue square in Figure 5.11) was more likely under the normality model than under the model describing the opening of the incubator (incubator open model). This explains why this segment was wrongly picked up by the normality model, rather than the incubator open model. From looking at the data, an easy way to discriminate between normal humidity recordings and the stabilisation phase after the opening of the incubator (top blue square in Figure 5.11) is that normally, the humidity is close to a specific 'set value', whereas during the stabilisation phase it exhibits values further away from this set value. Based on this intuition, we tried to address this issue by multiplying the likelihood of the normality model by a penalty term; the size of the penalty increasing with the distance of the humidity level from the set value. To do this, we used a 'product of expert' model: a model that combines individual component models (the experts) by taking their product and normalizing the result [47]. In this case, we used two individual component models. The first component model is the normality model describing the normal incubator humidity readings, with log-likelihood given by:

$$\log P(\tilde{X}) = -\frac{1}{2\sigma^2}\left[x_1^2 + \sum_{t=2}^{N}(x - ax_{t-1})^2\right] + C \qquad (5.20)$$

Note that in this equation, the set value of the humidity level was set to zero. The second component model corresponds to the penalty model, and was set to the following log likelihood, so that the further away the data is from the mean (or set humidity level), the smaller the likelihood (with the mean of the process set to zero):

$$\log P_2(\tilde{X}) = -\sum \frac{x^2}{2\sigma_0^2} + C \tag{5.21}$$

with the total likelihood of the resulting normality model resulting from the addition of the two log-likelihoods (minus a normalization term):

$$\log P(\tilde{X}) \propto \log P_1(\tilde{X}) + \log P_2(\tilde{X}) - \log Z \tag{5.22}$$

In order for the resulting model of normality to remain an autoregressive process, its inverse covariance matrix (equal to the addition of the two inverse covariance matrix) must satisfy the following property:

$$\Sigma^{-1} = \frac{1}{2\sigma^2}\begin{pmatrix} 1 & -a & 0 & 0 \\ -a & 1+a^2 & \ddots & 0 \\ 0 & \ddots & \ddots & -a \\ 0 & 0 & -a & 1+a^2 \end{pmatrix} + \frac{1}{2\sigma_0^2} I = \frac{1}{2\tilde{\sigma}^2}\begin{pmatrix} 1 & -\tilde{a} & 0 & 0 \\ -\tilde{a} & 1+\tilde{a}^2 & \ddots & 0 \\ 0 & \ddots & \ddots & -\tilde{a} \\ 0 & 0 & -\tilde{a} & 1+\tilde{a}^2 \end{pmatrix} \tag{5.23}$$

with $\tilde{a}$ and $\tilde{\sigma}$ corresponding to the parameters of the new normality model. Solving equation 5.23 is equivalent to finding the solutions to the following equations

$$\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2} = \frac{1}{2\tilde{\sigma}^2} \tag{5.24}$$

$$\frac{a}{2\sigma} = \frac{\tilde{a}}{\tilde{\sigma}^2} \tag{5.25}$$

$$\frac{(1+a^2)}{\sigma^2} + \frac{1}{2\sigma_0^2} = \frac{(1+\tilde{a}^2)}{2\tilde{\sigma}^2} \tag{5.26}$$

To avoid the only solution being $\tilde{a} = a$ and $\tilde{\sigma} = \sigma$, we neglect equation 5.24, (which only accounts for one term of the covariance matrix).The solution to these equations can be written as:

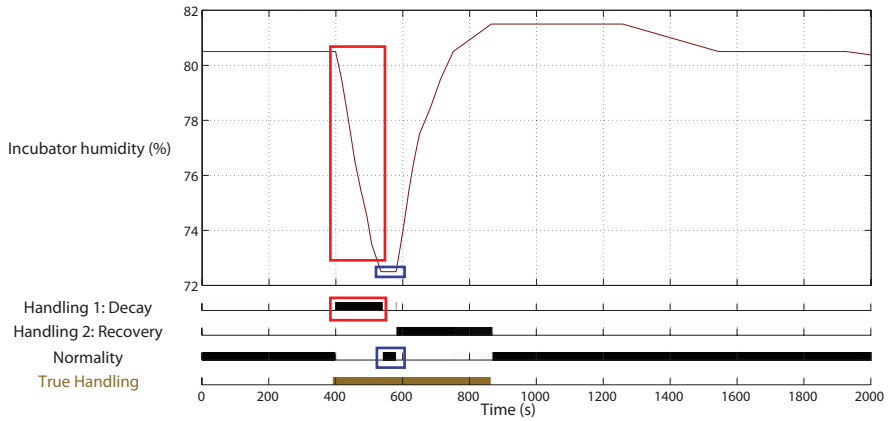
$$\tilde{\lambda} - \left(\left(1+a^2\right)\lambda + \lambda_0\right)\tilde{\lambda} + a^2\lambda^2 = 0 \tag{5.27}$$

$$\tilde{\lambda} = \frac{-\lambda(1+a^2) + \lambda_0 + \sqrt{\Delta}}{2} \tag{5.28}$$

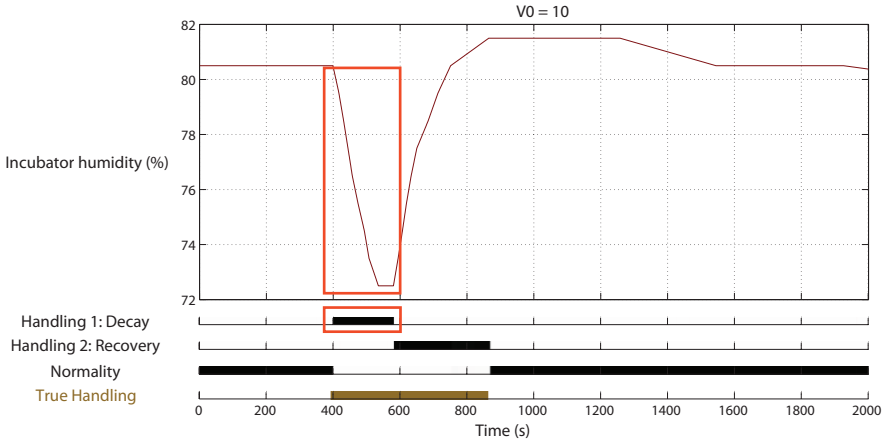
(where $\lambda = \frac{1}{\sigma^2}$, $\lambda_0 = \frac{1}{\sigma_0^2}$, $\tilde{\lambda} = \frac{1}{\tilde{\sigma}^2}$ and $\Delta = \lambda^2(1-a^2)^2 + 2\lambda(1+a^2)\lambda_0 + \lambda_0^2 > 0$). Finally, it is straightforward to obtain the parameters $\tilde{a}$ and $\tilde{\sigma}^2$ using equation 5.28.

After setting up the product of expert model, we looked at how it was able to infer the incubator open factor events from the physiological data: in particular, the stabilisation segments. Therefore, we looked at the probability distribution that the IO factor was firing, during selected data segments, with varying the penalty term in the model (σ_0). Although isolated examples showed that this model could account well for the data segments which had previously been a problem (the stabilization segments; see figure 5.12), the addition of a free parameter (σ_0) created additional problems. That is, it was not clear from our initial work how to best set σ_0, or indeed, learn it from the data. Hopefully however, future work should be able to address this issue, and successfully utilize the advantages provided by the presented product of experts model.

In the previous chapter, we set out a system which could gather feedbacks, on the basis that their can provide expert knowledge which is crucial to improving the model. This idea is reinforced in this chapter: we were motivated to build a better model of the incubator opening based on initial feedback gathered informally from the clinicians. This shows the value in developing close collaborative ties between the machine learning lab and the clinicians; as their feedback is crucial to the further development of the system. Hopefully, in the future, gathering such feedback automatically should guide the development of further extensions to the system.



(a) Illustration of the issue: the stabilization segment (see top blue square) is being wrongly picked up by the normality model (see bottom blue square).



(b) Issue addressed using a PoE model ($\sigma_0 = 10$)

Figure 5.12: Example of inferred switch settings for each stage of the handling factor (incubator open (decay) model, the recovery model and the normality model), during an handling episode. (a) This example illustrates the issue discussed, that is, the stabilization segment of humidity just before the closing of the incubator (see top blue square) is being wrongly picked up by the normality model (see bottom blue square). (b) In this example, we apply the PoE model described in the text, with σ_0^2 set to the value of 10. The incubator open (decay) model picks up now the stabilisation of the humidity readings.

Chapter 6

Conclusion

In this thesis, I described the system constructed by my laboratory, designed to infer the real state of health of premature babies, based on the vital parameters (e.g. heart rate, oxygen saturation) recorded from them in neonatology units. The goal of this system was to reduce the number of false alarms, which are very high in the devices currently used in neonatology units, by allowing the system to discriminate between changes in the recorded vital parameters that were due to real physiological problems, and changes caused by artefacts (e.g. probe drop-out).

In my project, I contributed to this project by setting up a method to evaluate the system using new monitoring data. The aim here was to asses its current performance, and find where it could be improved. As evaluating this system requires knowing the true clinical interpretation of the physiological data, clinical expertise had to be used, requiring close collaborations with clinical experts.

An interesting and important feature of the model developed by my host laboratory is that, as well as attributing each segment of data as due to some known physiological or artefactual causes (i.e. bradycardia, probe drop out), it is also able to identify physiological intervals that the model can not classify well - these intervals are labelled by the 'X-factor'. This is particularly useful for evaluating the system performance, allowing us to gain feedback only on those intervals where the model need to be improved.

The evaluation system constructed in this project takes the form of a feedback webform which gathers clinical interpretations of 'X-factor' intervals. In setting up this application, we had to face diverse issues such as making the form user friendly and clear to the specialised language used by clinicians. The relatively small number of feedbacks obtained, as well as the biases introduced by evolving the webform through time means that we have to be careful not to draw too strong conclusions. Nevertheless, some results stand out.

First, we confirmed that the model was behaving as intended: artefactual factors known by the system were able to accurately predict the events they were intended to model. Moreover, the analysis of feedbacks from the clinical experts suggested that some additional factors could be usefully added to the system,

particularly to model the heart rate and the oxygen saturation channels. The fact that the initial feedbacks clearly showed where the model needed improving are encouraging, and suggest that this approach could help discover new regimes claimed by the "X-factor" which, as a result, could then be added as factors in the model.

In the near future, further work should include the adaptation of the webform to correct for some of the issues discussed in this report. Notably, the categories extracted from the free text comments that were flagged repeatedly should be added in the webform as labelled events with a tick box, whereas the labelled events with a tick box which have hardly been flagged should be removed from the list of labelled events in the webform. At the same time, some clear guidelines should be written for the clinicians answering the feedback forms, explaining clearly what each question means, with illustrative examples of physiological data corresponding to each kind of events listed in the webform.

The second part of my project was motivated by early feedbacks form clinical experts, suggesting that the model was not able to account for the humidity readings recovering after the opening and closing of the incubator, causing the X-factor to fire often during these events. As the goal of the FSLDS model was to reduce the number of false positive, it is important that this kind of data is accounted for by an artefactual factor rather than the X-factor. To address this, the re-modelling of the handling factor was undertaken, so that this factor would be able to account for the recovery of the humidity readings to the set level, after the incubator is closed. This re-modelling gave reasonably good results, and seemed to decrease the likelihood of the X-factor episodes for segments of data just after the incubator was closed. In the discussion, we described how this model could be extended in the future to account for occasional occurrence of data segments where the humidity stabilised at steady values far away from its normal value.

Bibliography

[1] C. E. Poets, "Monitoring in the NICU," in *Respiratory Controls and Disorders In The Newborn*, ch. 9.

[2] J. A. Quinn and C. K. I. Williams, "Known Unknowns: Novelty Detection in Condition Monitoring," pp. 1–6, 2007.

[3] B. Larroque, P.-Y. Ancel, S. Marret, L. Marchand, M. André, C. Arnaud, V. Pierrat, J.-C. Rozé, J. Messer, G. Thiriez, A. Burguet, J.-C. Picaud, G. Bréart, and M. Kaminski, "Neurodevelopmental disabilities and special care of 5-year-old children born before 33 weeks of gestation (the EPIPAGE study): a longitudinal cohort study.," *Lancet*, vol. 371, pp. 813–20, Mar. 2008.

[4] R. L. Goldenberg, J. F. Culhane, J. D. Iams, and R. Romero, "Epidemiology and causes of preterm birth.," *Lancet*, vol. 371, pp. 75–84, Jan. 2008.

[5] AboutKidsHealth, "Premature Babies Resource Centre," 2004.

[6] J. G. Long, J. F. Lucey, and A. G. S. Philip, "Noise and Hypoxemia in the Intensive Care Nursery," *Pediatrics*, vol. 65, no. 1, pp. 143–145, 1980.

[7] V. Ahlborn, B. Bohnhorst, C. S. Peter, and C. F. Poets, "False alarms in very low birthweight infants: comparison between three intensive care monitoring systems.," *Acta paediatrica (Oslo, Norway : 1992)*, vol. 89, pp. 571–6, May 2000.

[8] C. Williams, "Condition monitoring in premature babies," 2006.

[9] D. Barber, *Bayesian Reasoning and Machine Learning*. 2011.

[10] J. Quinn, C. K. Williams, and N. McIntosh, "Factorial switching linear dynamical systems applied to physiological condition monitoring.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1537–1551, 2009.

[11] A. Guyton, "Fetal and Neonatal Physiology," in *Textbook of medical physiology*, ch. 83, Elsevier Saunders, 2006.

[12] D. Wright, "Birth," in *Human physiology and health*, ch. 9.8, Heinemann, 2000.

[13] S. T. Lawless, "Crying wolf: false alarms in a pediatric intensive care unit.," *Critical care medicine*, vol. 22, pp. 981–5, June 1994.

[14] G. G. HADDAD*, "Respiratory Control in the Newborn," in *Respiratory Controls and Disorders In The Newborn*, ch. 1, 2005.

[15] A. Guyton, "Pulmonary ventilation," in *Textbook of medical physiology*, ch. 37, Elsevier Saunders, 2006.

[16] A. Guyton, "Regulation of Respiration," in *Textbook of medical physiology*, ch. 41, Elsevier Saunders, 2006.

[17] P. Nightingale, "Paediatric and neonatal anaesthesia," in *Oxford Handbook of Anaesthesia. 1st Edn*, vol. 89, ch. 33, pp. 757–762, Aug. 2002.

[18] H. Rigatto, "Periodic Breathing," in *Respiratory Controls and Disorders In The Newborn*, ch. 10, p. 266, CRC Press, 2003.

[19] C. E. Poets, "Pathophysiology of Apnea of Prematurity," in *Respiratory Controls and Disorders In The Newborn*, ch. 12.

[20] A. Guyton, "Heart Muscle; The Heart as a Pump and Function of the Heart Valves," in *Textbook of medical physiology*, ch. 9, Elsevier Saunders, 2006.

[21] P. M. Oomen, "Apnea, Bradycardia and Desaturation: Clinical Issues," in *Respiratory Controls and Disorders In The Newborn*, ch. 11, 2005.

[22] Task Force on Prolonged Apnea, "Prolonged Apnea," *Pediatrics*, vol. 61, no. 4, pp. 651–652, 1978.

[23] C. F. Poets, V. A. Stebbens, M. P. Samuels, and D. P. Southall, "The relationship between bradycardia, apnea, and hypoxemia in preterm infants.," *Pediatric research*, vol. 34, pp. 144–7, Aug. 1993.

[24] C. F. Poets, G. A. Rau, K. Neuber, M. Gappa, and J. Seidenberg, "Determinants of lung volume in spontaneously breathing preterm infants.," *American journal of respiratory and critical care medicine*, vol. 155, pp. 649–53, Feb. 1997.

[25] G. P. Heldt, "Development of stability of the respiratory system in preterm infants," *J Appl Physiol*, vol. 65, no. 1, pp. 441–444, 1988.

[26] M. A. Dimaguila, J. M. Di Fiore, R. J. Martin, and M. J. Miller, "Characteristics of hypoxemic episodes in very low birth weight infants on ventilatory support.," *The Journal of pediatrics*, vol. 130, pp. 577–83, Apr. 1997.

[27] J. G. Long, A. G. Philip, and J. F. Lucey, "Excessive handling as a cause of hypoxemia.," *Pediatrics*, vol. 65, pp. 203–7, Feb. 1980.

[28] R. J. Martin and A. A. Fanaroff, "Neonatal apnea, bradycardia, or desaturation: does it matter?," *The Journal of pediatrics*, vol. 132, pp. 758–9, May 1998.

[29] A. Guyton, "The Normal Electrocardiogram," in *Textbook of medical physiology*, ch. 11, Elsevier Saunders, 2006.

[30] C. F. Poets and V. A. Stebbens, "Detection of movement artifact in recorded pulse oximeter saturation," *European Journal of Pediatrics*, vol. 156, pp. 808–811, Sept. 1997.

[31] J. van der Sloten, P. Verdonck, and M. Nyssen, "Evaluation of Conventional and Non-Conventional Pulse Oximeter," in *4th European Conference of the International Federation For Medical and Biological Engineering*, Springer, 2008.

[32] A. Guyton, "Vascular Distensibility and Functions of the Arterial and Venous Systems," in *Textbook of medical physiology*, ch. 15, 2006.

[33] C. M. Bishop, *Patter recognition and machine learning*. Springer, 2006.

[34] C. K. I. Williams, "Probabilistic Modelling and Reasoning Time Series Modelling : AR , MA , ARMA and All That." 2010.

[35] C. Chatfield, *The Analysis of Time Series: An Introduction*. Chapman & Hall, 1989.

[36] C. K. I. Williams, "Time Series Modelling and Kalman Filters." 2010.

[37] V. Pavlovi, J. Rehg, and J. MacCormick, "Learning Switching Linear Models of Human Motion," *Advances in Neural Information Processing*, 2000.

[38] A. F. Smith and M. West, "Monitoring renal transplants: an application of the multiprocess Kalman filter.," *Biometrics*, vol. 39, pp. 867–78, Dec. 1983.

[39] J. Ma, "A mixed-level switching dynamic system for continuous speech recognition," *Computer Speech & Language*, vol. 18, pp. 49–65, Jan. 2004.

[40] A. Cemgil, H. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 679–694, Mar. 2006.

[41] J. A. Quinn and C. K. I. Williams, "Bayesian Time Series Models," in *Computing* (C. S. Barber D., Cemgil T., ed.), ch. 12, pp. 1–26, Cambridge University Press, 2011.

[42] C. K. I. Williams and I. Stanculescu, "Automating the Calibration of a Neonatal Condition Monitoring System," *System*.

[43] C. Williams and N. Mcintosh, "Factorial Switching Linear Dynamical Systems for Physiological Condition Monitoring Premature Baby Monitoring." 2008.

[44] G. Lorenzi-Filho, H. R. Dajani, R. S. Leung, J. S. Floras, and T. D. Bradley, "Entrainment of blood pressure and heart rate oscillations by periodic breathing.," *American journal of respiratory and critical care medicine*, vol. 159, pp. 1147–54, Apr. 1999.

[45] C. H. Colwell, "PhysicsLAB."

[46] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, June 2006.

[47] G. E. Hinton, "Training products of experts by minimizing contrastive divergence.," *Neural computation*, vol. 14, pp. 1771–800, Aug. 2002.