# Machine Learning for Prediction of Extubation Readiness in Extremely Preterm Newborns

Charles Chijioke Onu

Computer Science
McGill University, Montreal

July 26, 2018

## Dedication

To my mother, who is no longer of this world, and to my father - I am grateful for the decades-worth of sacrifices you both made to get me here.

**Acknowledgements**

I thank Prof. Doina Precup, my supervisor whose solid advice, guidance and immense support played a critical role in the accomplishment of this work. A big thanks as well to the APEX Team - Prof. Robert Kearney, Prof Guilherme Sant'Anna, Dr Karen Brown, Dr. Carlos Robles-Rubio, Wissam Shalish, Lara Kanbar, Smita Rao, and Samantha Latremouille - for the huge effort put into data acquisition and signal processing, and for the enlightening conversations which steered this work. Last but not least, I thank the patients and parents involved in the APEX study for volunteering to help improve medicine for the future.

## Abstract

Infants who are born premature usually require breathing support through invasive mechanical ventilators in the early weeks of life. The duration of ventilation must be minimised in order to prevent the development of chronic lung disease. However, weaning the baby (extubating) too early is also dangerous and could lead to death. Presently, there are no objective protocols for physicians to make this decision, as it is an extremely difficult problem. This has led to high rates of extubation failures, up to 70% in some cases. This study leveraged methods of *machine learning* to explore predictors of extubation readiness for premature infants. We first sought to understand the kind of breathing transitions the infants undergo during a time of minimal support from the ventilators. To explore this question, we employed a framework called Markov chain modelling, which revealed interesting similarities and differences between newborns who succeeded extubation and those who did not. We converted this information into actionable knowledge through predictors built upon the Markov chains and by using powerful discriminative models known as support vector machines. Results showed that we could identify with high accuracy infants who succeeded but not those who failed extubation. Secondly, using richer data about heart and respiratory variability, we built a slightly more sophisticated model known as hidden Markov model and benchmarked it against another discriminator - Random Forests. By making critical design changes to default structure of these models and combining clinical variables (like infant age and weight), we developed a predictor that could have identified 71% of extubation failures ahead of time. The work presented in this thesis advances our understanding of respiratory patterns of premature newborns, in general, and in particular of those who may fail extubation; develops rigorous machine learning methods for physiological time-series data; and effectively brings us closer to developing accurate predictors of extubation readiness.

## Résumé

Les enfants qui sont nés prématuré souvent ont besoin de soutenir avec ventilateur mécanique invasif dans les premières semaines de la vie. La durée de ventilation doit être minimiser enfin de prévenir le développement de maladie pulmonaire chronique. Cependant, sevrer le bébé (extubation) trop tôt est également dangereux et pourrait entraîner la mort. Actuellement, les médecins n'ont pas les protocoles objectifs à faire cette décision si difficile. Celui a conduit à haut taux de échecs d'extubation, d'ici 70% dans quelques cas. Cette étude a profité des methodes de apprentissage automatique pour explorer prédicteurs de l'état de préparation à exubation chez les enfants prématuré. D'abord, nous tentons de comprendre le manière de transition respiratoire que les enfants prennent durant un temps de soutien minimale de les ventilateurs. À explorer cette question, nous avons employé un cadre entitré "Markov chain modelling" qui a révélé similitudes et différences intéressantes entre les nouveau-nés qui a reussi extubation et les autres qui ne l'a pas fait. Nous transformons cette information à connaisances exploitables par prédicteurs bâti sûr le "Markov chains" et en utilissant les modeles discriminatifs puissants connu comme "support vector machines". Les résultats ont demonstré que nous pourrions identifer avec haut , les enfants qui ont reussi mais pas celles qui ont échoué extubation. Deuxièmement, en utilisant des données plus riches sur la variabilité cardiaque et respiratoire, nous avons construit un modèle légèrement plus sophistiqué connu sous le nom de "hidden Markov model" et l'avons comparé à un autre discriminateur - "Random Forests". En modifiant de manière critique la structure par défaut de ces modèles et en combinant des variables cliniques (comme l'âge et le poids du nourrisson), nous avons développé un prédicteur qui aurait pu identifier 71% des échecs d'extubation à l'avance. Le travail présenté dans cette thèse avance notre compréhension des modèles respiratoires des nouveau-nés prématurés, en général, et en particulier de ceux qui peuvent échouer à l'extubation; développe des méthodes d'apprentissage machine rigoureuses pour les données de séries

chronologiques physiologiques; et nous rapproche effectivement de l'élaboration de prédicteurs précis de la préparation à l'extubation.

# Contents

**Contents** **v**

**1 Introduction** **1**

1.1 Scope of Clinical Problem . . . . . 1

1.2 The APEX Study . . . . . 2

1.3 Contributions and Outline of the Thesis . . . . . 3

**2 Background** **5**

2.1 Prediction of Extubation Readiness . . . . . 6

2.1.1 Spontaneous Breathing Trials and Analysis of Clinical Variables 6

2.1.2 Analysis of Cardiorespiratory Variability . . . . . 7

2.2 Machine Learning for Classification of Physiological Signals . . . . . 9

2.2.1 Sequence Data . . . . . 9

2.2.2 Scalar Covariates . . . . . 11

2.3 Developing Machine Learning Predictor of Extubation Readiness . . . 12

**3 Data Acquisition and Signal Processing** **14**

3.1 Data Acquisition . . . . . 14

3.1.1 Acquisition of cardiorespiratory data . . . . . 15

3.1.2 Acquisition of clinical data . . . . . 16

3.1.3 Extubation failure . . . . . 16
3.1.4 Database size . . . . . 17
3.2 Signal Processing . . . . . 17
3.2.1 Cardiorespiratory Metrics . . . . . 18
3.2.2 Respiratory Patterns . . . . . 18
3.2.3 Statistical and Spectral Features . . . . . 19

**4 Predicting from Patterns of Breathing 21**
4.1 Methods . . . . . 22
4.1.1 Discrete-time Markov Chain . . . . . 22
4.1.2 Discrete-time Semi-Markov Chain . . . . . 24
4.1.3 Support Vector Machine . . . . . 25
4.2 Experimental Design . . . . . 26
4.2.1 Feature Extraction for Discriminative Classification . . . . . 26
4.2.2 Experimental Protocol . . . . . 27
4.3 Markov-based Modelling and Prediction . . . . . 28
4.3.1 Modelling Transitions . . . . . 28
4.3.2 Modeling of Dwell/Sojourn Time Distributions . . . . . 29
4.3.3 Prediction . . . . . 30
4.4 Support Vector Machine Prediction . . . . . 33
4.4.1 Model Selection and Parameter Search . . . . . 33
4.4.2 Prediction . . . . . 34
4.5 Discussion . . . . . 35

**5 Predicting from Metrics of Cardiorespiratory Variability 37**
5.1 Methods . . . . . 37
5.1.1 Gaussian Hidden Markov Model . . . . . 37
5.1.2 Balanced Random Forest . . . . . 41
5.1.3 Incorporating Clinical Decision . . . . . 42

5.2 Experimental Design . . . . . 43
5.2.1 Features . . . . . 43
5.2.2 Experimental Protocol . . . . . 44
5.3 Hidden Markov Modelling and Prediction . . . . . 47
5.4 Random Forests Estimation and Prediction . . . . . 50
5.5 Discussion . . . . . 51

**6 Conclusion** **54**

**7 Appendix** **58**
7.1 Metrics of Predictor Performance . . . . . 58
7.1.1 Sensitivity . . . . . 58
7.1.2 Specificity . . . . . 58
7.1.3 Positive predictive value (PPV) . . . . . 58
7.1.4 Negative predictive value (NPV) . . . . . 59
7.1.5 Balanced Classification Accuracy and Misclassification Loss . . 59
7.2 Probability Distributions for Dwell Time . . . . . 59
7.3 Symmetric KL Divergence . . . . . 60
7.4 Ranges of Hyper-Parameters for Random Forests . . . . . 61
7.5 Learning curves for GHMM and CD-GHMM . . . . . 62
7.6 Distribution of Likelihood Scores for GHMM and CD-GHMM . . . . 64

**Bibliography** **67**

# 1

# Introduction

## 1.1 Scope of Clinical Problem

At birth, extremely preterm infants (gestational age $\leq$ 28 weeks) are usually at high risk of respiratory failure due to lung immaturity [63]. Most require endotracheal intubation and invasive mechanical ventilation (IMV) within the first days of life to survive [63, 62]. IMV is breathing support mechanism which involves the insertion of a tube (*intubation*) into the infant's trachea while oxygen is provided at intervals through a mechanical ventilator.

Physicians must minimize the duration of IMV because, even though it is a life-saving procedure, it is an independent risk factor for short- and long-term morbidities such as broncho-pulmonary dysplasia (BPD) - a chronic lung disease [63, 42]. On the other hand, one must take care to not *extubate* the infant too soon as this may ultimately result to the need for reintubation. Reintubation in such small infants is technically challenging due to inflammation of the trachea, among other factors. As such it could cause traumatic injury, infection to the upper airway, among other hazards which could result to irreversible, long-term disability or death in this population [2, 10, 23].

Unfortunately, there exists no standardized tests for determining extubation readiness in preterm infants. The decision to extubate is a very subjective one, usually

based on physician judgement and observation of bedside parameters such as blood gases, oxygen needs, etc. There is significant practice variations across institutions leading to high and inconsistent rates extubation failures from 10% to 70% (depending on the time frame and criteria used to define failure) [23, 21].

Given these observations, it is critical to develop objective tools for determining the optimal timing of extubation, in order to reduce extubation failure rates while minimizing the duration of IMV. In this work, we apply machine learning to address the task of predicting whether a patient is ready or not for extubation. We develop predictive models that combine clinical and cardiorespiratory time-series data to determine the readiness for extubation in extremely preterm newborns.

## 1.2 The APEX Study

The APEX study [59] is an ongoing multicenter, prospective, observational study aimed at developing tools for Automated Prediction of EXtubation readiness (APEX) in extremely preterm infants (clinicaltrials.gov identifier: NCT01909947). The project envisions to "develop an automated predictor to help physicians determine when extremely preterm infants are ready for extubation, using the combination of clinical tools along with novel and automated measures of cardiorespiratory variability". APEX has 3 main objectives:

1. Generate a library of clinical data and cardiorespiratory signals in preterm infants prior to extubation;

2. Develop a robust model for prediction of extubation readiness, i.e. referred to as APEX (Automated prediction of extubation readiness);

3. Prospectively validate the clinical utility of this prediction model

Work on objective 1 began in September 2013, with data being acquired from 5 tertiary-level NICUs in North America. The work presented in this thesis documents progress on objective 2 of the APEX protocol [59].

## 1.3 Contributions and Outline of the Thesis

This work provides several analyses using machine learning to predict extubation readiness in extremely preterm infants. Previous studies [49, 22] had focused on only one input modality (clinical variables or cardiorespiratory metrics), was based on smaller patient populations and employed a single type of predictor. This work developed generative and discriminative machine learning models for dealing with time-series measures of breathing patterns and cardiorespiratory behaviour. In addition, we make the first known attempt at combining multiple modalities (clinical covariates, breathing patterns, cardiorespiratory metrics) into a sophisticated, predictive model. The results of this work increases our understanding of infant breathing patterns, identifies useful features and presents a practical model for the prediction of extubation readiness.

The rest of this thesis is organized as follows. Chapter 2 presents a background on existing approaches to predicting extubation readiness, machine learning methods developed for the broader technical problem of classification of physiological time-series data, and how the current work builds upon these. Chapter 3 describes the data acquisition and signal processing steps applied to the data as part of the APEX project. Chapter 4 describes the models we developed and applied to predicting from breathing patterns. Chapter 5 describes the models we developed and applied to predicting from cardiorespiratory metrics as well as its combination with clinical covariates and breathing patterns. And finally, Chapter 6 provides a unifying discussion

on the results, its implications and paths for future work.

The author of this thesis was fully responsible for the review of relevant literature (chapter 2), the design of the machine learning analysis procedure and methodology, the implementation of same, and the interpretation of results (chapters 4, 5, 6, and 7). Some of the work described in this thesis has been published by the author in [47] and in co-authorship with Lara Kanbar in [46], [32][1]. The co-investigators of the APEX project - Prof. Doina Precup, Prof. Robert Kearney, Prof. Guilherme Mendes Sant'Anna, Dr Karen Brown - and Wissam Shalish contributed through reviews, feedback and advice. The author was responsible for writing all chapters of this thesis.

[1]Paper has been accepted at the time of writing this thesis but pending publication.

# 2

# Background

The respiratory management of extremely preterm infants (birth weight, BW $\leq$ 1250g) is challenging as these infants are born with underdeveloped lungs and an inability to maintain spontaneous breathing. Endotracheal intubation and invasive mechanical ventilation (IMV) is a life-saving therapy in the first few days of life [63], but when used for long periods could lead to morbidities [42]. Physicians must wean as early as possible and prevent extubation failure since reintubation is technically difficult and has been associated with adverse effects such as lung trauma, infection, lung collapse, and death [10, 23]. Currently, there is no consensus on an objective weaning protocol. As such decisions to extubate are based on clinician judgement and observation of bedside parameters (such as blood gases, ventilator settings, etc), leading to immense practice variation and reintubation rates (10% to 70%) depending on the population studied and the time frame used to define extubation failure [23, 21].

In the following sections, we review the literature on the problem of predicting extubation readiness motivating our focus on a machine learning approach. We then summarise approaches that have been taken in employing machine learning for prediction tasks involving physiological time-series data. Finally, we put this together, highlighting the type of models we developed to predict extubation readiness based on physiological time-series measures (heart rate and respiratory variability) and clinical

variables.

## 2.1 Prediction of Extubation Readiness

### 2.1.1 Spontaneous Breathing Trials and Analysis of Clinical Variables

Spontaneous breathing trials (SBT) have been widely studied in the search for objective prediction tools for extubation readiness. Infants are put under period of no or minimal ventilator support via endrotracheal tube-continuous positive airway pressure (ETT-CPAP), while physicians observe a number of bedside measures including changes in heart rate, oxygen saturation (SpO2) and/or oxygen requirements. The infant is declared ready if certain criteria are met. In [28], an SBT failure was recorded if the infant had either a bradycardia lasting longer than 15 s, defined as a drop in heart rate below 100 beats per minute, and/or a fall in oxygen saturation below 85% despite a 15% absolute increase in the fraction of inspired oxygen. SBTs have had limited success. First, in late 1980s-1990, SBTs of 6 to 24 h were common practice. However, evidence emerged that the trials' prolonged length and low pressures increased the risk of respiratory failure [17].

More recently, clinicians shifted towards the use of shorter 3 - 5 min SBTs [30, 14]. Kamlin, Davis, and Morley [30], studied the predictive ability of the SBT in low birth weight (<1250g) preterms who all had respiratory distress at birth. The SBT achieved a high sensitivity and specificity of 97% and 73% at predicting extubation success, and was adopted as a standard of care in the institution. However subsequent prospective audits found that the routine use of SBTs did not improve weaning times or extubation success rates [31]. Chawla et al. [14] prospectively tested the usefulness

of a 5-min SBT. This was found to have high sensitivity and positive predictive value (PPV), but limited specificity and negative predictive value (NPV)[1].

Beyond SBTs, the usefulness of several clinical variables as predictors of extubation readiness has also been examined. Kamlin, Davis, and Morley [30] employed statistical tests (student t and fisher test) to evaluate the significance of tidal volumes, expired minute ventilation, and the ratio of minute ventilation between ETT-CPAP and IMV. These variables showed potential in separating infants who failed and succeeded extubation but performed worse than the SBT. A recent secondary analysis of data from a clinical trial was conducted by [13], who identified strong markers of extubation success to include higher 5-minute Apgar score, and pH prior to extubation and lower peak fraction of inspired oxygen. The authors however developed no prediction tool based on this information.

### 2.1.2 Analysis of Cardiorespiratory Variability

Clinical studies have pointed to the potential utility of measures of heart rate and respiratory variability (HRV and RV), for predicting extubation readiness [18, 8, 60], although most were conducted on adult patients. Our group was one of the first to analyse HRV and RV in preterm infant populations.

Kaczmarek et al. [28] conducted a retrospective analysis of RV in the 44 infants studied by [30]. The variability index (VI) was found to be significantly decreased in infants who failed extubation. As a predictor of extubation readiness, RV indices gave perfect sensitivity but limited specificities. By combining the VI with SBT criteria, the best specificity of 75% was obtained. A later prospective observation study of 56 preterm infants (BW $\leq$ 1250 g) indicated that HRV as well was significantly lower in infants who failed extubation [27]. HRV measures had perfect specificity and positive predictive value, but limited sensitivity and negative predictive value in predicting

[1]We refer the reader to appendix 7.1 for definitions of the metrics - sensitivity, specificity, PPV, NPV - introduced in this paragraph

extubation readiness, indicating that perhaps HRV and RV contain complementary information.

More robust analysis of RV measures was restrictive since at the time it required manual breath-by-breath analysis (the most common method) of respiratory signals. This motivated the development of AUREA - Automated Unsupervised Respiratory Event Analysis [52]. AUREA automates the process of computing RV indices in repeatable, standardized fashion that requires no human intervention. AUREA was originally developed for older infants recovering from anaesthesia following surgery, but then retuned and validated on the preterm infant population [53, 54].

Several important metrics of cardiorespiratory behaviour are computed by AUREA on a sample-by-sample basis, making it non-trivial to decide on metrics to use, what time samples to consider and how to combine these into a predictor. This motivated the first work that applied *machine learning* (ML) methods [9] to the problem of predicting extubation readiness in preterm newborns. Precup et al. [49] developed support vector machine (SVM) classifiers based on cardiorespiratory metrics computed by AUREA at the middle minute of the ETT-CPAP. This classifier yielded a sensitivity of 83% and specificity of 74% on a set of 53 (42 successes and 11 failures) preterm infants.

Though these works highlight several useful variables and approaches for the prediction of extubation readiness, the results are not to be considered conclusive since they mostly involve single-center studies of very small, heterogeneous infant populations. In addition, the predictive ability of HRV and RV indices is far from being fully explored.

## 2.2 Machine Learning for Classification of Physiological Signals

Related to our goal of developing ML-based predictors of extubation readiness from measures of heart and respiratory variability, is the body of work applying ML to predict disease, clinical events, etc from physiological time-series signals. Physiological signals (such as speech, electromyography, electroencephalography, respiratory movement, etc) naturally come as sequence data in which the samples at each instance in time are correlated. We group ML approaches that have been applied to physiological sequence data in 2 based on the input representation: approaches based on modelling complete time-series, time-agnostic models operating on (transformed) scalar covariates or windows. We review research under both categories.

### 2.2.1 Sequence Data

One of the key challenges in modelling sequence data arise from the fact that permitting long range dependencies in instances across time involves exponential blow up of parameters [37]. Several sequence modelling approaches apply the Markov property [39] to resolve this challenge. The Markov property states that a sample in time is conditionally independent of its history before the immediately preceding observation, conditioned on this observation. This is a simplifying assumption that has been found to be effective in many problems especially those involving sequence data from biological systems. Alinovi et al. [1] developed a Continuous-Time Markov Chain (CTMC) model of breathing patterns in infants experiencing disorders such as apneas. They demonstrated that the learned CTMC models accurately described respiratory rate and simulated realistic sequences of respiration of normal and apnaeic infants.

More sophisticated methods have been built on top of Markov chains such as hidden Markov models (HMM) [6], conditional random fields [34], maximum entropy Markov models [41], etc. HMMs in particular have been very successful in the analysis of speech signals for speaker recognition. By carefully selecting features based on mel-frequency ceptral coefficients (MFCC) or perceptual linear predictive coefficients, investigators developed highly accurate HMM-based speech recognisers[3, 4, 65], leading to its widespread use at industry scale. In [43] an HMM combined with a Gaussian Mixture Model density estimator was used to classify motion from rectified and filtered electromyograph signals with up to 91.25% accuracy.

Other methods that have been explored include similarity measures like dynamic time warping (DTW) which has been used to analyse sensor data from an inertial measurement unit (IMU) for gait recognition in healthy individuals[29] and for the study of gait in patients with Parkinson's disease [64].

One drawback to most of the *classical* methods above is that they require carefully, hand-engineered features from domain experts in order to work well. Deep neural networks (DNN) [35] which have gained widespread use in the last 5 years have the ability to automatically extract rich feature representations from raw data [35]. In additions, Recurrent Neural Networks such as those that use long-short term memory (LSTM) cells can capture arbitrarily long-range dependencies in sequence data and have surpassed the performance of HMMs in speech recognition [24].

For predicting surgical outcomes for cerebral palsy patients, [20] developed a novel LSTM which models joint angles obtained during the subject's gait cycle to surpass some classical models by as much as 14%. [38] conducted the first empirical study using LSTMs to predict multiple diagnoses given multivariate paediatric intensive care unit (PICU) time series (including body temperature, heart rate, diastolic and systolic blood pressure, and blood glucose, among others). The developed LSTM achieved a micro and macro AUC of 85.60% and 80.75%, respectively in detecting conditions such as diabetes mellitus, scoliosis and asthma. DNN approaches, however, have a

key drawback. DNNs generally entail thousands of free parameters, and consequently require at least as much labelled data for learning to be possible. Unfortunately in many situations in the clinical domain (as in ours) only a few hundred examples may be available.

### 2.2.2 Scalar Covariates

Overall, the ML methods described above which model sequence data directly, including non-DNN methods, can be computationally expensive to run, require fair amount of data, and are usually non-trivial to tune and train [35]. Hence approaches have been developed which either completely ignores time as a factor or hand-engineers new scalar features that summarise sequential data (e.g., using statistics, spectral information, etc).

Support vector machines (SVM) [16] are a powerful and common choice for learning from such data. Indeed, Precup et al. developed an SVM classifier to validate the hypothesis that extubation readiness can be predicted from measures of heart and respiratory variability. Time-varying measures such as respiratory frequency, cardiac frequency, etc were computed. Feature vector at each instant in time were used independently to train the SVM. Predictions were then made by a majority vote of the instance classifications. This system gave a sensitivity of 83.2% and specificity of 73.6% on a dataset 53 patients. In [45], speech signals of newborns were computed as MFCC then supplied as input to an SVM yielding sensitivity and specificity of 86% and 89%, respectively in detecting newborn asphyxia. In [44] statistical, Hjorth, amplitude and spectral measures, computed from patient electroencephalography records, were used with an SVM to identify the presence of insomnia by an accuracy of 81%.

Random Forests (RF) [12], which builds ensembles of decision trees to make non-linear, robust estimators have also been explored. Similar to [20], Schwartz et al. attempted to predict surgical outcomes for cerebral palsy patients from gait data.

They built RFs based on de-correlated feature vectors at each time instant. And as in [49] predictions were made based on the fraction of votes assigned to a given class in the sequence. The investigators demonstrated that this approach was strongly predictive of good and poor pelvis-hip outcomes (82% sensitivity and 73% specificity) for limbs undergoing surgery.

## 2.3 Developing Machine Learning Predictor of Extubation Readiness

It has been established that several clinical features and cardiorespiratory metrics contain predictive information about an infant's readiness for extubation. However, it is not clear how exactly these can be combined to develop accurate tests for extubation readiness. There appears to be complex interactions between the different variables. Moreover, high heterogeneity in the characteristics of preterm infants, type (and amount) of ventilatory support received, and other factors, make the derivation of a simple rule for extubation readiness extremely difficult. In this work, we focus on the use of ML methods, a range of tools which have been developed for turning data into actionable knowledge through fitting of complex functions and models.

The use of ML tools for the prediction of extubation readiness was first studied by our group leveraging cardiorespiratory metrics[49] and clinical variables [22] to give promising results. These were on very small patients sets and explored only one modality and a single ML method - SVM. As the patient database of APEX grew, it became necessary to validate existing results, investigate other models especially those tailored to time series modelling and combine several modalities into an accurate predictor of extubation readiness in preterm infants.

In this work we develop models based on Markov chains, HMMs, SVMs, and RFs, using a variety of inputs and input modalities including respiratory patterns,

metrics of cardiorespiratory variability and clinical variables. We adopt techniques for addressing practical issues such as class imbalance and feature selection. Through this work, we developed an increased understanding of preterm infant breathing patterns and transition behaviour. We demonstrate empirically that multi-modal classifiers could help boost performance and reliability of predictors. This thesis details work that has been recently published by the primary author in [46], [47] and [32][2], in addition to new experiments and analyses.

[2]Paper has been accepted at the time of writing this thesis but pending publication.

# 3

# Data Acquisition and Signal Processing

Data acquisition for the APEX study began in September 2013 and is ongoing at 5 tertiary-level Neonatal Intensive Care Units (NICU) in North America. The study protocol for APEX has been published in [59]. In this chapter, we present the different modalities and types of data acquired, how they were acquired and the signal processing applied to the data; with a focus on aspects relevant to the machine learing phase described in subsequent chapters.

## 3.1 Data Acquisition

Data is acquired from five NICUs: the Royal Victoria Hospital, the Montreal Children's Hospital, the Jewish General Hospital (Montreal, Quebec, Canada); the Detroit Medical Center (Detroit, Michigan, USA), and the Women and Infant's Hospital (Providence, Rhode Island, USA). Approval was obtained from the Ethics Review Boards of each institution.

Infants were eligible if they had Birth Weight (BW) $\leq$ 1250g, receiving invasive mechanical ventilation (IMV) at the time of enrollment and undergoing their first extubation attempt. Infants were excluded if they had any major congenital anomalies, or were receiving any vasopressor or sedative drugs at the time of extubation. Written informed parental consent was obtained prior to enrollment. The attending

clinician was responsible for determining extubation readiness, and data were collected immediately prior to extubation.

### 3.1.1 Acquisition of cardiorespiratory data

The following cardiorespiratory signals were acquired from each infant:

1. Chest and abdominal movements using uncalibrated Respiratory Inductance Plethysmography (RIP) recorded with Respitrace QDC system ® (Viasys ® Healthcare, USA). One RIP band is placed around the infant's ribcage (RCG) at the level of the nipple line and the other band around the abdomen (ABD) at 0.5cm above the umbilicus;

2. Electrocardiography (ECG) using 3 electrodes (Vermed, USA, © 2010) placed on infant's chest or limbs;

3. Photoplethysmography (PPG) and oxygen saturation (SAT) signals recorded with a pulse oximeter (Radical, Masimo Corp, Irvine, LA) placed on infant's hand or foot.

All signals were anti-alias filtered at 500Hz and sampled at 1000Hz using a portable analog-digital data acquisition system (PowerLab version 7.3.8, ADInstruments, Dunedin, New Zealand, © 2009). Figure 3.1 shows a representative example of the cardiorespiratory signals from one infant.

Upon the decision of the physician to extubate, these signals are acquired during 2 continuous recording periods, before extubation:

1. A 60-minute period while the infant receives any mode of conventional IMV.

2. A 5-minute period during which ventilation is switched to ETT-CPAP.

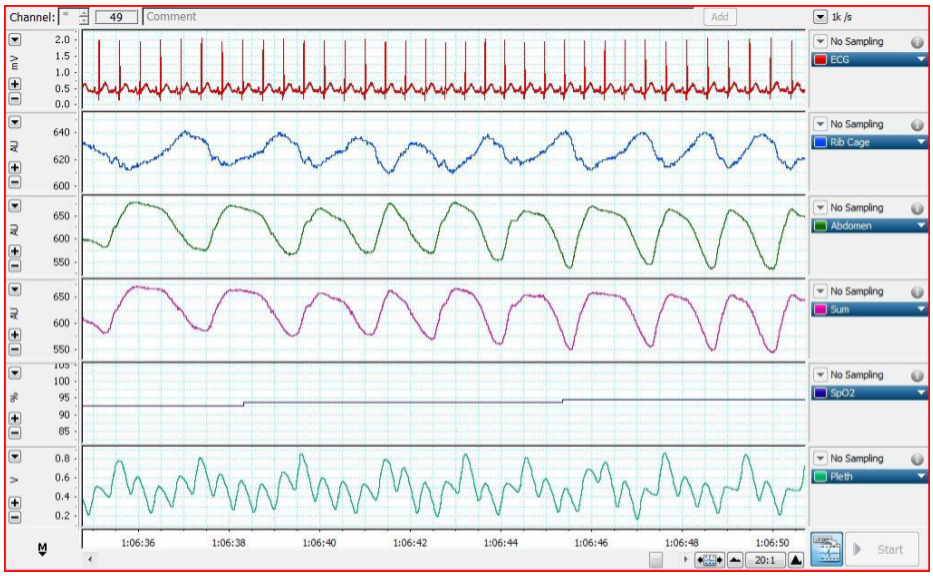


Figure 3.1: Representative example of a cardiorespiratory recording from a preterm infant. The signals displayed, from top to bottom, are: electrocardiogram, rib cage movements, abdominal movements, sum of rib cage and abdominal movements, oxygen saturation and photoplethysmography. Source: [59]

In this work, the cardiorespiratory signals recorded during the 5-minute ETT-CPAP are the main focus of analysis since this period better reflects the infants' ability for spontaneous breathing.

### 3.1.2 Acquisition of clinical data

Over 100 clinical variables are recorded during the infants stay in the NICU. These are read off the nursing flow chart and blood gas records. They include maternal characteristics, infant peri-extubation characteristics and extubation outcome measures [1]. For this work, we considered only infants' birth weight (BW) and gestational age (GA).

### 3.1.3 Extubation failure

Different investigators have defined extubation failure by various criteria and time windows. In the APEX study protocol, it is defined as the occurence of one or more

[1]The full list of clinical variables acquired can be found on Table 2 of the APEX protocol [59].

of the following criteria within 72 h after extubation: (a) FiO2 $>$ 0.5 to maintain SpO2 $>$ 88% or PaO2 $>$ 45 mmHg (for 2 consecutive hours); (b) PaCO2 $>$ 55–60 mmHg with a pH $<$ 7.25, in two consecutive blood gases done at least 1 h apart; (c) one episode of apnea requiring positive pressure ventilation with bag and mask; (d) Multiple episodes of apnea ($\geq$ 6 episodes/6 h). In this work we chose to explore another definition - *reintubation within 72 h* - based on guidance from clinicians. That is, a patient is considered to have failed extubation if the attending physician reintubates him/her within 72 h after the first extubation attempt. This choice followed observations in the data acquired so far of low correlation between the protocol definition of extubation failure and actual reintubations. The latter is a simpler definition and may be more inclined to predicting the possibility of reintubation which is an independent risk factor for morbidity and mortality in this group of infants [56, 19]. This definition is a common choice in the literature [28, 27, 30].

### 3.1.4 Database size

At the time of this work, the database contained 189 patients with BW $882 \pm 201g$ and GA $26.5 \pm 1.9$ weeks. A total of 28 (14.8%) infants failed extubation, i.e., required reintubation within 72hrs. Infants were extubated at $13.3 \pm 15.6$ days post-birth when deemed 'ready' by the attending physician.

## 3.2 Signal Processing

The raw signals acquired for each infant were further processed into cardiorespiratory metrics and respiratory patterns which characterise heart and breathing behaviour of the subject. The following subsections describes what these metrics and patterns are:

### 3.2.1 Cardiorespiratory Metrics

To obtain moving measurements of cardiorespiratory behavior, the signals were processed at every time instant into sample-by-sample metrics (at 50Hz) of power, respiratory frequency, cardiac frequency, and thoraco-abdominal synchrony, as described in [49, 52]. The metrics computed for this study include:

- Pause power in the RCG ($rp^{rc}$) and ABD ($rp^{ab}$): the power in the 0-2Hz band in a short sliding window relative to the median power in a preceding long window.
- Respiratory frequency ($rf^{ab}$): the frequency (in a sliding window) at which the highest power occurs in the 0-2Hz band, using a bank of band-pass filters with 0.2Hz bandwidth.
- Cardiac frequency using the ECG ($cf^{ec}$) or PPG ($cf^{pp}$): the frequency with the most power in the 1.5-3.5Hz band, using the Short Time Fourier Transform (STFT).
- Root-mean-square ($rms^{+}$): the sum of the RMS of the RCG and ABD in sliding windows.
- Thoraco-abdominal phase ($\Phi$): the phase difference between the RCG and ABD.
- Movement artifact power in the RCG ($bmp^{rc}$) and ABD ($bmp^{ab}$): the power in the 0-0.4Hz movement artifact band relative to the 0.4-2Hz breathing band.
- Cross-Correlation coefficient between the cardiac frequency and respiratory frequency ($\rho_0^{rf-cf}$), computed over a sliding window.

### 3.2.2 Respiratory Patterns

RIP signals sampled at 50Hz were analyzed using AUREA to extract the sequence of respiratory patterns each infant went through during the 5-minute ETT-CPAP. The 5 patterns extracted by AUREA are:

- Pause (PAU): cessation of breathing indicated by low RCG and ABD power in the breathing band (0.4-2Hz).
- Movement Artifact (MVT): periods during which there is power in the movement artifact band (0-0.4Hz) due to infant movement or nurse handling.
- Synchronous Breathing (SYB): periods during which RCG and ABD are in synchrony.
- Asynchronous Breathing (ASB): periods during which RCG and ABD are out of synchrony.
- Unknown (UNK): Ambiguous patterns not belonging to any other pattern category.

The following patterns were directly computed from the ECG and PPG signals:

- Bradycardia (BDY): artifact-free periods during which and the heart rate was below 100 beats/min.
- Desaturation (DST): artifact-free periods during which the oxygen saturation was less than 85%. Moving artifact was detected using a PPG movement artifact detector [14].

An example of RIP signals and corresponding patterns assigned by AUREA to the different samples is shown in Fig. 3.2.

### 3.2.3 Statistical and Spectral Features

A number of statistical and spectral measures were computed on each cardiorespiratory metric and breathing pattern to summarise them. This resulted to a total of 77 scalar cardiorespiratory features:

- Median, IQR, Median power, IQR of power of all AUREA cardiorespiratory metrics ($rp^{rc}$, $rp^{ab}$, $rf^{ab}$, $cf^{ec}$, $cf^{pp}$, $rms^{+}$, $\Phi$, $bmp^{rc}$, $bmp^{ab}$, $\rho_0^{rf-cf}$) - 40 features.

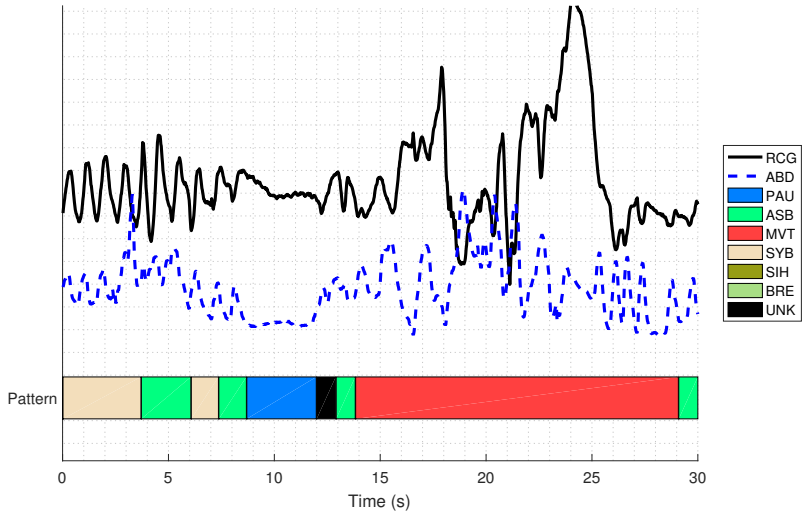


Figure 3.2: Example of a RCG and ABD signal segment and the corresponding respiratory patterns computed by AUREA.

- Kurtosis, Skewness, Median power, IQR of power - of the oxygen saturation (SAT) signal - 4 features.

- Standard deviation of the time interval between R-peaks (SDNN); the standard deviation of successive differences in the interval between R-peaks (SDSD); and the triangular index of the ECG signal - 3 features.

- Number of occurrences ($N^P$), total duration ($T^P_{tot}$), maximum length ($T^P_{max}$), pattern density ($D^P$), pattern frequency ($F^P$) of the AUREA patterns (excluding UNK), and of the BDY, DST patterns - 30 features.

where pattern density is defined as the fraction of the ETT-CPAP time spent in a pattern, and pattern frequency is defined as the number of pattern occurrences divided by the total duration of ETT-CPAP.

# 4

# Predicting from Patterns of Breathing

In this chapter, we cover methods developed to analyse the breathing patterns extracted by AUREA during the 5-minute ETT-CPAP. Our objective was to gain an empirical understanding of the transitions that infants who succeed and fail extubation make, and to leverage this information to create accurate predictive models of extubation readiness.

This chapter contributes the following: 1) we demonstrate empirically the more robust modeling capability of semi-Markov over Markov chain models for discrete time-series data, 2) we use semi-Markov chain models to understand transition structure of breathing patterns revealing key similarities and differences between infants who succeed and fail extubation, 3) we show that, in addition to generative classification via maximizing joint likelihood, the parameters of semi-Markov chains can be exploited in discriminative classifiers to improve predictive performance.

In the following sections, we describe our formulation of the methods used, the experimental setup and a discussion of results.

# 4.1 Methods

### 4.1.1 Discrete-time Markov Chain

Consider a system, illustrated by the state transition diagrams in Figure 4.1, which can generate sequences of two classes. Sequences classified as positive examples are generated by the state machine on the left while negative examples are generated by the one on the right. In both, a sequence can start from any state, lasts for 5 time steps and every time step results to a change of state (i.e., no dwell time). The primary difference between the two state machines is that the probability of transitioning to state $C$ is 0.8 for positive examples and 0.2 for negative example. Given this an example positive example could be: C -> A -> C -> B -> C while an example negative example could be A -> B -> C -> B -> A.

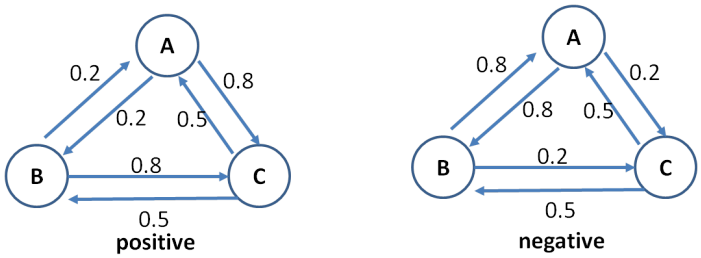


Figure 4.1: State transition diagrams for a system which generates sequences of 5 time steps which may be of a positive or negative class

Sequences generated from such a system typifies a Markov chain - one in which the next state is only dependent on the one preceding it, and independent of all other previous states. Markov chain modeling provides a tractable framework for characterizing time-series data. The values recorded at the each time step correspond to the state of the underlying Markov chain.

In order to model infant breathing patterns, we adopt discrete-time Markov chains (DTMC), in which at time $t$, the state $x_t$ takes a value from a finite set of states $S$ (in our case, the 5 respiratory patterns provided by AUREA).

### Modeling

Modeling time-series data as a Markov chain involves estimating 2 sets of parameters: the probability distribution over initial states (a vector) $\pi$ and the transition probabilities between states (a matrix) $A$. Fitting or learning the model of a Markov chain involves estimating these parameters from data. Their maximum likelihood estimates are given by [5]:

$$\pi_j = \frac{\#\ of\ sequences\ starting\ in\ j}{\#\ of\ sequences} \quad \forall j \in S \tag{4.1}$$

$$A_{i,j} = \frac{n_{ij}}{\sum_j n_{ij}} \quad \forall\ i, j \in S \tag{4.2}$$

where $n_{ij}$ is the number of time steps during which a transition from state i to j occurred. Given a time-series of observations $x_1, x_2, ..., x_T$, the joint likelihood of the sequence according to the Markov chain is given by:

$$P(x) = P(x_1) \prod_{t=2}^{T} P(x_t|x_{t-1}) = \pi_{x_1} \prod_{t=2}^{T} A_{x_{t-1},x_t} \tag{4.3}$$

Note that in our data, the start state distribution $\pi$ is unreliable due to infant and device handling at the beginning of data collection episodes, so in fact we did not include it in the model.

### Prediction

In order to apply DTMC models for classification, separate transition models $A^s$, $A^f$ were first fit to the data coming from the success and failure patients, respectively. The classification for a new sequence $x$ is done by computing its posterior likelihood with respect to both models (using Eq.4.3), and selecting the class, c whose model gives the higher likelihood:

$$\arg\max_c L(x|A^c) \tag{4.4}$$

where

$$L(x|A^c) = \prod_{t=2}^{T} A^c_{x_{t-1},x_t} \tag{4.5}$$

By this, we explore the hypothesis that the respiratory patterns of infants who succeed extubation follows a different Markov chain from those who fail.

### 4.1.2 Discrete-time Semi-Markov Chain

A discrete-time semi-Markov chain (DTSC) model differs from the DTMC in how it models the duration (number of time steps) spent in a state until a transition out of the state occurs, also known as *dwell or sojourn time*. In a DTMC model, dwell time is implicitly treated as a transition from a state to itself, whereas the DTSC model fits individual probability distributions over the dwell times in each state. Consequently, the transition matrix $A$ of the DTSC model only represents cross-state transitions with all the diagonal elements ($A_{11}$, $A_{22}$,..., $A_{|S||S|}$) set to 0.

#### Modeling

Concretely the semi-Markov chain model is characterized by 3 parameters: a start state distribution vector $\pi$; the transition matrix $A$, which stores only cross-state transition probabilities (i.e., diagonal elements are 0); and a set of dwell or sojourn time distributions $F$, which model the duration spent in each state, until a transition out of that state occurs.

The joint likelihood of a sequence of observations under a semi-Markov chain is given by:

$$P(x) = \pi_{x_1} \prod_{t=2}^{T} A_{x_{t-1},x_t} F_{x_t}(|x_t|) \tag{4.6}$$

where $F_{x_t}(|x_t|)$ is the probability of sojourning in the state $x_t$ for the duration $|x_t|$. [50]

In modeling the infant respiratory pattern sequences as a semi-Markov chain, the maximum likelihood estimates of $\pi$ and $A$ remain as before Eqs 4.1 and 4.2. To

fit the dwell time distributions, all dwell times in a breathing pattern (e.g., PAU) for one population (e.g. success patients) were obtained. Several known probability distributions (see appendix 7.2 for list) were fit to this data. The distribution which minimized the Bayesian Information Criterion (BIC) [58] was selected. This steps were repeated for all states in both success and failure groups to obtain 10 separate dwell time distributions.

**Prediction**

Classification of a new example sequence as success or failure was done as in previous section Eq. 4.4 by selecting the class of larger posterior likelihood. However in DTSC model the likelihood function is:

$$L(x|A^c) = \prod_{t=2}^{T} A^c_{x_{t-1},x_t} F_{x_t}(|x_t|) \tag{4.7}$$

It should be noted, that the framework of DTSC was useful in our application for several reasons. First, Markov chains implicitly model dwell times as an exponential distribution [9] which could introduce bias into the model if underlying data is not actually exponential. Secondly, in data characterized by very long dwell times, the transition probabilities of cross-state transitions (off-diagonal elements) go to 0, making it very difficult to get any useful information from the model. Finally, a Markov chain is highly susceptible to changes in the sampling rate of the data. Semi-Markov chains address all of these issues.

### 4.1.3 Support Vector Machine

Using the Markov model likelihood for classification can be sub-optimal if the model structure or some of the model parameters are imprecise with respect to the underlying mechanisms of the data. Discriminative models do not make probabilistic assumptions about how the inputs were generated, but rather attempt to learn a (linear or non-

linear) boundary between the groups. Support vector machines (SVMs), in particular, learn a maximum margin decision boundary [16]. In order to compare our results to the Markov and semi-Markov chain cases, we derived summary statistics from the respiratory pattern sequence of each patient and used these as inputs to train an SVM. In particular, a radial basis function (RBF) SVM was used. The key hyperparameters of the RBF SVM - box constraint, $C$ which penalizes the error function to manage overfitting, and kernel scale $\gamma$, which controls the width of the Gaussian, were tuned as described in section 4.2, the Experimental Setup.

## 4.2 Experimental Design

We developed an experimental framework to model the respiratory patterns as DTMC and DTSC, and then to compare the predictive ability of these generative methods, with a discriminatory one - SVM. The input to the Markov-based models were the sequences of respiratory patterns, summarised in table 4.1. While the derived features for the SVM are described in section 4.2.1.

Table 4.1: The 5 breathing patterns extracted by AUREA from respiratory inductive plethysmography (RIP) signals in ribcage and abdomen

| **Pattern Name** | **Code** | **Description** |
|---|---|---|
| Pause | PAU | A cessation of breathing |
| Synchronous Breathing | ASB | Ribcage and abdomen and ABD are in phase |
| Asynchronous Breathing | MVT | Ribcage and abdomen are out of phase |
| Movement Artifact | SYB | Associated with infant moving or nurse handling |
| Unknown | UNK | Ambiguous patterns not belonging to any other pattern category |

### 4.2.1 Feature Extraction for Discriminative Classification

The following features, motivated from the DTSC model, were extracted from each subject.

- Total dwell time in each respiratory pattern as a fraction of the total sequence duration, (*Dw-All*) - 5 features

- Number of transitions from pattern $i$ to pattern $j$ (where $i \neq j$) as a fraction of the total dwell time in pattern $i$, $\forall i \in S$, (*Tr-All*) - 20 features

- Number of occurrences of each respiratory pattern as a fraction of the number of occurrences of all patterns, (*Oc-All*) - 5 features

### 4.2.2 Experimental Protocol

For hyperparameter tuning of the SVM, 10-fold cross validation was employed to evaluate candidate models. The best model was selected as that which minimised the *loss* on the validation set. The *loss* is defined as *1 - balanced classification accuracy (BCA)* (see appendix 7.1.5 for a motivation on the use of balanced classification accuracy). The Markov-based models have no hyperparameters.

For evaluation of all 3 (DTMC, DTSC and SVM), leave-one-out cross validation was used. In each case, the model is fit to the data on all but one example, the validation example. The fitted model is then evaluated on the validation example. This is repeated until every example has served as validation exactly once. The evaluation metrics - sensitivity, specificity, and BCA - are then computed over the predictions on the validation examples. Due to the small nature of the dataset, no subset was left-out as a standalone test set.

Experiments were written in MATLAB [40].

## 4.3 Markov-based Modelling and Prediction

In this section, we present and discuss results obtained when using the Markov-based approaches - DTMC and DTSC - to model infant respiratory patterns and make predictions of extubation readiness.

### 4.3.1 Modelling Transitions

The transition probabilities were estimated from data via maximum likelihood. Given one transition, say PAU-MVT, its number of occurrences in the data is counted and divided by the number of occurences of the prior state (PAU) in that transition. This is repeated for all combinations of state pairs to give a total of 25 transition probabilities for each of the success and failure groups.

The Markov chain transition matrices for the success and failure populations are shown in Tables 4.2 and 4.3, respectively. Each cell in the matrix represents the probability of transitioning from the state labeled on the row to that on the column. It can be seen that the probability of self transitions (diagonal elements) account for nearly all of the transition probability on each row, leaving the probabilities of cross-pattern transitions (off-diagonal elements) close to 0. This is a reflection of extremely long dwell times relative to cross-state transitions in the data.

The symmetric KL divergence ($D_{KLS}$) between the success and failure transition distributions was close to zero 0.0019, indicating the distributions were almost identical.

The transition probabilities of the respiratory patterns were further estimated as semi-Markov chains. The transition matrices for the success and failure populations are shown in Tables 4.4 and 4.5. By collapsing self-transitions, it can be seen that resolution was greatly increased in the off-diagonal elements of the matrices. This

Table 4.2: Respiratory state transition probabilities for the **Success** population modeled as a DTMC. Self-state transitions are in **bold**. They account for a very high proportion of transitions, thereby reducing resolution in cross-transition probabilities.

| | **PAU** | **ASB** | **MVT** | **SYB** | **UNK** |
|---|---|---|---|---|---|
| **PAU** | **0.9936** | 0.0019 | 0.0004 | 0.0022 | 0.0018 |
| **ASB** | 0.0006 | **0.9953** | 0.0010 | 0.0013 | 0.0018 |
| **MVT** | 0.0012 | 0.0029 | **0.9931** | 0.0020 | 0.0007 |
| **SYB** | 0.0003 | 0.0005 | 0.0003 | **0.9977** | 0.0012 |
| **UNK** | 0.0015 | 0.0031 | 0.0003 | 0.0055 | **0.9895** |

Table 4.3: Respiratory State Transition probabilities for **Failure** population modeled as a DTMC. Self-state transitions are in **bold**. They account for a very high proportion of transitions, thereby reducing resolution in cross-transition probabilities.

| | **PAU** | **ASB** | **MVT** | **SYB** | **UNK** |
|---|---|---|---|---|---|
| **PAU** | **0.9920** | 0.0022 | 0.0007 | 0.0020 | 0.0032 |
| **ASB** | 0.0005 | **0.9955** | 0.0007 | 0.0013 | 0.0020 |
| **MVT** | 0.0007 | 0.0021 | **0.9934** | 0.0028 | 0.0010 |
| **SYB** | 0.0001 | 0.0005 | 0.0003 | **0.9978** | 0.0012 |
| **UNK** | 0.0013 | 0.0027 | 0.0004 | 0.0056 | **0.9899** |

revealed interesting results. It was observed that the *most probable transition* from the breathing patterns (SYB and ASB) and UNK pattern was the same in both infants who succeeded and those who failed extubation (shown in bold, black font). Whereas it differed for the non-breathing states (PAU and MVT) (shown in bold, red font).

Further, the symmetric KL divergence ($D_{KLS}$) between the 2 transition matrices for the semi-Markov model was 0.27. This increase from the DTMC case suggests that the DTSC model resulted to the learning of more discriminating characteristics between both groups of infants.

### 4.3.2 Modeling of Dwell/Sojourn Time Distributions

The dwell time distributions of the DTSC were estimated as described in section 4.1.2. The results are summarised in Table 4.6. It was observed that in each pattern, the *distribution type* which best fit the dwell times in both populations were same,

whereas the *distribution parameter values* differed. This suggests that the 2 groups of infants do not differ in the manner in which they sojourn in a single state but in the duration spent in that state before switching to a different one. This may also be an indication of some underlying consistency in breathing behaviour of premature infants in spite of extubation outcome.

It should also be noted that the dwell time was distributed exponentially only in the Pause pattern. As discussed earlier, the use of a DTMC model implicitly assumes an exponential distribution for all patterns. The DTSC framework has thus allowed for a more expressive and accurate representation. Detailed plots of the probability density functions (PDF) of sojourn times in all states are shown in Fig 4.2, as well as the distributions of best fit based on the Bayesian information criterion (BIC).

### 4.3.3 Prediction

As described, leave-one-out cross validation was employed to estimate the predictive capability of the models. The results using the DTSC model are summarised in Table

Table 4.4: Respiratory state transition probabilities for the Success population modeled as a Semi-Markov chain

| | **PAU** | **ASB** | **MVT** | **SYB** | **UNK** |
|---|---|---|---|---|---|
| **PAU** | 0 | 0.27 | 0.09 | 0.26 | **0.38** |
| **ASB** | 0.10 | 0 | 0.16 | 0.29 | **0.45** |
| **MVT** | 0.12 | 0.32 | 0 | **0.43** | 0.14 |
| **SYB** | 0.06 | 0.25 | 0.15 | 0 | **0.54** |
| **UNK** | 0.13 | 0.28 | 0.04 | **0.55** | 0 |

Table 4.5: Respiratory State Transition probabilities for Failure population modeled as semi-Markov chain

| | **PAU** | **ASB** | **MVT** | **SYB** | **UNK** |
|---|---|---|---|---|---|
| **PAU** | 0 | 0.28 | 0.06 | **0.39** | 0.28 |
| **ASB** | 0.12 | 0 | 0.21 | 0.28 | **0.40** |
| **MVT** | 0.17 | **0.41** | 0 | 0.32 | 0.09 |
| **SYB** | 0.14 | 0.21 | 0.14 | 0 | **0.52** |
| **UNK** | 0.15 | 0.30 | 0.03 | **0.52** | 0 |

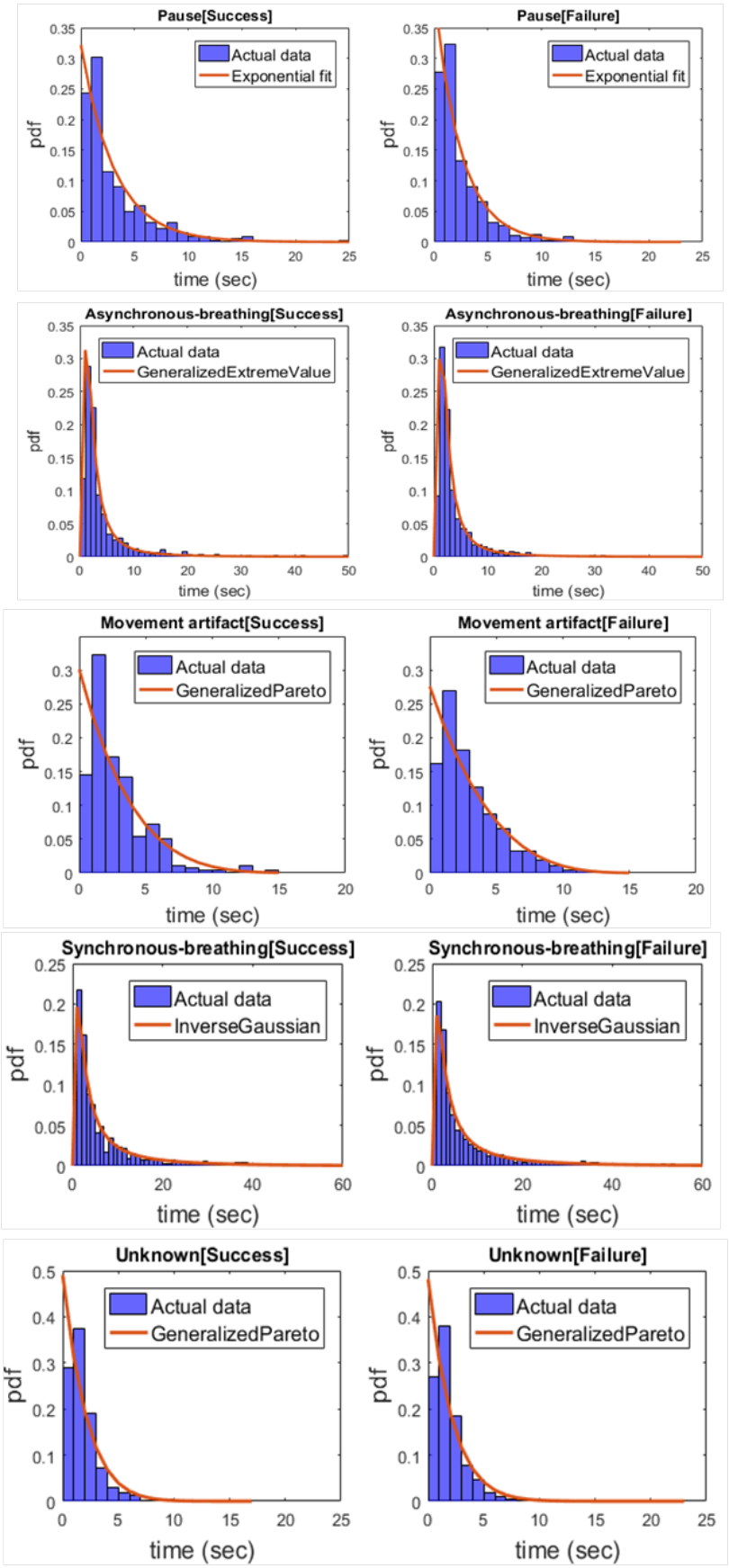


Figure 4.2: Probability Density Functions (PDF) of Dwell Time Distributions in all 5 respiratory patterns for success and failure patients

Table 4.6: The type and parameters of the distributions of best fit to the dwell (or sojourn) times in each respiratory pattern for success and failure patients.

| | **Success** | **Failure** |
|---|---|---|
| **Pause** | Exponential<br>$\mu = 2.51$ | Exponential<br>$\mu = 2.94$ |
| **Asynchrony** | GeneralizedExtremeValue<br>k=0.63, $\sigma$ = 1.30, $\mu$=1.85 | GeneralizedExtremeValue<br>k=0.65, $\sigma$ = 1.36, $\mu$=1.81 |
| **Movement** | GeneralizedPareto<br>k=-0.22, $\sigma$ = 3.62 | GeneralizedPareto<br>k=-0.11, $\sigma$ = 3.31 |
| **Synchrony** | InverseGaussian<br>$\mu$ =8.61, $\lambda$ = 3.61 | InverseGaussian<br>$\mu$ =7.83, $\lambda$ = 3.41 |
| **Unknown** | GeneralizedPareto<br>k=-0.07, $\sigma$ = 2.07 | GeneralizedPareto<br>k=-0.10, $\sigma$ = 2.05 |

4.7. *Lk-ALL* refers to the standard form of the likelihood function (Eq 4.7) in which all patterns along the sequence are considered. Success patients were identified at a rate (sensitivity) of 73% while specificity was 50%.

Further, we examined the predictive value of individual patterns/states. In particular, to compute the likelihood of a test sequence based on one pattern, the product of cross-state transitions emanating from only that state are taken. As before, this likelihood is computed with respect to the transition models for the 2 classes, and a prediction is made by selecting the class whose model gave higher likelihood. Results are shown accordingly in Table 4.8 where *Lk-STATE* represents prediction made using likelihood of the "STATE" specified. The best performance was obtained by the Pause pattern which gave the lowest misclassification loss of 0.37 and the highest specificity of 68%.

Table 4.7: Performance of Semi-Markov chain model. Lk-ALL or Lk-STATE refers to classification using likelihood of chain considering all states or a specified "STATE".

| Approach | Sensitivity | Specificity | Loss |
|---|---|---|---|
| **DTSC Model** | | | |
| Lk-ALL | 0.73 | 0.50 | 0.38 |
| Lk-PAU | **0.58** | **0.68** | **0.37** |
| Lk-ASB | 0.48 | 0.63 | 0.45 |
| Lk-MVT | 0.50 | 0.68 | 0.41 |
| Lk-SYB | 0.53 | 0.68 | 0.40 |
| Lk-UNK | 0.44 | 0.61 | 0.48 |

## 4.4 Support Vector Machine Prediction

In the section, we present and discuss results RBF SVM to predict extubation readiness based on features extracted from respiratory patterns of infants.

### 4.4.1 Model Selection and Parameter Search

Different combinations of the feature set were evaluated. First all 30 features described in 4.2.1 (Dr-Oc-Tr-ALL) from all patterns were used. Then, similar to the generative case, the predictive value of each individual pattern/state was evaluated. Concretely, for each state, the dwell time in that state, *Dw-STATE* (1 feature), the cross-transitions, *Tr-STATE* (4 features) and the occurrence count, *Oc-STATE* (1 feature) were combined, *Dw-Oc-Tr-STATE* (6 features) to train the classifier.

10-fold cross-validation in a grid search to find the best pair of hyper-parameters (box constraint $C$ and kernel scale $\gamma$) values by optimising for the balanced misclassification loss. This grid search was repeated for each feature set since each would have different optimal values of $C$ and $\gamma$.

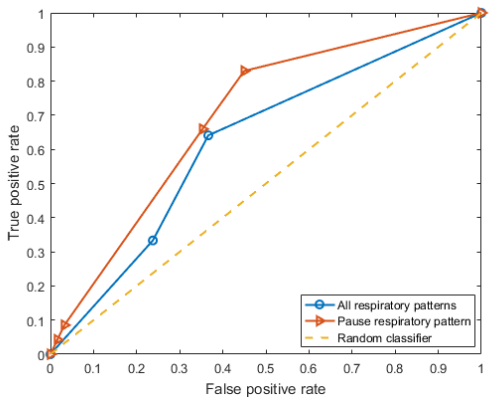


Figure 4.3: Receiver-operating characteristic (ROC) curve for support vector machine trained on summary features of all respiratory patterns (AUC=0.62) and on summary features of only the Pause pattern (AUC=0.70)

### 4.4.2 Prediction

The optimal values for $C$ and $\gamma$ were then finally evaluated using leave-one-out cross validation. The results are summarised in Table 4.8. The highest specificity of 84% was obtained when using features of only the Pause pattern, which also gave the lowest loss of 0.31. It could also be observed that whereas PAU and MVT patterns gave higher specificities , SYB and ASB gave higher sensitivities . This is likely an indication that the Pause and Movement patterns characterise better patients who may fail extubation while the breathing patterns better characterise patients who succeed.

In Fig. 4.3, we show the receiver-operating characteristic (ROC) curve for the best model which uses only PAU features *Dw-Oc-Tr-PAU* compared with the model which uses all pattern features *Dw-Oc-Tr-ALL*. The ROC curves were obtained by fixing $C$ at the optimal value and varying $\gamma$. It can be seen that the Dw-Oc-Tr-PAU provides a more reliable predictive surface with a high area under the curve (AUC) of 0.70.

Table 4.8: Performance of best support vector machine model. Dw, Oc, Tr refer to features extracted based on dwell time, occurrence count and transitions in states.

| Approach | Sensitivity | Specificity | Loss |
|---|---|---|---|
| **Discriminative (SVM)** | | | |
| Dw-Oc-Tr-ALL | 0.63 | 0.64 | 0.37 |
| Dw-Oc-Tr-PAU | **0.54** | **0.84** | **0.31** |
| Dw-Oc-Tr-ASB | 0.75 | 0.38 | 0.44 |
| Dw-Oc-Tr-MVT | 0.58 | 0.60 | 0.41 |
| Dw-Oc-Tr-SYB | 0.81 | 0.26 | 0.46 |
| Dw-Oc-Tr-UNK | 0.43 | 0.62 | 0.48 |

## 4.5 Discussion

We demonstrated the practical application of semi-Markov chains for modeling and classification of respiratory pattern behaviour of preterm infants in the period prior to extubation. We showed that semi-Markov chain models provide more expressive and robust details about the underlying time series compared to Markov chain models. In terms of sojourn time behaviour, the model revealed consistency between the success and failure groups in all respiratory states. Differences were highlighted primarily in transition behaviour arising from the Pause and Movement Artifact patterns.

Prediction results confirmed that these 2 patterns provide more discriminating information (especially for patients who failed extubation) than any other pattern. That the Pause pattern is a strong indicator of infants not ready for extubation is well aligned with existing clinical knowledge. However, it was interesting to observe that the Movement Artifact pattern is also a good indicator, suggesting that infants prone to fail are more restless and require more nurse handling.

The best performance obtained was specificity of 84% and sensitivity of 54% using Pause features in a support vector machine (SVM) classifier. This shows a very good failure detection rate but at a fairly high cost to prediction of success patients.

The use of automatically extracted respiratory patterns for prediction provides an approach that unveils intuition and enhances interpretable models. We emphasize

that all babies used in this study were deemed ready for extubation by an attending clinician, so these results constitute an improvement in detecting problem cases over current practice. The advantage of using an automated approach is that we can provide a quantified measurement of the breathing patterns, which supports more repeatable and precise clinical decisions.

# 5

# Predicting from Metrics of Cardiorespiratory Variability

In this chapter, we present models and experiments developed to leverage measures of heart and respiratory variability in order to predict extubation readiness. These time-series measures differ from the breathing patterns in that they are continuous-valued and multivariate, and potentially entail more detailed observations of infants' cardiorespiratory behaviour.

## 5.1 Methods

We developed classifiers based on *Hidden Markov Models*, a generative approach to modelling sequence data, and *Random Forests*, a discriminative classifier. These methods and how they were applied to the problem of predicting extubation readiness are described in the following sub-sections.

### 5.1.1 Gaussian Hidden Markov Model

#### Structure

Hidden Markov Models (HMM) were first introduced in late 60s by Baum et al. [6], and gained popularity as a state-of-the-art model for automatic speech recognition

[3, 25, 51, 7, 24] in subsequent decades. HMMs provide a framework for modelling sequence data as a probabilistic process driven by latent or hidden states.

HMMs make 2 conditional independence assumptions. The sequence of hidden states is assumed to be a Markov chain, i.e, a state is only dependent on the one preceding it. And the probability of an observation at a given time step in a sequence is dependent only on the hidden state at that time. Figure 5.1 illustrates the HMM structure as a graphical model. These 2 conditional independence properties of the HMM allow for data-efficient and tractable algorithms for inference on sequence data.

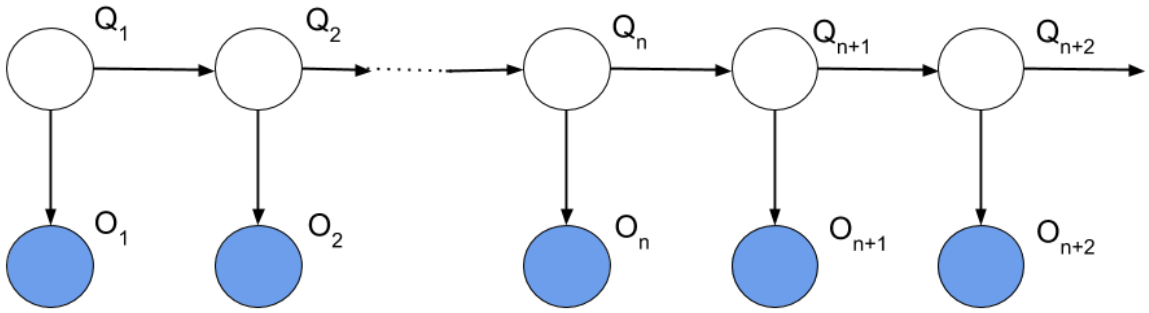


Figure 5.1: Graphical structure of a hidden Markov model (HMM)

### Parameters

Formally, consider that we have a set of observations for $T$ time steps $O_1, O_2, ..., O_T$ denoted as $O$. The observations may be multi-variate such that each $O_t$ is a vector of continuous values of dimension $D$ representing the number of features being observed. We also have a corresponding sequence of hidden states $Q_1, Q_2, ..., Q_T$ denoted $Q$. Each $Q_t$ usually takes from one of $K$ discrete state values. Given the structure of the HMM (5.1), the joint probability distribution of the observations and hidden states decomposes as:

$$P(O, Q) = P(Q_1) \prod_{t=1}^{T-1} P(Q_{t+1}|Q_t) \prod_{t=1}^{T} P(O_t|Q_t) \tag{5.1}$$

The HMM is thus defined by a set of parameters $\theta$ which include:

1. The distribution of start states, $P(Q_1)$

$$\pi = \{\pi_1, \pi_2, ...\pi_K\}$$

2. The transition probabilities from $Q_t$ to $Q_{t+1}$, $P(Q_{t+1}|Q_t)$

$$A = \{a_{ij}\} \qquad where \quad i, j \in [1, K]$$

3. The emission densities of $O_t$ from $Q_t$, $P(O_t|Q_t)$

$$B = \{b_j(O_t)\} \qquad where \quad j \in [1, K]$$

**Density Representation and HMM Variants**

In our problem we assumed discrete hidden states, thus $pi$ and $A$ were represented as discrete probability densities. The observed variables however are continuous-valued vectors - the multivariate metrics of cardiorespiratory variability. We had to apply some restriction on the form of the probability density function of the emission densities to ensure that $B$ can be estimated in a consistent way. As is common in the literature [51, 36], we modelled $B$ as a multivariate Gaussian distribution, hence the term Gaussian density hidden Markov model (GHMM). The use of mixture models such as Gaussian mixture model (GMM) has also been applied extensively [26] in cases where it is suspected that the conditional distributions of observations over the hidden states is a mixture distribution. The applicability of GMM was not investigated in this study.

Other HMM variants exist based on the structure of the transition matrix $A$. Left-to-right HMMs [9], for instance, enforce a transition model in which states are ordered such that the state $Q_{t+1}$ must either be same or of higher order than that at $Q_t$. In this work, We used an ergodic transition model, i.e., one in which every possible combination of state transitions is allowed. Evidence from Markov chain modelling of the underlying breathing patterns (chapter 4) suggested no special transition structure.

**Inference**

Given a specific HMM, there are 3 main questions that one could seek to answer [51]: how to estimate the model parameters $\theta$ given a set of observed sequences; how to compute the likelihood of a sequence of observations given a model $\theta$; and how to obtain the state sequence which best explains an observed sequence. The experiments carried out in this work were concerned primarily with the first and second, as it was of interest to fit HMMs to our data and to also utilise the learned model for prediction via maximising sequence likelihood.

The conditional independence structure in the HMM has allowed for the development of efficient and well-studied algorithms to address these inference questions. For the problem of parameter estimation (or model fitting), we employed the *Baum-Welch algorithm* [6]. Baum-Welch is a special case of expectation-maximisation (EM) algorithms which finds a maximum likelihood solution through iterative updates from initial (random) guesses. For computing the likelihood of a sequence, the *forward procedure* [51] was used.

**Prediction with HMM**

In order to use this framework for prediction, we fit an HMM each to the two populations of infants - extubation successes $\theta^s$ and failures $\theta^f$ using the Baum-Welch algorithm. Given a test sequence $O$, its likelihood or posterior probability with respect to both models. The sequence is then predicted as the class of the model which gave a higher likelihood, i.e., we assign sequences to the class it is more likely to have come from under the specific HMM.

$$\arg\max_c L(O|\theta^c) \tag{5.2}$$

where $L(O|\theta^c)$ is the likelihood of sequence $O_1, O_2, ..., O_T$ under the model $\theta^c$, computed efficiently using the forward procedure.

### 5.1.2 Balanced Random Forest

For discriminative classification, we used Random Forest (RF) classifiers [12]. The random forest classifier is a bagging machine learning method [11] which works by training an ensemble of decision trees in parallel. Each tree is trained on a subset of the examples and features in the dataset. This approach permits each decision tree to learn something new and different about the dataset. During testing, each tree in the forest makes an independent prediction on the new example. The predictions from all trees are then averaged to obtain a single prediction for that example. Such bagging methods help to reduce variance and the chances of overfitting to data. RF was selected because it had been shown in previous work that linear classifiers are inadequate for this difficult problem of predicting extubation readiness [49]. By leveraging multiple decision trees, RFs have the ability to learn complex, non-linear decision boundaries. Additionally, because the number of correctly classified samples at each leaf of the decision trees can be examined, feature importance ratios which indicate a feature's contribution to the classification output can be ultimately computed from RFs. This information can inform feature selection [12].

The RF classifier, like many machine learning algorithms (such as logistic regression and support vector machines) encounters difficulty in making good predictions when the number of examples in the different classes is imbalanced. In the case of RFs, the skew in the dataset could be worsened in some or all of the subsets passed to the trees, potentially leading to trees that are only good at predicting the majority class. Our dataset has a high class imbalance with about 85% being data of infants who succeeded extubation. We addressed the class imbalance challenge through random undersampling of the majority group (success examples) before training each decision tree. In particular, we ensured that the subsets passed to the decision trees have equal number of success and failure examples. This type of random forest has been presented in literature as a *balanced random forest* (BRF) [15].

### 5.1.3 Incorporating Clinical Decision

Clinically, it is quite common that infants who are older and larger at birth tend to be extubated successfully, and that the difficulty in deciding when to extubate lies primarily in the younger and smaller infants. To analyze this empirically, we examined the gestational age as a function of the birth weight of our infant population Figure 5.2. Of the 80 babies who were at least 27 weeks old or weighed above 1000g, 76 (95%) were extubation successes. We applied a rule to encode this choice - all infants above 27 weeks or 1000g were automatically classified as success. Our predictors (GHMM and BRF) were then trained on only the population of young and small infants. In doing this, we encode the choice of the clinician to extubate the low-risk population of older babies and to focus the efforts of the classifier on the difficult, younger segment of the population. This classifier involves a 2-stage process – the clinical decision/stratification rule, followed by a GHMM or BRF classifier (CD-GHMM or CD-BRF), illustrated in Figure 5.3

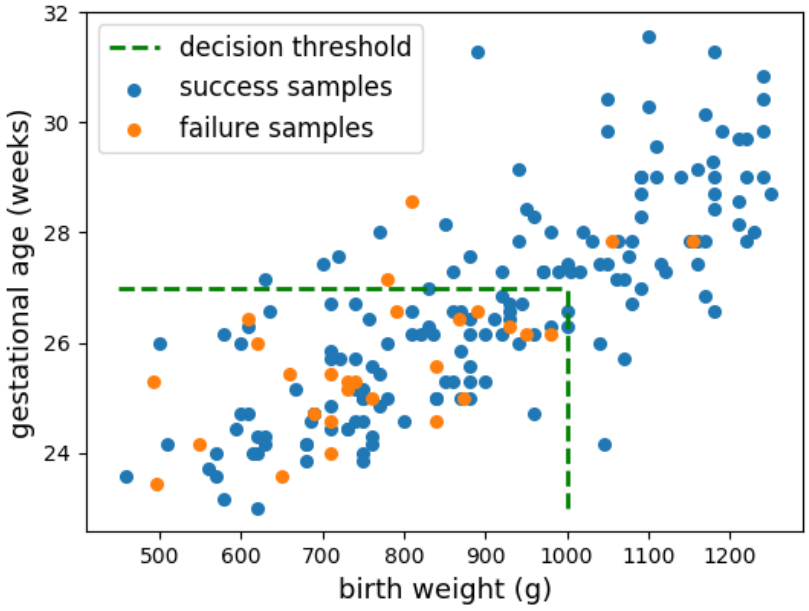


Figure 5.2: Gestational Age (GA) vs Birth Weight (BW) of the patient population showing decision threshold separating the older, larger patients. 95% of patients above threshold were successfully extubated.

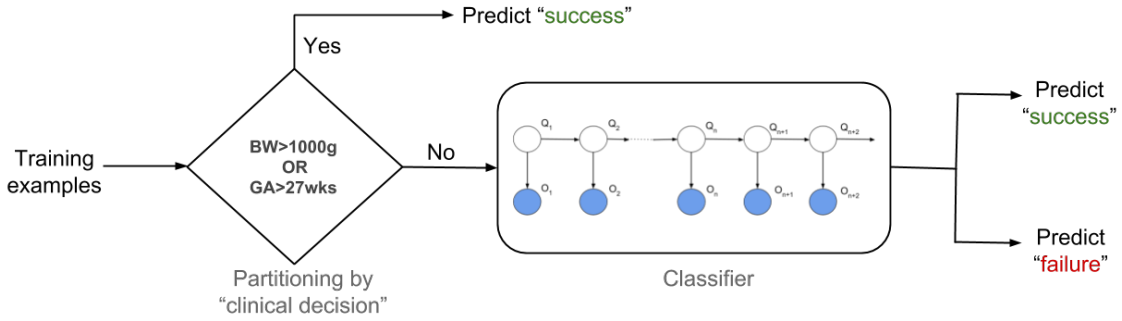


Figure 5.3: Structure of CD-GHMM and CD-BRF classifiers which incorporate clinical decision to exclude low-risk patients from training process.

## 5.2 Experimental Design

We setup an experimental framework to analyse and compare the predictive ability of hidden Markov models and random forests for correctly determining extubation readiness in extremely preterm infants. Two sets of experiments were conducted evaluating a total of 5 methods. The first set evaluated the hidden Markov model approaches of GHMM and CD-GHMM; while the second set evaluated random forest classifiers: RF, BRF and CD-BRF.

### 5.2.1 Features

As features, we used the 12 metrics computed by AUREA during the 2nd minute [1] of the 5-minute ETT-CPAP period. Each metric, sampled at 50Hz, represents a unique measure of variability relating to either the cardiac or respiratory system. Table 5.1 summarises all metrics.

For training the GHMM, the data from each patient was a 12 x 300 dimension signal. However for training the RF classifiers, it was necessary to summarise the signals to a 12 x 1 vector since RFs treats each instance (vector of observed variables)

[1] Previous work [49] had found the 2nd minute to be the most predictive time period.

Table 5.1: Time-series metrics of cardiorespiratory variability used in the hidden Markov Models - GHMM and CD-GHMM

| Metric Code | Description |
|---|---|
| $rp^{rc}$ & $rp^{ab}$ | Pause power in the RCG and ABD |
| $rf^{ab}$ | Respiratory frequency |
| $cf^{ec}$ & $cf^{pp}$ | Cardiac frequency using ECG and PPG |
| $rms^{+}$ | Sum of the root mean-square of RCG and ABD |
| $\Phi$ | Phase difference between RCG and ABD |
| $bmp^{rc}$ & $bmp^{ab}$ | Movement artifact power in RCG and ABD |
| $\rho_0^{rf-cf}$ | Cross-Correlation coefficient between the cardiac frequency and respiratory frequency |

in time as IID. Table 5.2 summarises the 77 scalar cardiorespiratory features that were used to train the RFs.

Table 5.2: Scalar cardiorespiratory features used in random forest classifiers - RF, BRF, CD-BRF

| | Computed features | Count |
|---|---|---|
| **Metrics** $rp^{rc}$, $rp^{ab}$, $rf^{ab}$, $cf^{ec}$, $cf^{pp}$, $rms^{+}$, $\Phi$, $bmp^{rc}$, $bmp^{ab}$, $\rho_0^{rf-cf}$ | Median, IQR, Median power, IQR of power | 40 |
| **SAT** | Kurtosis, Skewness, Median power, IQR of power | 4 |
| **ECG** RR Intervals | SDNN, SDSD, triangular index | 3 |
| **Patterns** PAU, MVT, ASB, SYB, BDY, DST | $N^P$, $T^P_{tot}$, $T^P_{max}$, $D^P$, $F^P$ | 30 |

### 5.2.2 Experimental Protocol

To use our relatively small dataset as efficiently as possible, we did not leave out a fixed test set. Instead we employed 5-fold stratified cross-validation (SCV) for model selection and evaluation. The dataset is split into 5 subsets (folds) of roughly equal number of examples and which maintain approximately same success-failure

proportion. To evaluate a given model, 4 folds are used to train while 1 fold is left for testing. This is repeated until all folds have been used exactly once as a test fold. This process enables the use of each example at least once for both training and testing.

For each model (i.e. hyperparameter setting) trained, we tracked 3 performance metrics: sensitivity or the success detection rate, specificity or failure detection rate and balanced accuracy , which is the average of the two. We selected the model that gave the best balanced accuracy. The selected model was further investigated to gain deeper understanding of how it makes correct and false predictions. For the HMMs, this involved computing and visualising the predicted conditional likelihoods while for the RFs we looked at the average predicted probabilities from all trees in the forest. In addition, receiver operating characteristic (ROC) curve and the area under the curve (AUC) were calculated for the optimal RF models to get a sense of reliability. The AUC was generated by varying a threshold on the average predicted probability. Performance metrics are always reported on the test set.

### Model Selection - RF, BRF, CD-BRF

Model selection involves finding the settings of hyper-parameters which gives optimal performance from a given model. In the random forest classifiers (RF, BRF and CD-BRF), the hyper-parameters include settings that affect the underlying decision trees – such as the number of features to randomly draw at each node of the tree, the maximum depth before terminating the tree and minimum number of examples required at a leaf node – as well as settings that govern the random forest itself – such as the number of trees to train. These parameters generally control the bias-variance tradeoff between fitting an overly complex model and an overly simplistic one [12]. We performed an exhaustive search of several combinations of these hyperparameters using the 5-fold SCV procedure described above (The complete range of hyperparameters explored is summarised in Appendix 7.4). The best hyperparameter setting was chosen as that which gave the best balanced accuracy on the test set averaged over

all folds.

### Model Selection - GHMM, CD-GHMM

The HMMs have a number of hyperparameters namely the number of hidden states, the *covariance type* for the emission densities, the initialisation for the HMM's parameters ($\pi$, $A$ and $B$) and the termination criterion for the Baum-Welch algorithm. The number of hidden states and initialisation were fixed, while the optimal setting for the others was found via hyperparameter search with 5-fold SCV. Concretely, the number of hidden states was set to 5 corresponding to the number of known underlying breathing states (from chapter 4). Considering that HMMs are very sensitive to initialisation and prone to converging at local optima, the initial values of the HMM's parameters ($\pi$, $A$ and $B$) must be set smartly [51]. We initialised $\pi$ as a uniform multinomial distribution and $A$ and $B$ as the transition matrix of the respiratory pattern sequences and as the conditional densities of the observed data given each of the 5 patterns, respectively. Using 5-fold SCV, all models were trained for a full 30 epochs[2] of Baum-Welch. A range of 4 covariance types (Table 5.3) for the emission density were explored controlling the bias-variance trade-off with "full" being the most complex model. Performance was recorded at the end of every epoch. Learning curves (plots of performance metrics as a function of number of epochs) were utilised to pick the best model based on the test set.

All experiments were written in Python [55] with the scikit-learn library [48].

[2]Based on preliminary tests, 30 was chosen as a good upper limit for the termination criterion

Table 5.3: Description of the 4 types of covariance matrices explored in GHMM and CD-GHMM classifiers

| **Covariance type** | **Description** |
|---|---|
| Spherical | Each hidden state uses a single variance value that applies to all observation features |
| Diagonal | Each state uses a diagonal covariance matrix |
| Full | Each state uses a full (i.e. unrestricted) covariance matrix |
| Tied | All states use the same full covariance matrix |

## 5.3 Hidden Markov Modelling and Prediction

We present the results of experiments using the HMM-based classifiers - Gaussian hidden Markov model (GHMM) and clinical decision with Gaussian hidden Markov model (CD-GHMM).

In Table 5.4, we summarise the performance of the GHMM and CD-GHMM for each covariance type. The best performing GHMM was based on a spherical covariance matrix with sensitivity 69% and specificity 59%. On the other hand the best performing CD-GHMM used a tied covariance matrix attained sensitivity 79% and specificity 57% on the test set. Overall, the CD-GHMM performs 10% better than the GHMM in detecting success patients, but at a 2% cost in detecting failure patients. It was not possible to fit full covariance matrices for the CD-GHMM classifier as the number of examples in each fold was greatly reduced due to the clinical stratification.

Learning curves for the best performing GHMM (spherical covariance) is shown in Figure 5.4. It can be seen that the test accuracy peaks at epoch 8 and decreases afterwards, even through the training accuracy continues to rise steadily up to the final epoch. This is an instructive example of why model selection must not be based on training set performance.

In Figure 5.5 we show the learning curves for the best performing CD-GHMM

Table 5.4: Performance of Gaussian hidden Markov model (GHMM) and clinical decision with Gaussian hidden Markov model (CD-GHMM) classifiers using different covariance types

| | GHMM | | | CD-GHMM | | |
|---|---|---|---|---|---|---|
| Cov. type | Sensitivity | Specificity | Bal Acc | Sensitivity | Specificity | Bal Acc |
| Spherical | **0.69** | **0.59** | **0.64** | 0.91 | 0.21 | 0.56 |
| Diagonal | 0.92 | 0.19 | 0.55 | 0.91 | 0.14 | 0.53 |
| Full | 0.94 | 0.07 | 0.51 | - | - | - |
| Tied | 0.58 | 0.56 | 0.57 | **0.79** | **0.57** | **0.68** |

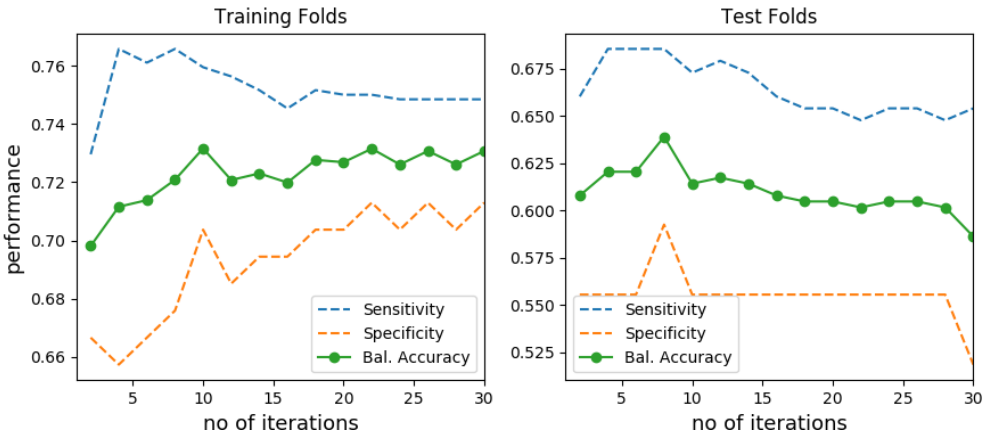


Figure 5.4: Learning curves for best GHMM (spherical covariance)

(tied covariance). The balanced accuracy on the test set increases slightly, peaks at 68% (14th epoch) before dropping to a flat value of 65%. The CD-GHMM also shows clear signs of overfitting in its failure detection model (left plot on Fig. 5.5). This is likely due the attempt to fit such an expressive generative model to the even reduced number of failure examples resulting from clinical stratification. Learning curves for other models tried are shown in Appendix 7.5.

To better understand the resulting hidden Markov models trained to predict success and failure examples, we visualised the distribution of the likelihoods assigned by the models to the examples at test time. Figure 5.6 shows this for the CD-GHMM classifier - the left plot is the distribution of likelihoods assigned by success model to all patients while the right plot is that assigned by the failure model. As expected, the

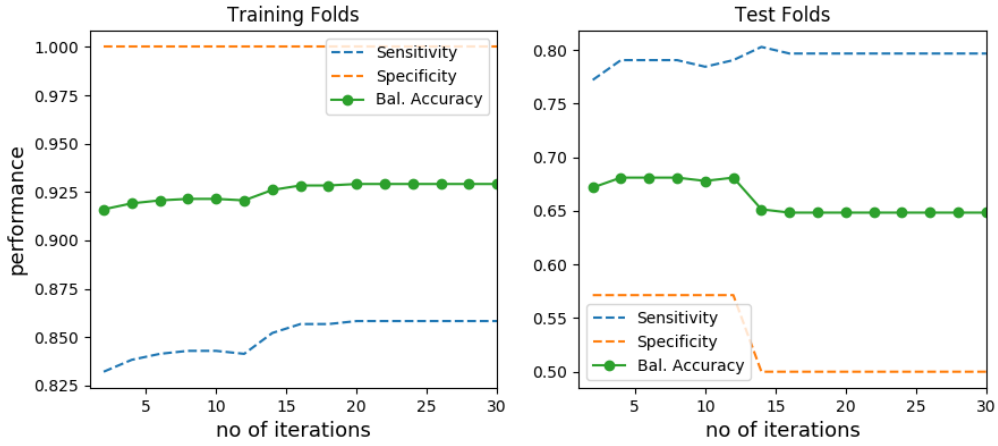


Figure 5.5: Learning curves for best CD-GHMM (tied covariance)

number of success patients with high likelihoods on the success model is more than the number of success patients with high likelihoods on the failure model. However, the failure patients also appear to accumulate high likelihood scores on the success model than on the failure model. This suggests that the problem of class imbalance precluded the learning of a more robust failure prediction model. A similar trend in seen in all likelihood distributions for the other models tried (see Appendix 7.6)

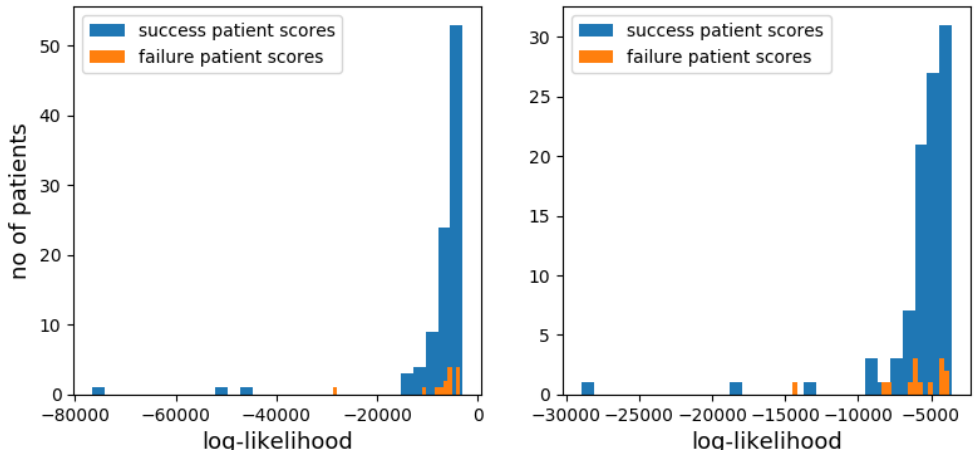


Figure 5.6: Distribution of likelihood scores of success and failure patients in the test folds considering the Success (**Left**) and Failure (**Right**) models of the best performing hidden Markov model - CD-GHMM (tied covariance).

Table 5.5: Performance of random forest (RF), balanced random forest (BRF) and clinical decision with balanced random forest (CD-BRF) classifiers

| | **Sensitivity** | **Specificity** | **Balanced Accuracy** | **AUC** |
|---|---|---|---|---|
| RF | 0.86 | 0.43 | 0.65 | 0.65 |
| BRF | 0.60 | 0.75 | 0.68 | 0.66 |
| CD-BRF | **0.78** | **0.71** | **0.75** | **0.74** |

## 5.4 Random Forests Estimation and Prediction

Here we present the results using 3 variants of the random forest classifier - the standard random forest (RF), balanced random forest (BRF), and clinical decision with balanced random forest (CD-BRF) - to learn cardiorespiratory behaviour and predict extubation readiness.

Table 5.5 details the performance of all 3 at the optimal hyperparameter setting. First, it can be seen that the CD-BRF performs best, attaining the highest balanced accuracy of 75% among all 3 classifiers. Second, whereas the standard random forest classifier (RF) learns a skewed model with a high false positive rate (as seen in the low 43% specificity), BRF and CD-BRF use random undersampling of the majority class and attain a better balance between true positive and false positive rates (with specificity of 75% and 71% respectively).

The receiver operating characteristic curves for each are shown in Figure 5.7. The CD-BRF classifier also has the highest area under the curve (AUC) of 0.74 indicating that it had the best performance over a wider range of values.

The importance weights assigned by the best classifier (CD-BRF) were examined to understand what features contributed to prediction outcome. It was found that only 17 of the 77 cardiorespiratory variability features had non-zero weights, suggesting that the remaining 60 features had no correlation with the outcome. These 17 features and their corresponding weights are shown Figure 5.8. It was interesting to observe

that among the top 6 features selected were the number and frequency of asynchrony in breathing, and features of the ribcage (pause power and movement). These 5 features are important with the clinical understanding of the work of breathing.

## 5.5 Discussion

We presented an approach for predicting extubation readiness from automated, novel cardiorespiratory variability features using expressive generative models based on hidden Markov models and non-linear, discriminative classifiers based on random forests. This work extended upon the pilot work [49] published previously by incorporating novel features sensitive to breathing patterns in a larger, multi-institutional population.

Our best classifier combined clinical domain knowledge with a BRF to give a success detection rate of 78% and failure detection rate of 71%. Whereas incorporating clinical decision led to 7% improvement in the balanced classification accuracy for the

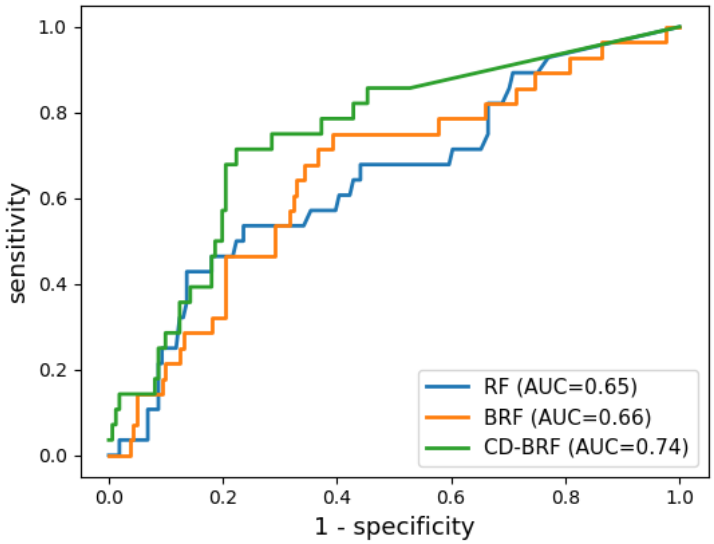


Figure 5.7: Receiver operating characteristic (ROC) curves for random forest (RF), balanced random forest (BRF) and clinical decision with balanced random forest (CD-BRF) classifiers. AUC – area under the curve

random forests, the same was not observed for the hidden Markov model counterparts. As seen in the predicted likelihood scores of the CD-GHMM, this is likely due to class imbalance. In the balanced random forest (BRF), class imbalance was explicitly accounted for by randomly undersampling the majority class in order to train each decision tree with an equal number of success and failure examples. It may also be possible to develop HMM-based predictors which handle class imbalance in a similar way as balanced random forests. This is an area of current research interest for the author.

As with most generative models, HMMs come with major assumptions about how data was generated: independence of observations given hidden states, the representation of densities as Gaussians and the Markov property. These assumptions may be too strong for the underlying generating process of the data. It may be useful to explore other options such as the use of mixtures of Gaussians to model the emission densities. In addition, other graphical models like conditional random fields (CRF) even though theoretically respects the Markov property is very flexible to the incor-

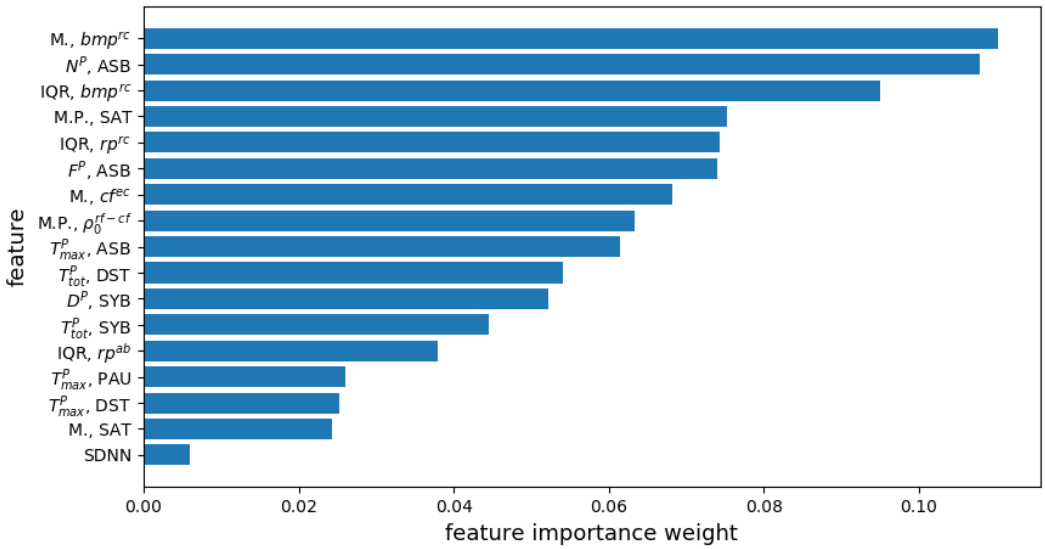


Figure 5.8: Importance weights of features selected by the best CD-BRF classifier. M. is Median, M. P. is Median Power. All other symbols are same as described in section 3.2.3.

poration of features across arbitrary time horizons.

Overall, performance in this work suggests that there were signs in the cardiorespiratory behavior of these infants which, if considered by the physicians, could have prevented 71% extubation failures. Previous work using cardiorespiratory variability features achieved a failure detection rate of 83.2% and success detection rate of 73.6% [49]. This was carried out on a much smaller sample size of 53 babies. The performance observed in current work may be a more realistic measure given the increased heterogeneity in the population. In this work, the best AUC of 0.75 compares with that of [22] which used only clinical variables, and is better than results in [47] which used only respiratory patterns. Overall, this highlights the difficulty of predicting extubation readiness in such high-risk population.

There were a few limitations with this work. Only the 2nd minute of the ETT-CPAP was considered for consistency and to allow direct comparison with previous work. Given that only 17/77 features showed importance, it is crucial to explore new features or longer ETT-CPAP periods. Future work will evaluate how these features differ in the two groups of patients and determine what new, potentially useful features can be developed to boost performance.

It is important to ultimately learn the clinical decision (CD) step from data. This will guarantee the generalisability of the approach to future unseen patients. Also, other clinical variables such as the birth weight and gestational age recorded *at extubation* should be taken into account to improve the stratification in the CD-GHMM and CD-BRF classifiers. Future work will examine these possibilities. Finally, as the number of patients (especially failure cases) was quite small, it will be important to test the models developed here on a held-out validation set. This is part of the data acquisition protocol and will be tested in future work.

# 6

# Conclusion

The goal of this work was to apply machine learning to develop predictors of extubation readiness in extremely preterm infants based on cardiorespiratory behaviour. To this end, we explored several input modalities and developed both generative and discriminative predictive models.

The use of discrete-time Markov and semi-Markov chains (DTMC and DTSC) in chapter 4 were well-suited to the univariate, discrete time-series of respiratory patterns. The DTSC models helped view breathing patterns of preterm infants as sequences of transitions that can be inspected for distinguishing characteristics. These models revealed interesting similarities and differences between infants who succeeded and those who failed extubation. Predictive models built on the DTSC indicated that information in this signal may not be sufficient for accurately predicting extubation readiness. Features inspired by the DTSC were further used discriminative support vector machine (SVM) classifiers. Among the predictive models using the respiratory patterns as input, this gave best results but showed limitation in detecting failure patterns at a high rate, suggesting yet again that more information was needed from other modalities.

In chapter 5, we further designed and developed models for learning from multivariate continuous metrics of cardiorespiratory variability. First, a Gaussian hidden Markov model (GHMM) was designed which assumed 5 hidden states corresponding

to the 5 respiratory patterns. This model performed better than the Markov chains in predicting both success and failure patients. Discriminative classifiers based on random forests (RF) were equally developed for transformed scalar representations of the metrics. We selected RF due to its non-linearity and capability for embedded feature selection. Class imbalance was accounted for via undersampling of majority class in each DT, thus extending the standard RF into balanced random forest (BRF). The BRF gave the best performance on this problem especially in correctly identifying patients that failed extubation. Performance was further boosted when the BRF was focused on the *high risk* population of younger and smaller infants who are of clinical relevance. The best performing classifier was CD-BRF with sensitivity 78% and specificity 71%. It is worth noting that the population of infants in this study represents infants who were deemed ready for extubation. Thus our classifier represents the possibility of preventing the premature extubation of 71% of infants who eventually required reintubation. This would come, though, at a cost to prolonging IMV for some babies who were ready. The actual health and economic costs of this trade-off is yet a matter for clinical debate.

Breathing patterns provide a clinically relevant representation of the respiratory behaviour of preterm infants under IMV. However, results in this work suggest that, used alone, it may have limited applicability as a predictor of extubation readiness. It is possible that significant predictive information is lost in the process of converting the *multivariate, continuous time series* of cardiorespiratory metrics into the *univariate, discrete time series* of breathing patterns. In addition, the clinical validation of AUREA showed that it has an overall agreement score of 0.75 in comparison with the gold standard [52]. Even though this was shown to be better and more repeatable than expert human scorers, it nevertheless indicates that a fair amount of noise is introduced into the low-variance breathing patterns. It is also worth noting that no method was applied to address class imbalance in the Markov chains, leaving open the question as to whether it could have faired better. Given the empirical results

obtained, the author of this thesis recommends a focus on the use of the rich cardiorespiratory metrics to build predictive models of extubation readiness given its success in this work. Such models should necessarily be augmented by clinical variables and measures of breathing patterns about the patients to boost performance and generalisation.

The APEX database has taken almost 5 years to acquire and currently is one of the largest databases of cardiorespiratory signals of preterm infants receiving IMV. Yet, one of the key limitations of this study is the size of the database as it limits the full applicability of machine learning and statistical tools. For this reason, it was not practical to leave out a chunk of the data solely for testing. We applied k-fold stratified cross validation to mitigate bias but further work on separate independent data will be necessary to truly evaluate the generalising ability of the models developed. Indeed, this is part of the APEX data acquisition protocol to obtain 50 patient examples that would *only* be used to validate the models developed on the rest of the acquired data.

Another important issue is the lack of a universal definition of extubation readiness. Across several studies different criteria and time windows (72hr, 5days, 7days) have been used. In part this mirrors the significant practice variations in the care and management of preterm infants across physicians and institutions. It is important to continue to gather clinical and experimental evidence to come up with more evidence-based protocols. This is important not only for effective, standardised development and benchmarking of predictors, but also for to understand and mitigate factors which could cause irreversible long term morbidities and the consequent socioeconomic burdens on patients and families.

As the APEX database continues to grow, it will be necessary to design and develop improved predictors and smart approaches to convert such data into actionable knowledge. This could involve building other kinds of machine learning models. For example, deep neural networks (DNN) have shown strength in the design and formulation of features that humans would not have considered. A viable application of

DNN approach might be design an encoder framework which will leverage the huge untapped 60 minutes worth of IMV to learn new feature extractors that characterise preterm infant breathing. Such encoders when applied on the ETT-CPAP segment of data may extract relevant intermediate features that can then be used in some of the classical machine learning methods explored in this thesis.

In summary, this work has demonstrated empirically the design and development of machine learning approaches to understand cardiorespiratory behaviour of preterm infants and to develop accurate predictors of extubation readiness. It is the hope of the author that this will set pace for more progress on prediction tools for these delicate members of our population.

# 7

# Appendix

## 7.1 Metrics of Predictor Performance

We define metrics used to evaluate predictors of extubation readiness in this work and in the studies reported in the literature (chapter 2). Where $p$ is the number of positive examples (success patients) and $n$ the number of negative examples (failure patients); $tp$, $tn$, $fp$ and $fn$ are respectively the number of true positives, true negatives, false positives and false negatives:

### 7.1.1 Sensitivity

$$Sensitivity = \frac{tp}{tp + fn} = \frac{tp}{p} \tag{7.1}$$

### 7.1.2 Specificity

$$Specificity = \frac{tn}{tn + fp} = \frac{tn}{n} \tag{7.2}$$

### 7.1.3 Positive predictive value (PPV)

$$PPV = \frac{tp}{tp + fp} \tag{7.3}$$

### 7.1.4 Negative predictive value (NPV)

$$NPV = \frac{tn}{tn + fn} \tag{7.4}$$

### 7.1.5 Balanced Classification Accuracy and Misclassification Loss

Accuracy is defined as:

$$\begin{aligned} acc &= \frac{tp + tn}{tp + fn + tn + fp} && (7.5)\\ &= \frac{tp + tn}{p + n} = \frac{tp}{p + n} + \frac{tn}{p + n} && (7.6)\\ &= \frac{tp}{p}\left(\frac{p}{p + n}\right) + \frac{tn}{n}\left(\frac{n}{p + n}\right) && (7.7)\\ &= sensitivity\left(\frac{p}{p + n}\right) + specificity\left(\frac{n}{p + n}\right) && (7.8) \end{aligned}$$

The *balanced classification accuracy* measure, $acc_b$, corresponds to an equal weighting of the sensitivity and specificity measures:

$$acc_b = sensitivity * 0.5 + specificity * 0.5 \tag{7.9}$$

The *balanced misclassification loss* is:

$$loss_b = 1 - acc_b \tag{7.10}$$

## 7.2 Probability Distributions for Dwell Time

List of probability distributions considered when fitting sojourn time [61]

- Beta
- Birnbaum-Saunders

- Exponential
- Extreme value
- Gamma
- Generalized extreme value
- Generalized Pareto
- Inverse Gaussian
- Logistic
- Log-logistic
- Lognormal
- Nakagami
- Normal
- Rayleigh
- Rician
- t location-scale
- Weibull

## 7.3 Symmetric KL Divergence

The Kullback-Leibler (KL) divergence [33] is a measure of how well a distribution $Q$ is approximating another distribution $P$. It is defined as:

$$D_{KL}(P||Q) = \sum_n P_n \log \frac{P_n}{Q_n} \tag{7.11}$$

The KL-divergence is non-symmetric: $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, which is not desirable in our application. Hence, we use symmetrized KL-divergence to compare distributions over transitions between patterns:

$$D_{KLS}(P||Q) = D_{KL}(P||Q) + D_{KL}(Q||P) \tag{7.12}$$

$$D_{KLS}(P||Q) = \sum_{n}(P_n - Q_n)\log\frac{P_n}{Q_n} \tag{7.13}$$

## 7.4 Ranges of Hyper-Parameters for Random Forests

Table 7.1 shows the list of hyperparameters and the respective ranges searched for the RF, BRF and CD-BRF classifiers.

Table 7.1: Hyperparameters and Ranges search for random forests. Range format: *[start value:increment:end value]*

| **Hyperparameter** | **Range** |
|---|---|
| Number of estimators (decision trees) | [1:1:30] |
| Fraction of features to consider when looking for best split | [0.1:0.1:1] |
| Minimum fraction of samples required at leaf | [0.1:0.05:0.5] |
| Maximum depth of tree | [1:1:15] |

## 7.5 Learning curves for GHMM and CD-GHMM

Below are learning curves for all of the GHMMs and CD-GHMMs given different covariance matrix types.

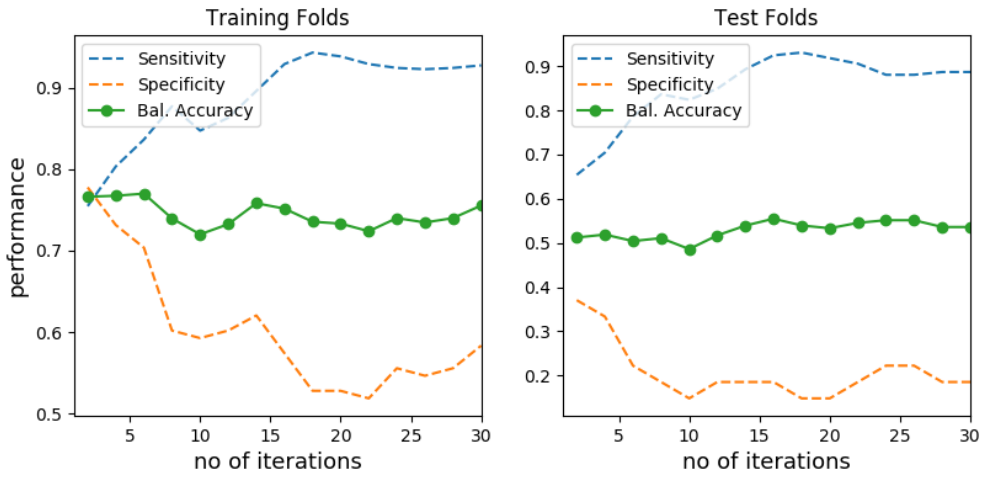


Figure 7.1: Learning curves for best GHMM (diagonal covariance)

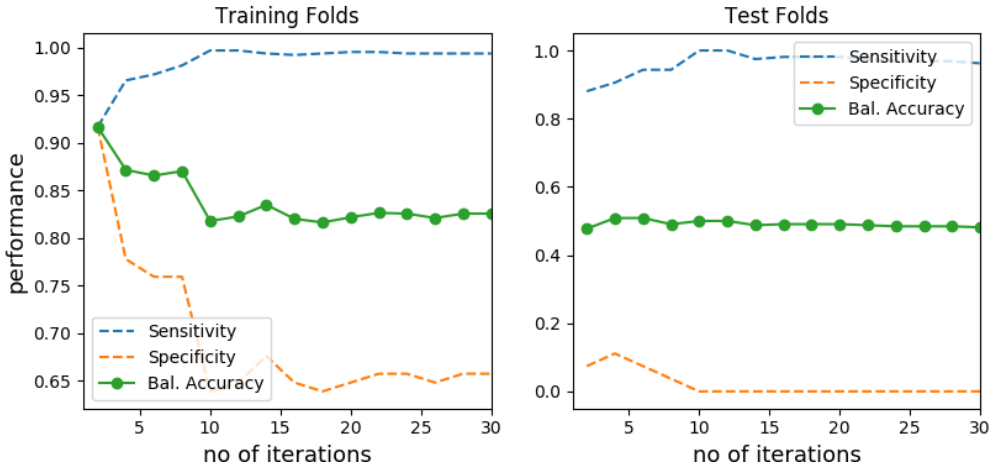


Figure 7.2: Learning curves for best GHMM (full covariance)

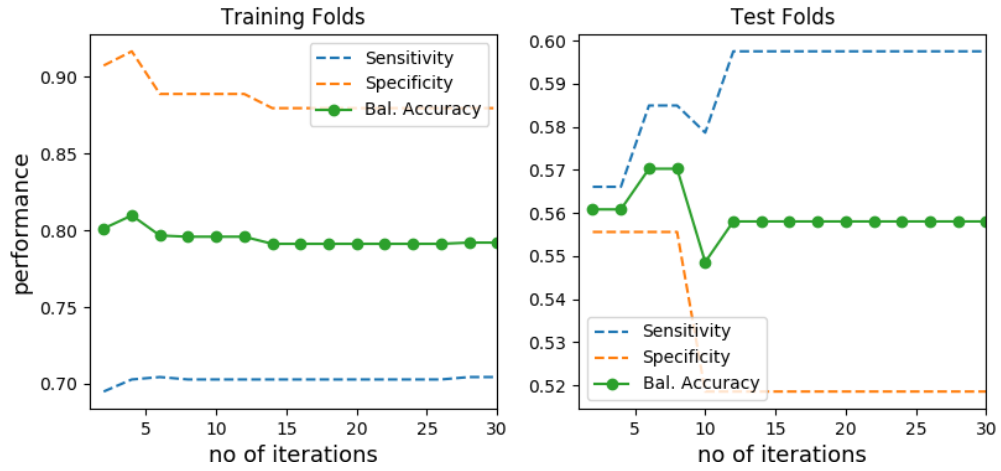


Figure 7.3: Learning curves for best GHMM (tied covariance)

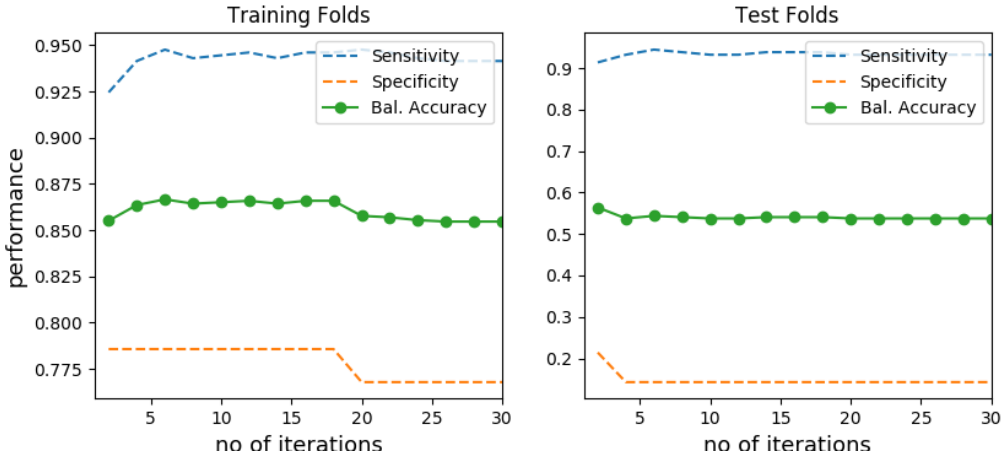


Figure 7.4: Learning curves for best CD-GHMM (spherical covariance)

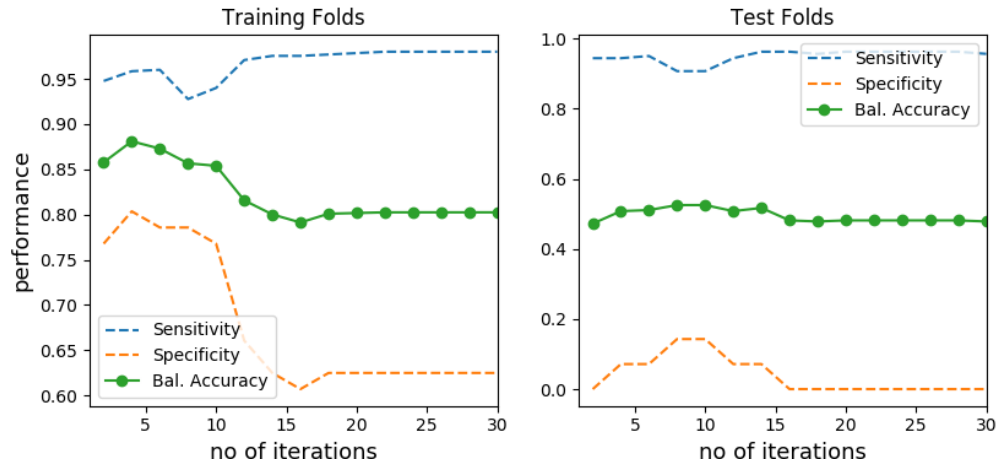


Figure 7.5: Learning curves for best CD-GHMM (diagonal covariance)

## 7.6 Distribution of Likelihood Scores for GHMM and CD-GHMM

The figures below show the distribution of likelihood scores for the all of the GHMMs and for the rest of the CD-GHHMs (excluding tied covariance shown earlier).

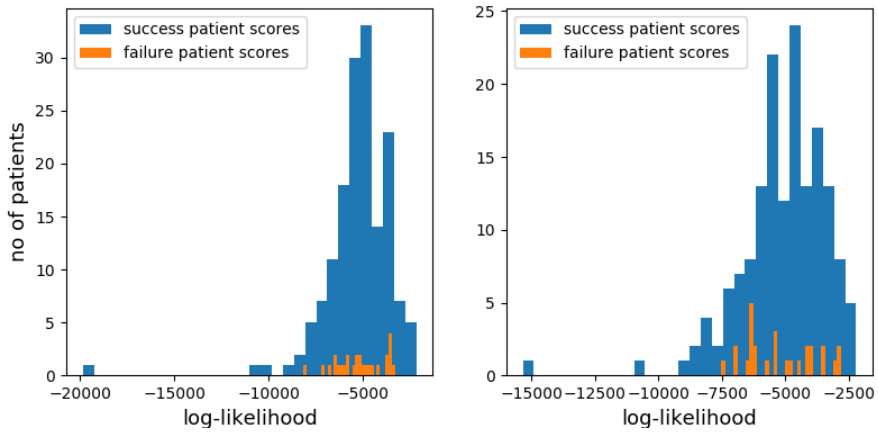


Figure 7.6: Distribution of likelihood scores of success and failure patients in the test folds considering the Success (**Left**) and Failure (**Right**) models of the GHMM (spherical covariance).

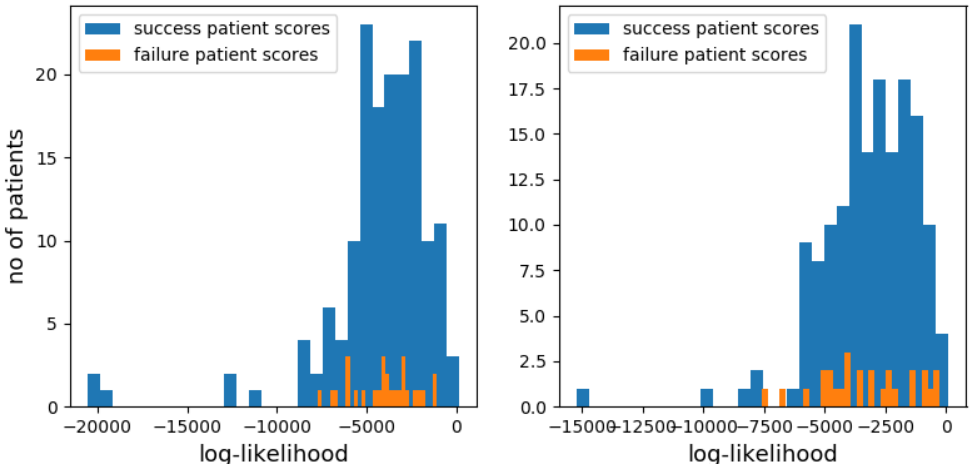


Figure 7.7: Distribution of likelihood scores of success and failure patients in the test folds considering the Success (**Left**) and Failure (**Right**) models of GHMM (diagonal covariance).

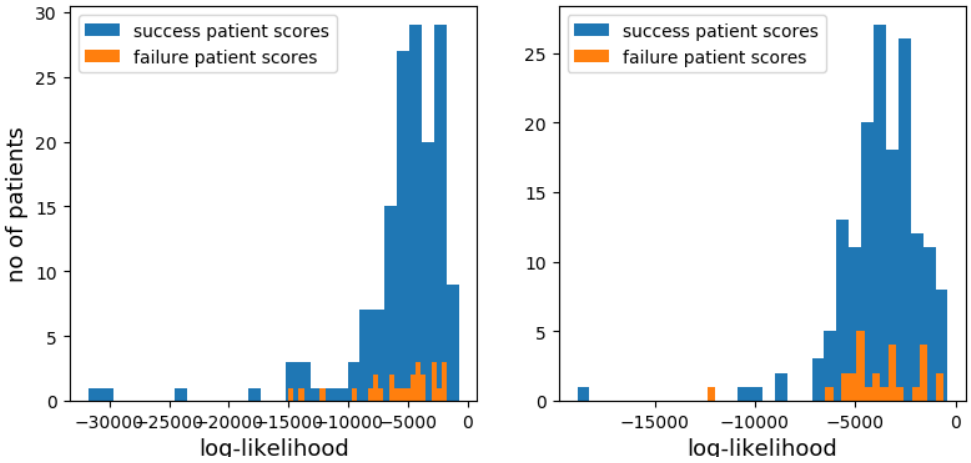


Figure 7.8: Distribution of likelihood scores of success and failure patients in the test folds considering the Success (**Left**) and Failure (**Right**) models of GHMM (full covariance).

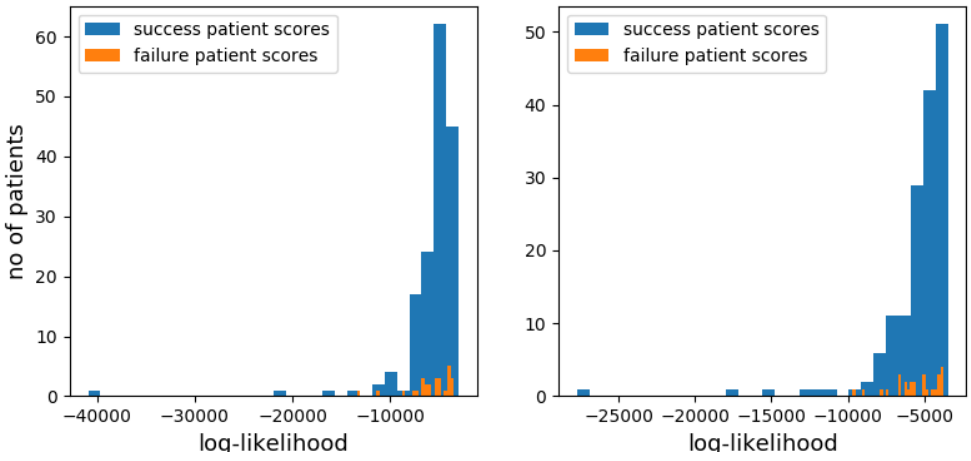


Figure 7.9: Distribution of likelihood scores of success and failure patients in the test folds considering the Success (**Left**) and Failure (**Right**) models of GHMM (tied covariance).

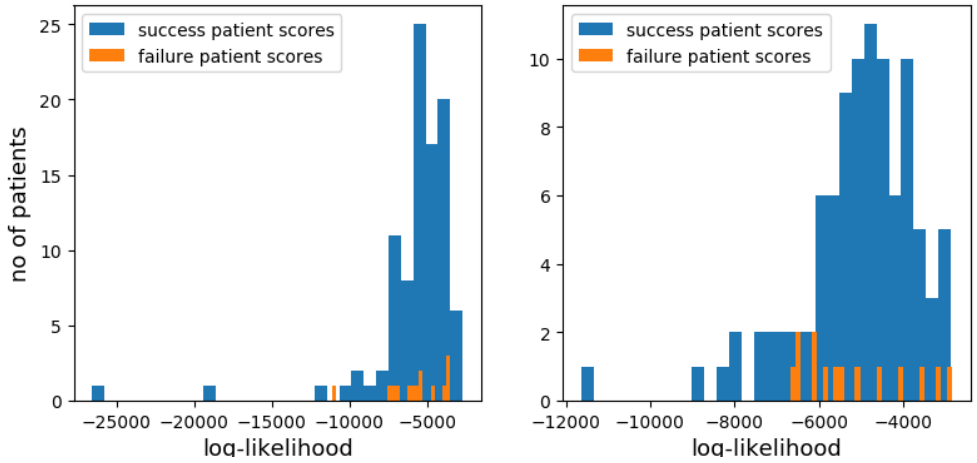


Figure 7.10: Distribution of likelihood scores of success and failure patients in the test folds considering the Success (**Left**) and Failure (**Right**) models of CD-GHMM (spherical covariance).

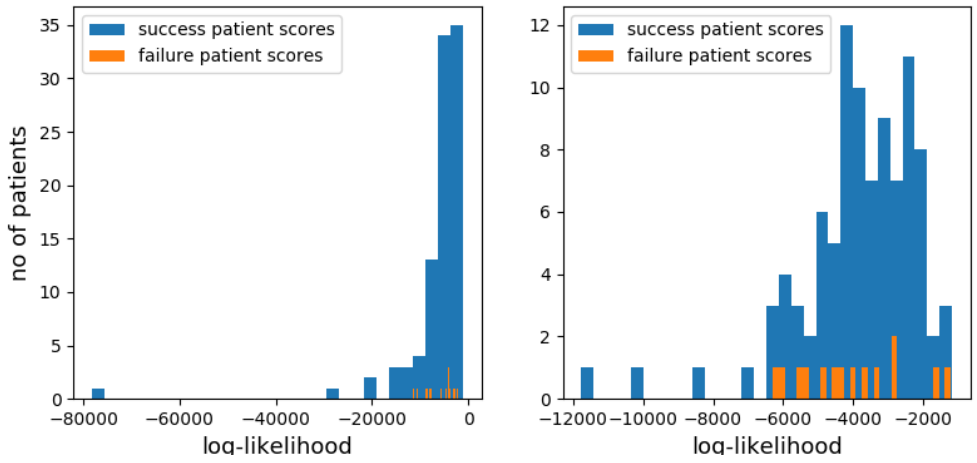


Figure 7.11: Distribution of likelihood scores of success and failure patients in the test folds considering the Success (**Left**) and Failure (**Right**) models of CD-GHMM (diagonal covariance).

# Bibliography


[1] Davide Alinovi et al. "Markov chain modeling and simulation of breathing patterns". In: *Biomedical Signal Processing and Control* 33 (2017), pp. 245 –254. ISSN: 1746-8094. DOI: http://doi.org/10.1016/j.bspc.2016.12.002.

[2] Steven D Baisch et al. "Extubation failure in pediatric intensive care incidence and outcomes". In: *Pediatric Critical Care Medicine* 6.3 (2005), pp. 312–318.

[3] James Baker. "The DRAGON system–An overview". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 23.1 (1975), pp. 24–29.

[4] Janet M Baker et al. "Developments and directions in speech recognition and understanding, Part 1 [DSP Education]". In: *IEEE Signal Processing Magazine* 26.3 (2009).

[5] Vlad Barbu and Nikolaos Limnios. *Semi-Markov Chains and Hidden Semi-Markov Models Toward Applications: Their Use in Reliability and DNA Analysis*. 1st ed. Springer Publishing Company, Incorporated, 2008. ISBN: 0387731717, 9780387731711.

[6] Leonard E Baum et al. "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains". In: *The annals of mathematical statistics* 41.1 (1970), pp. 164–171.

[7] Yoshua Bengio. "Markovian models for sequential data". In: *Neural computing surveys* 2.199 (1999), pp. 129–162.

[8] Mauo-Ying Bien et al. "Comparisons of predictive performance of breathing pattern variability measured during T-piece, automatic tube compensation, and pressure support ventilation for weaning intensive care unit patients from mechanical ventilation". In: *Critical care medicine* 39.10 (2011), pp. 2253–2262.

[9] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[10] Z Bismilla et al. "Failure of pediatric and neonatal trainees to meet Canadian Neonatal Resuscitation Program standards for neonatal intubation". In: *Journal of Perinatology* 30.3 (2010), p. 182.

[11] Leo Breiman. "Bagging predictors". In: *Machine learning* 24.2 (1996), pp. 123–140.

[12] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[13] Sanjay Chawla et al. "Markers of successful extubation in extremely preterm infants, and morbidity after failed extubation". In: *The Journal of pediatrics* 189 (2017), pp. 113–119.

[14] Sanjay Chawla et al. "Role of spontaneous breathing trial in predicting successful extubation in premature infants". In: *Pediatric pulmonology* 48.5 (2013), pp. 443–448.

[15] Chao Chen, Andy Liaw, and Leo Breiman. "Using random forest to learn imbalanced data". In: *University of California, Berkeley* 110 (2004), pp. 1–12.

[16] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297.

[17] Peter G Davis and David J Henderson-Smart. "Extubation from low-rate intermittent positive airway pressure versus extubation after a trial of endotracheal continuous positive airway pressure in intubated preterm infants". In: *the Cochrane library* (2001).

[18] Adina E Draghici and J Andrew Taylor. "The physiological basis and measurement of heart rate variability in humans". In: *Journal of physiological anthropology* 35.1 (2016), p. 22.

[19] Scott K Epstein and Ronald L Ciubotaru. "Independent effects of etiology of failure and time to reintubation on outcome for patients failing extubation". In: *American journal of respiratory and critical care medicine* 158.2 (1998), pp. 489–493.

[20] Madalina Fiterau et al. "ShortFuse: biomedical time series representations in the presence of structured information". In: *arXiv preprint arXiv:1705.04790* (2017).

[21] Annie Giaccone et al. "Definitions of extubation success in very premature infants: a systematic review". In: *Archives of Disease in Childhood-Fetal and Neonatal Edition* 99.2 (2014), F124–F127.

[22] Pascale Gourdeau et al. "Feature selection and oversampling in analysis of clinical data for extubation readiness in extreme preterm infants". In: *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE. 2015, pp. 4427–4430.

[23] Fernanda Hermeto et al. "Incidence and main risk factors associated with extubation failure in newborns with birth weight< 1,250 grams". In: *Jornal de pediatria* 85.5 (2009), pp. 397–402.

[24] Geoffrey Hinton et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 82–97.

[25] Frederick Jelinek, Lalit Bahl, and Robert Mercer. "Design of a linguistic statistical decoder for the recognition of continuous speech". In: *IEEE Transactions on Information Theory* 21.3 (1975), pp. 250–256.

[26] Bing-Hwang Juang, Stephene Levinson, and M Sondhi. "Maximum likelihood estimation for multivariate mixture observations of markov chains (corresp.)" In: *IEEE Transactions on Information Theory* 32.2 (1986), pp. 307–309.

[27] Jennifer Kaczmarek et al. "Heart rate variability and extubation readiness in extremely preterm infants". In: *Neonatology* 104.1 (2013), pp. 42–48.

[28] Jennifer Kaczmarek et al. "Variability of respiratory parameters and extubation readiness in ventilated neonates". In: *Archives of Disease in Childhood-Fetal and Neonatal Edition* (2013), fetalneonatal–2011.

[29] Amir Kale et al. "Gait analysis for human identification". In: *International Conference on Audio-and Video-Based Biometric Person Authentication*. Springer. 2003, pp. 706–714.

[30] C O F Kamlin, P G Davis, and C J Morley. "Predicting successful extubation of very low birthweight infants". In: *Archives of Disease in Childhood - Fetal and Neonatal Edition* 91.3 (2006), F180–F183. ISSN: 1359-2998. DOI: 10.1136/adc.2005.081083. eprint: http://fn.bmj.com/content/91/3/F180.full.pdf. URL: http://fn.bmj.com/content/91/3/F180.

[31] C Omar Farouk Kamlin et al. "A trial of spontaneous breathing to determine the readiness for extubation in very low birth weight infants: a prospective evaluation". In: *Archives of Disease in Childhood-Fetal and Neonatal Edition* 93.4 (2008), F305–F306.

[32] Lara J Kanbar et al. "Undersampling and Bagging of Decision Trees in the Analysis of Cardiorespiratory Behavior for the Prediction of Extubation Readiness in Extremely Preterm Infants". In: *Engineering in Medicine and Biology Society (EMBC), 2018 40th Annual International Conference of the IEEE*. IEEE. 2018.

[33] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[34] John Lafferty, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In: (2001).

[35] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444.

[36] L Liporace. "Maximum likelihood estimation for multivariate observations of Markov sources". In: *IEEE Transactions on Information Theory* 28.5 (1982), pp. 729–734.

[37] Zachary C Lipton, John Berkowitz, and Charles Elkan. "A critical review of recurrent neural networks for sequence learning". In: *arXiv preprint arXiv:1506.00019* (2015).

[38] Zachary C Lipton et al. "Learning to diagnose with LSTM recurrent neural networks". In: *arXiv preprint arXiv:1511.03677* (2015).

[39] Andreĭ Markov. "The theory of algorithms". In: ().

[40] MATLAB. *version 9.0.0 (R2016a)*. Natick, Massachusetts: The MathWorks Inc., 2016.

[41] Andrew McCallum, Dayne Freitag, and Fernando CN Pereira. "Maximum Entropy Markov Models for Information Extraction and Segmentation." In: *Icml*. Vol. 17. 2000. 2000, pp. 591–598.

[42] J Davin Miller and Waldemar A Carlo. "Pulmonary complications of mechanical ventilation in neonates". In: *Clinics in perinatology* 35.1 (2008), pp. 273–281.

[43] Takayuki Mukaeda and Keisuke Shima. "A novel hidden Markov model-based pattern discrimination method with the anomaly detection for EMG signals". In: *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*. IEEE. 2017, pp. 921–924.

[44] Lamana Mulaffer et al. "Comparing two insomnia detection models of clinical diagnosis techniques". In: *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*. IEEE. 2017, pp. 3749–3752.

[45] Charles C Onu. "Harnessing infant cry for swift, cost-effective diagnosis of perinatal asphyxia in low-resource settings". In: *Humanitarian Technology Conference-(IHTC), 2014 IEEE Canada International*. IEEE. 2014, pp. 1–4.

[46] Charles C Onu et al. "A semi-Markov chain approach to modeling respiratory patterns prior to extubation in preterm infants". In: *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*. IEEE. 2017, pp. 2022–2026.

[47] Charles C Onu et al. "Predicting extubation readiness in extreme preterm infants based on patterns of breathing". In: *Computational Intelligence (SSCI), 2017 IEEE Symposium Series on*. IEEE. 2017, pp. 1–7.

[48] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[49] D. Precup et al. "Prediction of extubation readiness in extreme preterm infants based on measures of cardiorespiratory variability". In: *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2012, pp. 5630–5633. DOI: 10.1109/EMBC.2012.6347271.

[50] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 1994.

[51] Lawrence R Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Readings in speech recognition*. Elsevier, 1990, pp. 267–296.

[52] Carlos A Robles-Rubio, Karen A Brown, and Robert E Kearney. "Automated unsupervised respiratory event analysis". In: *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE. 2011, pp. 3201–3204.

[53] Carlos A Robles-Rubio et al. "Automated analysis of respiratory behavior for the prediction of apnea in infants following general anesthesia". In: *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE. 2014, pp. 262–265.

[54] Carlos Alejandro Robles-Rubio et al. "Scoring Tools for the Analysis of Infant Respiratory Inductive Plethysmography Signals". In: *PLOS ONE* 10.7 (July 2015), pp. 1–31. DOI: 10.1371/journal.pone.0134182. URL: http://dx.doi.org/10.1371%2Fjournal.pone.0134182.

[55] Guido Rossum. *Python Reference Manual*. Tech. rep. Amsterdam, The Netherlands, The Netherlands, 1995.

[56] Guilherme Mendes Sant'Anna and Martin Keszler. "Weaning infants from mechanical ventilation". In: *Clinics in perinatology* 39.3 (2012), pp. 543–562.

[57] Michael H Schwartz et al. "Predicting the outcome of intramuscular psoas lengthening in children with cerebral palsy using preoperative gait data and the random forest algorithm". In: *Gait & posture* 37.4 (2013), pp. 473–479.

[58] Gideon Schwarz et al. "Estimating the dimension of a model". In: *The annals of statistics* 6.2 (1978), pp. 461–464.

[59] Wissam Shalish et al. "Prediction of Extubation readiness in extremely preterm infants by the automated analysis of cardiorespiratory behavior: study protocol". In: *BMC pediatrics* 17.1 (2017), p. 167.

[60] Hsiu-Nien Shen et al. "Changes of heart rate variability during ventilator weaning". In: *Chest* 123.4 (2003), pp. 1222–1228.

[61] Mike Sheppard. “Fit all valid parametric probability distributions to data”. In: *MATLAB Central File Exchange, Retrieved January 18, 2017* (2012).

[62] Barbara J Stoll et al. “Trends in care practices, morbidity, and mortality of extremely preterm neonates, 1993-2012”. In: *Jama* 314.10 (2015), pp. 1039–1051.

[63] Michele C Walsh et al. “Extremely low birthweight neonates with protracted ventilation: mortality and 18-month neurodevelopmental outcomes”. In: *The Journal of pediatrics* 146.6 (2005), pp. 798–804.

[64] Xingchen Wang et al. “Monitoring of gait performance using dynamic time warping on IMU-sensor data”. In: *Medical Measurements and Applications (MeMeA), 2016 IEEE International Symposium on*. IEEE. 2016, pp. 1–6.

[65] Steve Young. “A review of large-vocabulary continuous-speech”. In: *IEEE signal processing magazine* 13.5 (1996), p. 45.