Data Fusion Models for Detection of Vital-Sign Deterioration in Acutely Ill Patients



Sara Khalid

Oriel College

University of Oxford

Supervised by Professor Lionel Tarassenko and Dr. David Clifton

Submitted: Trinity Term, 2013

This thesis is submitted to the Department of Engineering Science in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Data Fusion Models for Detection of Vital-Sign Deterioration in Acutely Ill Patients

Sara Khalid

Oriel College, University of Oxford

Doctor of Philosophy

Hillary Term, 2013

ABSTRACT

Vital signs can indicate patient deterioration prior to adverse events such as cardiac arrest, emergency admission to the intensive care unit (ICU), or death. However, many adverse events occur in wards outside the ICU where the level of care and the frequency of patient monitoring are lower than in the ICU. This thesis describes models for detection of deterioration in acutely ill patients in two environments: a step-down unit in which patients recovering from an ICU stay are continuously monitored, and a general ward where patients are intermittently monitored following upper gastrointestinal cancer surgery.

Existing data fusion models for classification of vital signs depend on a threshold which defines a "region of normality". Bradypnoea (low breathing rate) and bradycardia (low heart rate) are relatively rare, and so these two types of abnormalities tend to be misclassified by existing methods. In this thesis, techniques for selecting a threshold are described, such that the classification of vital-sign data is improved. In particular, the proposed approach reduces the misclassification of bradycardia and bradypnoea events, and indicates the type of abnormality associated with the deterioration in a patient's vital signs.

Patients recovering from upper gastrointestinal (GI) surgery have a high risk of emergency admission to the ICU. At present in the UK, most intermediate and general wards outside the ICU depend on intermittent, manual monitoring using track-and-trigger systems. Both manual and automated patient monitoring systems are reported to have high false alert rates. The models described in this thesis take into account the low monitoring frequency in the upper GI ward, such that the false alert rate is reduced.

In addition to accuracy, early detection of deterioration is a highly desirable feature in patient monitoring systems. The models proposed in this thesis generate alerts for patients earlier than the early warning systems which are currently in use in hospitals in the UK. The improvements to existing models proposed in this thesis could be applied to continuous and intermittently acquired vital-sign data from other clinical environments.

Dedication

To Him, Who is the Light of the heavens and the earth, and Who wills it to be, and it is.

And to Amma, Aboo, Junaid, and our eight month old Imaan.

Acknowledgements

There are many people I must thank for helping me to reach the finish line. Foremost, Professor Lionel Tarassenko, who has supported me in matters academic and personal, throughout the roller-coaster ride that these past four years have been. It has been a privilege to have worked under his supervision, and to learn from him to view the grand(er) scheme of things. "Standing on the shoulders of giants" now has a new meaning for me. Thank you also to Dr. David Clifton for the many "exploratory" walks undertaken during the course of my research, which often led us to Bayesian destinations. Thanks is due to both supervisors for patiently sculpting the essayist in me into a scientific writer.

I would like to thank research nurses Debbie Evans, Sarah Vollam, and Breda Lynch for assisting with data collection, and Dr. Lei Clifton for her meticulous organisation of the data used in this thesis. Thanks is due also to my colleagues in the multi-continental BSP group. I am grateful to Christina Orphanidou, Dave Wong, and Oliver Gibson, and Carmelo Velardo, for being valuable sources of knowledge and tea bags. I am particularly grateful to Dr. Hammer for providing encouragement throughout my research, as well the occasional opportunity to revise my Urdu; and to the three CDT musketeers, Lise, Marco, and Mauro, for proof-reading this thesis.

I am grateful to the Rhodes Trust and to the Qualcomm Trust for providing financial support during the course of my DPhil. A special thank you to Mary Eaton at the Rhodes House for being a shoulder to unburden on, term in, term out.

I must mention my brothers Saad, Bilal, and Talha, for their insane jokes, sporting prowess (or lack thereof), and for keeping alive my internal competitive monster. And Nano for sharing with me her wisdom and embroidery patterns during many a long, sweltering Karachi summers.

There are two people to whom my gratitude, now and forever, cannot justly be summed up in words. Amma and Aboo – thank you, for bringing me this far.

Finally, I have the very great pleasure of thanking Junaid Hassan, who has been my rock these past four years, and whose life, in addition to my own and that of our little Imaan, has been revolving around my adventures in Oxford.

To all the above, and to my friends in Oriel College, among fellow Rhodes Scholars, and from Oxford-at-large, thank you for making my education in this city of dreaming spires much more than an purely scholastic one.

Related Publications

Khalid, S., Clifton, D.A., and Tarassenko, L.: A Bayesian Patient-Based Model for Detecting Deterioration in Vital Signs Using Manual Observations. *Foundations of Health Information Engineering and Systems. Lecture Notes in Computer Science*, 2013, pp 146-158.

Khalid, S., Clifton, D.A., Clifton, L.A., and Tarassenko, L.: A Two-Class Approach to the Detection of Physiological Deterioration in Patient Vital Signs, with Clinical Label Refinement. *IEEE Transactions on Information Technology in Biomedicine.* 16(6), 2012, pp. 1231-1238.

Clifton, D.A., Clifton, L.A., Alvi, M., Khalid, S., Meredith, D., Price, J., Watkinson, P., and Tarassenko, L.: Towards Assisted Living via Probabilistic Vital-Sign Monitoring in the Home. *IET Assisted Living*, London, UK, 2011, pp. 9-14.

Khalid, S., Clifton, D.A., Clifton, L.A., and Tarassenko, L.: Optimising Classifiers for the Detection of Physiological Deterioration in Patient Vital-Sign Data. *Biomedical Engineering Systems and Technologies*, Rome, Italy, 2011, pp. 425-428.

Khalid, S., Clifton, D.A., and Tarassenko, L.: Modelling Patient Vital-Sign Deterioration Trajectories Using Bayesian Inference. *IEEE PGBIOMED*, Glasgow, UK, 2011, pp. 1-2.

Contents

1 A Review of Patient Monitoring Systems **1**
- 1.1 Introduction . . . 1
- 1.2 Patient Deterioration Outside the ICU . . . 4
- 1.3 Interventions . . . 17
- 1.4 Impact of EWS-based Interventions . . . 22
- 1.5 Thesis Overview . . . 31

2 Automated Detection of Deterioration using Non-invasive Vital-Sign Monitoring **34**
- 2.1 Non-invasive Monitoring of Vital Signs . . . 34
- 2.2 Automatic Alarm Generation . . . 38
- 2.3 Data Fusion . . . 41
- 2.4 Discussion . . . 47

3 Constructing a Model of Normality **49**
- 3.1 Introduction . . . 49
- 3.2 The UPMC Study . . . 50
- 3.3 Labelling of Vital-Sign Data . . . 51
- 3.4 Labelling of UPMC Vital-Sign Data . . . 52
- 3.5 Constructing a Density-based Model of Normality . . . 53
- 3.6 Discussion . . . 60

4 One-Class Classification **61**
- 4.1 Introduction . . . 61
- 4.2 Calculating the Probability of Novelty . . . 62
- 4.3 Optimisation of T Based on Data Distribution . . . 65
- 4.4 Optimal Value of T for Identifying Abnormal Data . . . 71
- 4.5 Optimal Value of T for Minimising Classification Error . . . 73
- 4.6 Discussion and Conclusions . . . 84

5 Two-Class Classification **86**
- 5.1 Introduction . . . 86
- 5.2 Standard Two-class Classification . . . 87
- 5.3 Hybrid Two-class Classification . . . 87
- 5.4 Logistic Regression Model . . . 89
- 5.5 Multi-Layer Perceptron Model . . . 91
- 5.6 Support Vector Machines . . . 97
- 5.7 Application to 3-D and 4-D Datasets . . . 100
- 5.8 Results . . . 103
- 5.9 Conclusions . . . 110

6 Constructing Models of Abnormality **111**
6.1 Introduction . . . 111
6.2 Description of "Abnormal" Clusters . . . 112
6.3 Constructing Models of Abnormality . . . 117
6.4 Results . . . 119
6.5 Discussion . . . 121
6.6 Conclusions . . . 124

7 Deterioration in Patients on Upper GI Ward **129**
7.1 The Cancer Hospital Project . . . 130
7.2 The $\mathcal{CH}_2$ Dataset . . . 132
7.3 Comparison of Datasets $\mathcal{D}_1$ and $\mathcal{CH}_2$ . . . 136
7.4 Time-varying Trends . . . 145
7.5 Discussion . . . 153

8 Models for Detecting Deterioration on the Upper GI Ward **156**
8.1 Introduction . . . 156
8.2 Assessing Model Performance . . . 157
8.3 Datasets . . . 158
8.4 Parzen Windows Model . . . 160
8.5 Motivation for Patient-based Model of Normality . . . 164
8.6 Patient-Based Modelling . . . 166
8.7 A Model for Multi-dimensional Data . . . 177
8.8 EWS Models . . . 186
8.9 Overall Results . . . 188
8.10 Discussion and Conclusions . . . 190

9 Conclusion **194**
9.1 Summary of Results . . . 194
9.2 Future Work . . . 197

A Normal-Gamma Conjugate Distributions for Uni-dimensional Data **202**
A.1 Conjugate Prior and Posterior . . . 202
A.2 Hyper-Parameters . . . 203

B Normal-Wishart Conjugate Distributions for Multi-dimensional Data **205**
B.1 Conjugate Prior and Posterior . . . 205
B.2 Multi-dimensional Extension . . . 206

C Last Day of CALMS-2 EICU Patients **209**

Bibliography **209**

List of Symbols

Symbols	Description
C_l	the l-th class
$P(.)$	probability
y	class labels
$p(.)$	probability density function
$\mathcal{D}$	dataset, denoted by calligraphic font
$\boldsymbol{\Theta}$	multi-dimensional parameter set
θ	uni-dimensional parameter set
H_i	i-th hypothesis
$\mathbb{R}$	the set of real numbers
$\mathcal{N}(.)$	Normal (Gaussian) function
$\mathcal{NG}(.)$	Normal-Gamma function
$\mathcal{NW}(.)$	Normal-Wishart function
d	number of vital-sign dimensions in a dataset

List of Acronyms

BMN	Bayesian model of normality
BR	Breathing rate
CCO	Critical care outreach
DBP	Diastolic blood pressure
EICU	Emergency intensive care unit
EWS	Early warning score
GI	Gastrointestinal
HDU	High dependency unit
HR	Heart rate
ICU	Intensive care unit
MET	Medical emergency team
NR	Normal recovery
PW	Parzen windows
SBP	Systolic blood pressure
SDA	Systolic diastolic average
SDU	Step-down unit
SpO_2	Blood oxygen saturation
T**MP**	Temperature

Chapter 1

A Review of Patient Monitoring Systems

1.1 Introduction

Each year, 850,000 adverse events occur in hospitals in the UK [49] including 23,000 cardiac arrests [73] and 20,000 unplanned intensive care unit (ICU) admissions [150] that may be avoided with better care [104, 105, 44]. The survival-to-discharge rate after an in-hospital cardiac arrest in the UK is between 17-24% [65, 115]. In the USA, more than 150,000 cardiac arrests occur every year with a mortality rate of over 80% [124], and 100,000 deaths per year are considered avoidable [4]. According to the 2012 National Confidential Enquiry into Patient Outcome and Death (NCEPOD), it is now known that in 75% of the cases, cardiac arrest is preceded by premonitory signs of deterioration, and that over a third (38%) of these could be prevented if the premonitory signs were recognised and responded to in time [80].

According to a recent enquiry of all National Health Service (NHS), public and private hospitals in England, Wales and Northern Ireland, most cardiac arrests are predictable events preceded by period of progressive physiological deterioration, which is frequently poorly recognised and responded to [80]. Over 50% of these occur on lower-acuity general wards outside the ICU [31, 104]. It was found that in 7 out of 10 cases of cardiac arrests occurring on the general wards outside ICU, care was sub-optimal [80]. Poor care was ascribed to lack of (a) recognition of the premonitory physiological indicators, (b) appreciation of the urgency, and (c) proper and timely response by escalation of the case to senior doctors, among other factors. In another study [104] it was reported that

66% of unexpected deaths occurred on the general wards. This confirms the findings of earlier studies [7, 57, 146, 104, 105, 83, 73, 133, 71, 19] that the incidence of extreme adverse events including cardiac arrest, emergency ICU admission and unexpected death occurring on general wards outside the ICU is correlated with the level of care in wards outside the ICU.

It has been recognised that deterioration in patient condition is often preceded by physiological deterioration in one or more of the patient's vital signs[1], and that adverse events can be prevented by detecting this deterioration early [105, 104, 57, 144, 19]. As early as in the 1990s it was shown that abnormal changes in breathing rate, for example, are a typical precursor of cardiac arrest events (present in 38% of the patients) [140, 55, 7] . Several studies in the 1990s reported abnormalities in the respiration rate, heart rate, blood pressure and arterial oxygen saturation in patients prior to an adverse event [146, 57, 105, 140, 73, 113, 19]. The first large multi-national study [83] to confirm in 2004 the presence of physiological abnormality prior to adverse events such as cardiac arrest, unanticipated ICU admissions, and death was conducted across 90 hospitals from the UK, New Zealand, and Australia. It reported that during the 3-day study period, 60% of patients with an adverse event experienced deterioration in physiological vital signs within 24 hours of the event. This included 79.4% of the cardiac arrest cases, 54.5% of the unanticipated ICU cases, and 11.7% of the death cases. Other studies have also reported that premonitory indication of deterioration in vital signs can occur in the 24 hours before an adverse event such as an emergency ICU admission, cardiac arrest, and unexpected death [64, 6, 147]. Some research has reported that deterioration can be detected as early as 6-8 hours in advance [140, 19, 115, 71].

In 2011-2012, nearly 35% of all hospital admissions in the UK were categorised as emergency admissions[16]. A recent systematic review [169] of 27 studies on the incidence of adverse events prior to ICU admissions reported that up to 37.2% adverse events led to an ICU admission. Up to 18.3 % of adverse events ended up having re-admission to the ICU [169, 132]. The incidence of emergency admissions to the Intensive Care Unit (ICU) of hospitals has increased sharply over the last few decades [137]. From 1984 to 1997,

[1]The heart rate, breathing rate, oxygen saturation, temperature, and blood pressure are called the vital signs of a patient, and they can be measured non-invasively.

a 50% rise in the number of patients requiring emergency admission was reported [15]. Furthermore, unanticipated or emergency ICU admissions have a 50% mortality rate, which is significantly higher when compared with the 35% mortality rate of scheduled admissions [169, 104, 132], and a prolonged length-of-stay (LOS) [169]. If these adverse events are detected sufficiently early, intervention that could help reduce mortality rates and hospital LOS, as well as lessen the burden on ICUs [169], would be possible.

However, timely detection of adverse events is not always possible because most of these events occur outside the ICU on the general ward where patient vital signs are typically monitored by clinical staff every 8-12 hours [145, 44], with each nurse attending several patients. This often causes premonitory deviations in vital signs to go unnoticed [31, 73, 144]. A recent study [75] on 326 in-hospital patients in a United States (US) hospital showed that only 7 out of 59 clinically adverse events were detected by clinical staff. Additionally, an adverse event may not manifest itself through a deviation in any one particular vital sign, and it may instead be due to subtle, less obvious deviations in several vital signs. Furthermore, studies [57, 7] show that in cases where deteriorating trends in vital signs have been identified, they failed to elicit the necessary response from the clinical staff, possibly due to a lack of clearly defined response procedures.

To summarize, there continues to be a high incidence of high-mortality adverse events such as cardiac arrest, emergency ICU admissions, and unexpected death in hospitals in wards outside the ICU. Many of these events are preceded by early warning signs of physiological deterioration. With better recognition of these premonitory signs through vital-sign monitoring and adequate and timely response, these events may be predicted and prevented. Monitoring and care of patients on wards outside the ICU is not optimal. There is a need for patient monitoring systems that can assist clinical decision-making by detecting deterioration in time and improve patient outcomes.

In this chapter a review of the current state-of-the-art of patient monitoring systems is presented. Current interventions for the detection of deterioration in vital signs are discussed. Traditional and automated systems are described, and finally an overview of this thesis is presented.

	General Ward	Intermediate Ward	Intensive Care Unit
Level of care	1	2	3
Nurse:patient ratio	$1:5-10$	1:3-5	1:1-2
Monitoring frequency	>4* hours	2-4 hours	Continuous monitoring

$\longrightarrow$Increasing level of care

Table 1.1: Typical hierarchy of level of care in different wards in the hospital. In the UK, the intermediate ward is also known as the high-dependency unit (HDU). In the USA, it is called the step-down unit (SDU). * The monitoring frequency in general wards differs in hospitals. For instance, in the Oxford University Hospitals, UK, patients in the general wards are monitored every $8-12$ hours, and the frequency is increased if there is any concern.

1.2 Patient Deterioration Outside the ICU

1.2.1 Extreme Adverse Events

An adverse event is an instance of deterioration of the health of the patient. In the extreme case, an unattended adverse event may ultimately result in a cardiac arrest, unanticipated death, or in an unanticipated admission to the ICU which may or may not be accompanied by a cardiac arrest. Many extreme adverse events occur outside the ICU on general or intermediate wards [104, 31]. These are low-acuity wards where monitoring of patients is less intense than in the ICU, as shown in Table 1.1. Here the staff-to-patient ratio, as well as the frequency of monitoring is relatively reduced, and this has been linked to the incidence of most adverse events and corresponding low survival rate in these wards [65].

Although deterioration can occur at any time, certain patients are more at risk [6]. These include patients who have recently had an ICU admission, and patients who are recovering from a critical illness [44]. A 40.8 % mortality rate has been associated with post-ICU patients [123]. These patients may have been in a medical ICU or surgical ICU.

Three of the leading complications in these high-risk patients are cardiac or circulatory diseases, respiratory diseases, and gastrointestinal (GI) diseases [172]. These are the leading causes of death in the UK, responsible for 207,000 (cardiac arrest), 123,000 (respiratory), and 59,000 (GI) annual deaths, respectively [172]. Studies on the incidence of adverse events including cardiac arrest and unanticipated ICU admissions have found cardiac, respiratory, GI-related, sepsis-related, neurological, and metabolic complications to be the top underlying conditions associated with adverse events, but few have reported on the relative incidence of these associated conditions, as shown in Table 1.2. Not all

Condition	[140]	[144]	[133]	[170]
Cardiac	•9%	•	•32%	•63%
Respiratory/pulmonary	•38%	•	•12%	•15%
Neurological			•11%	
Metabolic	•11%		•	•2%
Gastrointestinal		•	•8%	•3%
Cancer		•	•11%	
Sepsis			•10%	•2%
Renal		•	•8%	

Table 1.2: The major underlying conditions associated with a cardiac arrest. A condition reported by a study is marked by a •. Where reported, incidence of a condition is shown as a percentage.

studies shown in this table reported relative incidence as a percentage. Further, some studies included fewer conditions than others, resulting (where reported) in the variation in percentage incidence associated with any single condition.

This thesis focuses on adverse events in post-surgical and post-medical ICU populations that have a high risk of emergency re-admission. Two of the leading causes of adverse events leading to emergency re-admission are cardiac arrests and GI complications.

1.2.2 Cardiac Arrest

In the event of a cardiac arrest, the heart fails to contract effectively, resulting in the cessation of the regular circulation of blood in the body, and the delivery of oxygen to tissues. Lack of oxygen supply to the brain results in loss of consciousness which in turn results in suppressed breathing. If untreated within minutes, this can lead to immediate death. The immediate treatment for a cardiac arrest is called cardio-pulmonary resuscitation (CPR), which in the most basic form involves the application of chest compressions in order to manually restore the circulation of blood in order to maintain the blood supply to the brain. To restore the normal functioning of the heart, however, defibrillation then has to be carried out, in which a electric shock is applied until spontaneous circulation resumes. The major physiological indicators are those which are involved with breathing and circulation and therefore delivery of oxygen; therefore the most commonly reported physiological indicators are the heart rate, breathing rate, oxygen saturation, and blood pressure [146, 71].

While the incidence of cardiac arrests in the UK is reported to be between 3.8 [114]

and 7.0 [57]events per 1000 beds annually, and in the USA more than 150,000 cardiac arrests are reported [31], the survival rates after a cardiac arrest are very low, with mortality rates being over 80% in the UK [65] and in the USA [31]. According to a large meta-analysis of 98 studies on in-hospital CPR [141], it was found that the survival-to-discharge ratio (which is the proportion of cardiac arrest patients successfully resuscitated and discharged from the hospital) was 15% (2,994 out of a total of 19,955 patients). It was further shown that the survival-to-discharge ratio was 12.2% for in-hospital cardiac arrest patients aged 70-79 years, 10.2% for patients aged 80-89 years, and 0% for those aged over 89 years [141, 113]. Moreover, in patients who survived to be discharged alive from the hospital, the mortality rate within 36 months after discharge was 66%. This shows that the morbidity and mortality associated with cardiac arrest remain high and the patients outcomes continue to be bleak [113].

1.2.3 Emergency ICU Admission

An unplanned or emergency ICU admission occurs when a patient who is on a general or intermediate ward outside the ICU deteriorates so seriously that critical care becomes necessary which is provided in the ICU[2]. Patients who have had a medical or surgical ICU admission are more vulnerable to death or to a re-admission to the ICU during the same hospitalisation, and are called high-risk patients. High-risk patients have a higher mortality and LOS. The risk of an emergency admission increases according to the severity of the illness [59].

According to a systematic review of studies on ICU admissions, 37.2% of adverse events led to an ICU admission, and between 17% and 76% could be avoided with better care [169]. ICU admissions have up to 58% mortality rate [169].

A recent large multi-centre study in the USA [2] conducted on data collected from 469,976 ICU patients from 219 hospitals showed that there were 17,874 (2.54%) emergency ICU re-admissions following a discharge from the ICU. Furthermore, there were 6,492 (0.92%) cases of death within 48 hours of the initial discharge from the ICU, and 499 (0.07%) cases of a re-admission to the ICU followed by death. Another study found the

[2]An planned ICU admission usually follows pre-scheduled surgery and is called an elective ICU admission.

incidence of re-admissions in the USA to be 2% within 48 hours and 4% within 128 hours of ICU discharge, with a median interval between the admissions of 3 days [18].

Although the reasons for the original admission to the ICU and re-admission can be different, the leading reasons have been identified as cardiac, respiratory, GI complications, renal, neurological, and sepsis, with cardiac, respiratory and GI leading the list [18, 139, 89, 132, 30, 81, 52, 5, 107, 138, 29]. Table 1.4 shows the most common underlying causes of ICU admissions identified in the literature.

In a study [89] of 97 ICUs in 35 hospitals in the USA 229, 375 ICU admission cases were investigated. Both medical and surgical ICUs were considered. It was found that 6.1% of the cases resulted in a re-admission to the ICU within a median time period of 3.4 days. The re-admission population had a four times greater mortality rate than the population without re-admission, and a 2.5 times greater hospital length-of-stay (LOS). The major conditions for re-admission in the medical ICU patients were cardiac, respiratory and GI. In the post-surgical ICU patients, the major reason for re-admission was GI complications. The factors linked with an increased likelihood of re-admission included physiological abnormalities following the initial discharge from ICU, age, and length of stay in the step-down unit (SDU) outside the ICU.

Pearse et al. [123] studied the high-risk post-surgery population in a study of 94 NHS hospitals in the UK between 1999 and 2004. The definition of high-risk was a population having a mortality rate of 5% or more. It was found that 12.5% of post-surgery patients were high-risk and that these 12.5% accounted for more than 80% of the deaths. Less than 15% of the high-risk patients are admitted to the ICU and most of these high-risk deaths occur on general or intermediate wards outside the ICU.

Staffing levels on general and intermediate wards have an impact on the incidence of re-admission [132] [107]. Renton et al. [132] conducted a retrospective longitudinal study of 38 Australian ICUs between 2000 and 2007. Their results showed that ICU admission rates had not decreased during these years. They also found that patients are more likely to have a re-admission if they are discharged from the ICU during the off-peak hours (i.e. 6 pm - 6 am), when there are fewer staff on duty and frequency of patient monitoring is reduced. The major underlying causes for ICU re-admission

were also reported. A majority (80%) of the ICU re-admission were found to be due to cardiovascular diseases (27%), respiratory diseases (25%), post op GI complications (12%), neurological complications (11%) or sepsis (8%).

Another study on ICU re-admission by [81] reported an ICU re-admission rate of 13.4%, and that mortality increased five times with re-admission. The two leading causes of re-admission were cardiac (14.3%) and respiratory (13%) complications. In most re-admissions within 24 hours, GI surgery was performed. Cardiovascular complications were noted as the main reason for most first re-admissions, and GI and respiratory complications were the main reasons for most subsequent re-admissions. Similar results were obtained in another study [5] which found the re-admission rate to be 3.6 %, while the mortality rate was much higher (17%). The main cause of re-admission was respiratory failure, and the main underlying conditions were respiratory (47%), cardiac(20%), sepsis (14%), and GI (10%).

Re-admission can occur any time between hours and days following the initial discharge from ICU. Metnitz et al.[107] reported a 5% re-admission rate, and found that there was no significant difference in mortality between early re-admission (within 48 hrs) and late re-admission (3-4 days). The highest risk of re-admission was reported to be cardiovascular surgery, renal surgery, and GI disease. It was indicated that a low nurse-to-patient ratio may be a contributing factor in re-admission.

A review of 20 ICU admission studies from countries including UK, US, Canada, Australia, Korea, Saudi Arabia, India, Israel, and Austria was performed by Elliot et al. [52]. The average rate of re-admission was 7.78%. The review found that there had not been a decrease in the incidence of re-admission from 1983 to 2005. GI disease was a potential risk factor identified repetitively by several studies. Cardiac and respiratory disease were the most common diseases in the ICU re-admission population. Abnormal heart rate, breathing rate, and oxygen saturation were identified as major indicators of deterioration.

Evidence of early warning signs preceding an emergency ICU admission or a re-admission has been presented in several studies [89, 59, 30, 52, 5, 107, 138] . Table 1.3 presents a summary of the main findings of studies on emergency ICU admissions and re-admission,

which suggest, in general, an increase in mortality rate and the presence of physiological indication in re-admission cases.

Study (Year)	Description	Key findings
Rosenberg [138] (2001)	Prospective cohort study of 4,684 ICU admissions in a US hospital.	Mortality rate of re-admission population was 5 times that of non-re-admission population, and LOS 2 times greater. Incidence of re-admission might be higher in patients who respond poorly to treatment.
Rosenberg [138] (2001)	Prospective cohort study of 4,684 ICU admissions in a US hospital.	Mortality rate of re-admission population was 5 times that of non-re-admission population, and LOS 2 times greater. Incidence of re-admission might be higher in patients who respond poorly to treatment.
Bardell [5] (2003)	Retrospective study of 2117 coronary bypass surgery patients.	ICU re-admission rate was 3.6 %. Re-admission mortality rate was 17%. Main cause for re-admission was respiratory failure.
Seward [144] (2003)	Study of 200 deaths following emergency admission.	Poor care led to unexpected deaths in 25 out of 200 deaths. Diagnosis was delayed in 64% of cases.
Pearse [123] (2006)	A study of ICU admissions in 94 NHS hospitals in UK.	Post-ICU mortality rate was 40.8%. 39% of these deaths were of patients sent to ICU from general wards. 12.5% of post-surgery patients accounted for 80% deaths; less than 15% of them reached the ICU.
Frost [59] (2009)	Meta-analysis of 11 studies from 1966 to 2008 on ICU re-admissions	Risk of re-admission increased by 43% with an increase of 1 standard deviation in the severity of illness score. Re-admission rate was directly proportional to the severity of illness.
Chrsuch [30] (2009)	Prospective cohort study of over 10,000 ICU admissions to a hospital in Canada, between 1989 and 1996.	ICU re-admission rate was 5.5 %. High occupancy in ICU increases risk of re-admission. Main readmission risk factors were cardiac, respiratory, sepsis, neurological, GI, trauma, and abnormal physiology.
Renton [132] (2011)	Retrospective study of 38 ICUs in Australia from 2000-2007.	5.5 % patients re-admitted to ICU, including some at night. 82% re-admissions were due to cardiac, resp, GI, neurological, sepsis. Mortality rate of patients with re-admissions was 5 times that of patients without re-admissions (20.7 % vs 4.4%).

Study (Year)	Description	Key findings
Vyalen [169] (2011)	Systematic review of 17 studies on adverse events before emergency ICU admission.	Up to 37.3% adverse events required ICU admission. ICU re-admission rate was up to18.3%. Between 17% and 76.5% adverse events were preventable.
Kramer [89] (2012)	Study of 97 ICU units across 35 US hospitals.	ICU re-admission rate was 6.1%. Patients with ICU re-admission had a 4 times higher mortality rate and a 2.5 times greater LOS than patients without re-admission. Physiological abnormalities were present in re-admission population.
Brown [18] (2012)	Retrospective study of 156 ICUs in 106 US hospitals.	2% and 4% re-admissions in 48 and 128 hrs, respectively. An average of 3 days in SDU before re-admission. Main risk factors for re-admission were GI, cardiovascular, renal, cancer, HIV.

Table 1.3: Summary of the recent studies on emergency ICU admissions (contd. from previous page).

1.2.4 Early Warning Signs of Deterioration

Studies [83, 146, 140, 55, 7, 57, 133, 104, 71, 73, 10, 19] have reported on various premonitory indicators of adverse events, including physiological, biochemical, and symptom-based indicators.

Physiological Indicators

Physiological indicators include heart rate, breathing rate, arterial oxygen saturation, systolic and diastolic blood pressures, and temperature, and urine output. Excluding urine output, these physiological indicators are called vital signs. Vital signs can be measured non-invasively.

Other indicators are biochemical or symptom-based. Biochemical indicators include the blood levels of creatinine, sodium, potassium, haemoglobin, white cell count, platelet count. These are measured by testing a sample of blood. Symptom-based indicators include lethargy, loss of consciousness (which is measured by the Glasgow coma scale (GCS)), anxiety, shallow breathing or shortness of breath (dyspnoea), chest pain, change in mental status, and nurse/doctor concern for the patient.

Condition	[81]	[5]	[107]	[138]	[29]	[89]
Cardiac	•	•	•	•	•	•
Respiratory/ pulmonary	•	•	•	•	•	•
Neurological					•	
Gastrointestinal	•	•	•	•	•	•
Sepsis		•	•	•		
Renal		•	•			

Table 1.4: Major conditions associated with emergency ICU admissions, marked by • where reported.

Early Warning Signs

Studies on early warning signs of deterioration have used different definitions and ranges of abnormality, e.g. the lower limit of tachypnoea[3]. Some have reported on the incidence of tachycardia (i.e., an abnormally high heart rate) and bradycardia (i.e., an abnormally low heart rate) whereas others have used the combined definition of "heart rate abnormality". Due to the heterogeneity of the methodology and statistics used in these studies, it is difficult to report on standardised results. Table 1.5 summarizes the main indicators of physiological deterioration reported by various studies which used similar categorisations of abnormality. Heart rate, breathing rate, and blood pressure appear in most studies, whereas other physiological indicators such as oxygen saturation, temperature, GCS, mental status, nurse / doctor concern, and urine output are reported by some studies only, illustrating the difficulty in obtaining consistent results.

In 1990, Schien et al. [140] found that 84% of the 64 patients who had had an in-hospital cardiac arrest had shown premonitory signs of clinical deterioration within 8 hours of the arrest, and further that these patients had a 5% survival rate. They noted that while no signs of abnormality were found in the laboratory data, an abnormality in respiratory or neurological function was found in overall 70% of the patients. The major abnormalities were found to be a high respiratory rate (38%), and reduced neurological function. The major underlying causes of the cardiac arrest were respiratory (38%), metabolic (11%), cardiac (9%), and neurological function (6%). Multiple causes were found in 27% of the cases. Respiratory function was also considered the major indicator by Bedell et al. [7] who reported that 28% of the patients who had an in-hospital cardiac arrest had had

[3]An abnormally high breathing rate is known as tachypnoea.

Physiological indicators	[147]	[73]	[83]	[7]
Heart rate	54% (tachycardia)	79%	9% (tachycardia)	
Breathing rate	58% (tachypnoea)	65%	11% (tachypnoea), 4% (bradypnoea)	50%
Blood pressure	46% (hypotension), 8% (hypertension)	86%	31%	
Oxygen saturation		71%		
Temperature	13% (pyrexia), 4% (hypothermia)	47%		
Mental status	46%			
Urine output	29%			
Consciousness (GCS)			24%	
Chest pain		37 %		
Nurse/doctor concern		41%		

Physiological indicators	[133]	[71]	[10]
Heart rate	7% (tachycardia), 5% (bradycardia)	7% (tachycardia), 6% (bradycardia)	13% (tachycardia), 6% (bradycardia)
Breathing rate	15% (tachypnoea), 1% (bradypnoea)	12% (tachypnoea), 0% bradypnoea	
Blood pressure	28% (hypotension) 1% (hypertension)	31% (hypotension)	
Oxygen saturation			
Temperature			
Mental status			
Urine output			
Consciousness (GCS)		4%	
Chest pain			14%
Nurse/doctor concern			

Table 1.5: Physiological indicators of cardiac arrest. Incidence of abnormality in an indicator is shown as a percentage, rounded off to the nearest unit. Where specified, the type of abnormality is shown in brackets. An abnormally high heart rate is known as tachycardia, while a low heart rate is known bradycardia. Abnormally high, and abnormally low breathing rate are called tachypnoea and bradypnoea, respectively. Pyrexia (fever) is the condition when core body temperature is higher than normal, and hypothermia is the condition when core temperature is abnormally low. Hypertension and hypotension refer to high and low blood pressure, respectively. Glascow Coma Score (GCS) is a measure of consciousness. Here a drop in GCS by 2 points or more is reported.

physiological signs to which there had been an inadequate response. They report that 9% of all arrests may have been avoided with proper care and timely response.

Another study by Fiesselmann et al. [55] also concluded that respiratory rate (or breathing rate) was a major predictor of cardiac arrest. They conducted a case-control study in which the last 72 hours of patients who had suffered a cardiac arrest were compared with the same time period for patients who had not had a cardiac arrest, and found that 54% of the cardiac arrest patients had a breathing rate of greater than 27 breaths per minute, whereas only 17% of the non-cardiac arrest population had a high breathing rate at admission. They found that respiratory rate had a sensitivity of 0.54 and a specificity of 0.83 in predicting cardiac arrest.

In one of the earliest studies, Franklin et al. [57] investigated 150 cardiac arrests that had a 91% mortality rate. It was found that 66% (99 out of 150) cardiac arrests had signs of deterioration up to 6 hours before the event, and that these signs were either not reported by the nurse to the doctor, or were not responded to by the doctor. The study concluded that cardiac arrests outside the ICU are preceded by early warning signs and that systems to recognise these signs early should be in place.

Hodgett et al. [73] conducted a control study including 118 cardiac arrest patients and 132 patients without cardiac arrests in order to determine the main indicators of cardiac arrest. Abnormalities were found in the cardiac arrest cases in the breathing rate (65%), heart rate (79%), systolic blood pressure (86%), temperature (47%), and oxygen saturation (71%). Chest pain and nurse/doctor concern was found in 37% and 41% of patients, respectively.

An important early study investigated the reasons for 778 deaths in three hospitals in Australia over a six-month period [71]. It was found that all deaths had been preceded by either severe vital-sign abnormalities (30%), cardiac arrest (22%) or an emergency ICU admission (21%). Serious vital-sign abnormality was present in the last 8 hours before death for 50% of the patients who did not have a DNR[4] order, and in half of them severe vital-sign abnormality was present between 8 and 48 hours prior to death. A third of the patients without a DNR order had severe vital sign abnormality in the

[4]A Do-Not-Resusciate (DNR) order is issued when no treatment is desired in the case of a cardiac arrest. A death without DNR may be assumed to be unexpected.

entire 48 hour period prior to death, indicating that they had been high-risk for that entire period without the situation being appreciated and appropriately responded to. The main indicators were hypotension (30%) and tachypnoea (17%).

Another study [146] had earlier reported that 51% of those who had had an in-hospital cardiac arrest had vital-sign abnormalities, and here too the top three indicators were tachycardia (54.0%), hypotension (45.8%), and tachypnoea (58.3%). 10.5% of patients were found to have vital-sign abnormalities in the last 8 hours before arrest, and 21.5 % between 8 and 24 hours before arrest. Over two-thirds of the patients (68%) were found to have a vital-sign abnormality in the 24-hour period prior to the arrest.

In 2004, the first large-scale multi-national investigation [83] conducted over a three-day period in 90 hospitals in he UK, Australia, and New Zealand reported that physiological abnormality was present in 60% of cases, including 54.5% of (DNR and unexpected) deaths, 79.4% cardiac arrest cases, and 54.5% of the unanticipated ICU admissions, between 15 minutes and 24 hours prior to the event.

Table 1.6 shows a summary of the studies on the incidence and timing of extreme adverse events outside the ICU (cardiac arrests, deaths and emergency ICU admissions). The majority of theses studies have found that an extreme adverse event is preceded by premonitory signs in the the last 24 hours, and that these events are predictable. Several of these studies have reported the level of care outside the ICU to be inadequate, and that with proper care these events are avoidable.

1.2.5 Monitoring Outside the ICU

At least half of the adverse events occur outside the ICU in general wards which are often infrequently monitored [158], and receive suboptimal care [7, 105]. According to the 2012 NCEPOD report covering NHS hospitals in the UK [80], although early warning signs were present in 75% of cases of cardiac arrest, these signs were not recognised in 35% of the cases, and not acted upon in 56% of cases. Almost half the cases (55%) were not referred to senior doctors. The report outlines the following reasons for the lack of proper care outside the ICU.

- Failure to properly assess the urgency of the situation. This included deficiencies

Study (Year)	Onset of deterioration	In-hospital mortality	% Cases with premonitory signs	Inadequate response	Avoidable events
Schien [140] (1990)	within 8 hours before arrest	92%	84%		•
Bedell [7] (1991)	-	-	28%	•	•
Fieselmann [55] (1993)		-	54%		•
Franklin [57] (1994)	6 hours within event.	91%	66%	•	
Smith [147] (1998)	$<$ 8 hours after admission (10.5% of 47 patients), 8-24 hours after admission (21.5%), $>$ 24 hours after admission (68.0%)	-	51%	•	•
Rich [133] (1999)	At least one physiological change 8 hours before event.	67 %	60%	•	•
McGloin [104] (1999)	-	52%	65%	•	•
Buist [19] (1999)	Median of 6.5 hours before event.	62%	76%		•
Hillman [71] (2001)	50% 0-8 hours before, and 50% 8-48 hours before event.	All deaths	100% of which 30% severe abnormality, 50% in the DNR group	•	•
Berlot [10] (2004)	6 hours before event.	84 %	86.4%	•	•
Kause [83] (2004)	15 minutes to 24 hours within event.		60%		•

Table 1.6: A summary of studies on the incidence and timing of premonitory signs before adverse events including cardiac arrest, emergency ICU admission, and unexpected death. The second right-most column shows a • if inadequate monitoring or response was reported. The right-most column shows a • if the study found events which were considered avoidable.

in initial assessment of the patient (48%), indecision about CPR status (90%), for instance. 9% of patients were considered to have wrongly been admitted to the general ward (Level 1), when they should have received Level 2 or Level 3 care.

- Infrequency of patient monitoring, poor recognition of physiological warning signs, and deficiency in timely escalation to senior doctors.

The report concluded that care on general and intermediate wards was inadequate in 70% of the cases.

Franklin et al. [57] reported that out of 99 cases of cardiac arrest where signs of physiological abnormality were present up to 6 hours in advance, 25 % of the cases were not escalated to the physician by the nurse. In 43% of cases the patient was seen by the physician but the physician did not inform the ICU in spite of the presence of physiological deterioration in the vital signs and in the mental status of the patients, and in 91 % of these cases the physician also did not take a blood gas sample. In the remaining 32% cases when the physician did inform the ICU, the physician was unable to stabilise the patient such that the patient suffered an arrest while in transit to the ICU.

In another study, Bedel et al. [7] also report that in 28% of the patients who had an in-hospital cardiac arrest, abnormalities in vital signs were inadequately responded to. They report that 9% of all arrests may have been avoided with proper care and timely response.

Hodgetts et al. [73] reported that the odds of potentially avoidable cardiac arrest in general wards was 5.1 times that of the odds in the ICU.

McGloin et al. [104] reported in a 6-month audit of a hospital that 66.5 % of the 477 deaths in the hospital had occurred on the general ward, 6% of which were unexpected deaths. 13 out of these 20 unexpected deaths had clear signs of deterioration which were not acted upon. ICU admission due to inadequate care on the general ward (i.e., non-recognition of premonitory signs and failure to respond) was seen in 36% of cases, and the corresponding mortality rate in these poorly-managed cases was much higher than in cases that were better managed (52% vs 35%).

Other studies have also reported poor management of patients on the general wards [133, 6, 57, 7, 104, 71, 10]. Rich et al. [133] reported that only 6 out of 37 adverse events

were reported to senior staff, although an abnormal change was present in at least one vital sign up to 8 hours in advance.

In [105] , McQuillan et al. summarize the reasons for inadequate care on general wards as being the lack of knowledge of organised procedures, inexperience and lack of communication between nurses and senior staff, and the lack of medical staff and resulting infrequency of monitoring due to low staff-to-patient ratio on the general ward. With better care and timely intervention, as many as 41% of ICU admissions were considered preventable.

The recognition of these factors led to the development of intervention mechanisms for timely response to deterioration. In the next sections existing and emerging interventions are introduced and their impact is discussed.

1.3 Interventions

Recognition of the significance of early intervention on general and intermediate wards outside the ICU has led to the development of rapid response systems, which are aimed at improving the clinical decision-making process by enabling monitoring of patient condition at regular intervals, and generating an alarm or "trigger" in the event of deterioration. A rapid response system has an afferent arm which is responsible for the identification of at-risk patients, and an efferent arm which responds to the call for help by the afferent arm and administers care [47].

1.3.1 Track-and-Trigger

In the UK, physiological abnormality in vital signs is identified, and responded to, through the use of Track-and-Trigger systems. Track-and-trigger has been developed following guidelines issued by the National Institute of Clinical Excellence (NICE) for assessing patient status [21]. Nurses record vital-sign values at regular intervals, and record the observations using either paper charts, or hand-held PDAs, enabling a patient's vital signs to be "tracked" through time. A track-and-trigger score is calculated on the basis of the observations, and if the score exceeds a certain threshold, a response by the staff is "triggered" and the patient's condition is reviewed. Frequency of track-and-trigger

observations varies according to patient condition and nurse workload, and in general track-and-trigger observations are typically taken every 8-12 hours in the general ward. The frequency of monitoring is raised to every 4 hours or more if staff are concerned about the patient.

The efferent arm or the clinical team responds according to the level of alarm, i.e. the level of deterioration in vital signs. A moderate deterioration is often responded to by increased vigilance by the duty nurses, i.e. an increase in the frequency of track-and-trigger observations. However, if the deterioration is more serious, the duty nurse may alert the Critical Care Outreach (CCO) team, which is nurse-led, or the Intensive Care Unit (ICU) team. These teams are trained to provide critical care by the bedside, and prepare the patient for transferral to ICU, or to other high-dependency units. Critical care outreach teams were developed in 2000 by the NHS in the UK [60].

1.3.2 Medical Emergency Team

In Australia and US, the dedicated physician-led response team is called the Medical Emergency Team (MET) [60]. The MET is slightly different from the CCO team in the sense that it can administer ICU level care on-the-spot even outside the ICU. Here too the "tracking" is performed by nurses who record the vital signs periodically. If and when some MET calling criteria are met, a MET call is issued and the MET responds on the ward.

All rapid response systems rely on some "calling criteria". If and when the criteria are fulfilled, an "alert" or the call for help is triggered. Over the last decade, calling criteria have evolved into early warning scores (EWS).

1.3.3 Early Warning Scores

Numerous calling criteria or scoring systems have been used in the literature [60]. The two main categories are

(i) single-parameter scoring systems, and

(ii) multi-parameter or aggregate scoring systems.

In single-parameter systems, the alarm is triggered whenever any one vital sign exceeds

Vital Sign	Lower Limit	Upper Limit
Heart rate (bpm)	40	140
Breathing rate (rpm)	8	36
Systolic blood pressure (mmHg)	80	200
Diastolic blood pressure(mmHg)	-	110
Arterial oxygen saturation (%)	85	-

Table 1.7: An example of single-parameter MET calling criteria used in the University of Pittsburgh Medical Centre in the USA in 2008 [75].

Vital Sign	Score						
	3	2	1	0	1	2	3
Heart rate (bpm)		≤ 40	41-50	51-100	101-110	111-129	≥ 130
Breathing rate (rpm)	≤ 8			9-18	19-24	25-29	≥ 30
Arterial oxygen saturation (%)	≤ 91			≥ 92			
Systolic blood pressure (mmHg)	≤ 90	91-99		101-179			≥ 200
Temperature (^{o}C)		≤ 35.0		35.1-37.9			≥ 38.0

Table 1.8: An example of a multi-parameter system. The early warning score (EWS) system used in the upper-GI ward of the Oxford University Hospitals, Oxford in 2011.

a pre-defined limit. Table 1.7 shows an example of the MET calling criteria used in one hospital in the USA. In addition to the vital signs shown in the table, the calling criteria included symptoms such as acute loss of consciousness, sudden collapse, seizure, sudden loss of mobility, extreme colour change and uncontrolled bleeding [75]. While such systems are simple to use, they do not take into account the fact that some deteriorations may occur as subtle changes in multiple vital signs, rather than a noticeable change in a single vital sign [159].

This limitation is overcome aggregate scoring systems, in which a weighted score is assigned to each vital sign, and the scores are combined into an overall score, known as the EWS, or the Patient At Risk (PAR) score.

The original EWS, called MEWS (modified early warning score), used criteria based on the Acute Physiology and Chronic Health status Evaluation (APACHE) system which

has become a standard for evaluating the degree of illness in ICU-admitted patients, and predicting ICU deaths [86]. It combines various physiological, biochemical, and physical indicators to assess patient status. MEWS assigns individual scores to vital signs recorded on the ward during routine checks, and combines them to produce the aggregate score. Clinical review is prompted, or, in some hospitals, an emergency team responds, if a predefined threshold is exceeded. Common triggers have been reported as being a nurse's concern about patient, changes in GCS, systolic blood pressure, mental state, and cardio-respiratory changes such as hypotension, hypoxia, and tachypnoea [27].

Table 1.8 shows an example of the EWS system in use in hospitals in Oxford, UK, at the time of the trial described in Chapter 7 of this thesis. If the total score for the vital signs exceeds 4, or if an individual score for any one vital sign exceeds 3, an alert is generated and a clinical response is triggered.

According to some studies [129, 146], EWS systems suffer from the drawback of being difficult to calculate, and are therefore susceptible to miscalculation errors. Thus EWS have received mixed reviews in the literature, with have criticising their effectiveness [50, 153], and others encouraging their use in hospitals after modifications [151, 62, 46].

The performance of an early warning system and its calling criteria is typically assessed in three ways: (1) retrospective analysis, (2) before-and-after analysis, and (3) randomised control trial.

Retrospective Analysis

Retrospective analysis is possible when annotations such as patient outcomes are available for data that has already been collected from a trial. Based on the outcomes the performance of a system can be evaluated in retrospect. For instance if the EWS generate an alert and we know (with hindsight) that the episode was an adverse event, then this is an instance of a true positive.

If the EWS fails to generate an alert for a known adverse event, this is a false negative. A false positive occurs if an alert is generated when no adverse event occurred. Finally, a case is deemed to be a true negative if an event did not occur and no alert was generated. This can be summarised into a "confusion matrix" as shown in Table 1.9.

	Actual True Event	Actual No Event
Predicted True Event	True Positive	False Positive
Predicted No Event	False Negative	True Negative

Table 1.9: A 2x2 confusion matrix for ROC analysis.

Often a fundamental problem is defining a true positive for an event such as patient deterioration, which is why hard outcomes such as mortality, ICU admission and LOS are used.

In the clinical setting it is not only desired that adverse events are detected, but also that false alerts are not generated. From this it follows that a better early warning system is one that maximises the number of true positives, i.e. has a high "sensitivity", and minimises the number of false positives, i.e. has a high "specificity", where sensitivity and specificity are defined as follows.

$$\text{Sensitivity} = \frac{\text{true positives}}{\text{true positives+false negatives}} \tag{1.1}$$

$$\text{Specificity} = \frac{\text{true negatives}}{\text{true negatives+false positives}} \tag{1.2}$$

Often there is a trade-off between the sensitivity and specificity of an EWS. Sensitivity can be plotted against (1-specificity) in what is known as a Receiver Operator Curve (ROC). The Area Under the Receiver Operator Curve (AUROC) can be used to determine the overall performance of an EWS system.

Before-and-After studies

Outcomes are compared in a "before" phase prior to the introduction of the intervention to the outcomes in the "after" phase when the intervention is implemented. Markers of outcome such as mortality, hospital length of stay, admission to ICU can be compared. With this approach, the control of time-based variables cannot be guaranteed. For instance, variables such as mortality and length of stay have seasonal trends. Further, staff alertness can also be impacted by their awareness of the phase of the study – i.e., staff may be more alert in the intervention phase than in the "before" phase which may impact the reliability of the results.

Randomised Control Trial

A randomised control trial (RCT) has a prospective design, and is commonly considered to be more reliable scientifically than other methods. Patients are randomly divided into a "control" group and an "intervention" group. In the intervention group, the intervention is administered but in every other way, the two groups are deemed to be the same. With all other variables being constant, the effect of the intervention alone can be assessed by comparison between the two groups.

1.4 Impact of EWS-based Interventions

Table 1.10 summarizes the key findings of some studies conducted in the last decade on the impact of the use of EWS systems. These include studies on MET and CCO teams using EWS systems.

Single-Parameter EWS systems

A review of single-parameter EWS was conducted by Smith et al. [149]. Seven studies were reviewed and it was found that there are at least 39 different categories of single-parameter EWS systems. Of these 30 were evaluated and it was found that (a) there is a very wide range of sensitivity and specificity of these EWS, and (b) although reported specificities are high, the sensitivities reported by these studies are too low to introduce single-parameter EWS at scale.

Multi-parameter EWS systems

In 1997, Morgan et al. [110] first reported an EWS that combined vital-sign and mental status abnormalities to detect deterioration. Since then several modifications have been adopted in different institutions. Hodgetts et al. [72] carried out a retrospective study to identify risk factors for a cardiac arrest in order to formulate an early warning score using statistical results rather than empirical rules as done in the initial early warning score systems. Cardiac arrest case notes of 118 patients were studied and associated risk factors were weighted according to relative predictive performance. The main risk indicators

were identified as respiratory rate, blood pressure, temperature, oxygen saturation and nurse/doctor concern. An aggregate EWS score was computed and it was shown that all patients with a score > 10 had suffered a cardiac arrest. A score of > 8 was found to have 99% specificity and 52% sensitivity.

The aforementioned EWS was used in a before-and-after design study conducted by Kenward et al. [85] in a UK hospital. During a 12 month intervention period, 136 MET calls were made. Risk factors identified were similar to those identified by Hodgetts et al. [72]. No significant reduction in the in-hospital mortality and and incidence of cardiac arrests was found in the "before" and "after" phases. A possible reason for this was late activation of the MET; 24% of the calls were made within 5 minutes of the cardiac arrest, whereas deterioration in multiple vital signs was reportedly present in the majority of these cases several hours in advance.

Bellomo et al. [9] were the first to report on an overall reduction in in-hospital mortality with EWS-based intervention. They conducted a before-and-after study at an acute-care hospital in Australia, with a 4-month "before" period and a 4-month "after" period, following an intermediate period of training the staff in the use of MET. The effect of seasonal variation was taken into account by ensuring that the "before" and "after" phases fell on the same 4 months of the year. The calling criteria included abnormal changes in vital signs including heart rate, breathing rate, oxygen saturation, blood pressure, and in the consciousness level and urine output. Nurse/doctor concern was also included. A total of 99 MET calls were recorded in the intervention period, with the majority being caused by nurse/doctor concern, and changes in blood pressure, oxygen saturation and consciousness. A relative risk reduction (RRR) of 65% (from 63 before to 22 after) was found in the incidence of the in-hospital cardiac arrests, and an RRR of 26% (from 302 before, to 222 after) in the in-hospital mortality. In a subsequent paper based on the same study [8] MET calls for a post-operative population were investigated. It was reported that, after excluding DNR cases, the post-intervention in-hospital survival rate was 89.4%. An overall RRR of 36.6% was reported in the in-hospital mortality for post-operative patients (from 73 before to 45 after the intervention). Finally a reduction in LOS was also reported from an average of 23.8 before to 18.9 days after the intervention, corresponding

to a reduction in 4000 bed days in the hospital. For individual adverse events, an RRR of 79.1%, 78.2%, 74.3%, and 88.5% was found in cases of respiratory failure, stroke, sepsis, and renal failure, respectively, resulting in an overall $> 50\%$ reduction in adverse events in post-operative patients associated with MET intervention. It was recognised that the before and after nature of the study imposed certain limitations and an RCT should ideally be conducted, but an RCT was not considered plausible for a single hospital.

The only RCT to date in a single-hospital set-up was reported 2004 by Priestley et al. [127]. A ward-randomised trial was conducted in 16 adult wards of a hospital in the UK in which the intervention was performed in each ward sequentially. A 24 hour CCO team led by nurses was implemented in the intervention group, with usual care in the control group. This study investigated only two outcomes - in-hospital mortality and LOS. The main finding of the study was that CCO intervention was associated with a reduced risk of in-hospital mortality. However, two main limitation were reported. One was the sample size of the control and intervention groups due to implementation in a single hospital. A multi-centre RCT was recommended. The second problem was that of the "Hawthorne effect" - staff were aware of the study and therefore unusually motivated during the study which may have impacted the results.

The Medical Emergency Response, Interventions and Therapy (MERIT) study [70] is the only multi-centre RCT to date. It was conducted across 23 hospitals in Australia. Outcomes were compared in 12 randomly selected hospitals using MET with 11 control hospitals where no MET was in place. The study outcomes were in-hospital mortality, and emergency ICU admission. No significant impact of the intervention was found on the outcomes including cardiac arrests, emergency ICU admissions, and unexpected deaths. It was found, however, that the existence of a MET team was associated with a higher rate of emergency calls (8.7 vs 3.1 per 1000).

Among recent studies ([164, 68, 120, 108, 93]), Moon et al. [108] did an 8 year before-and-after study in a UK teaching hospital. In the "before" phase between 2002 and 2005 there were 213,117 hospital admissions, and in the "after" phase between 2006 and 2009, there were 235,516 admissions. A 24/7 CCO team was used. A significant reduction in cardiac arrests as a proportion of all adult admissions was seen, from 0.4% to 0.2%.

Similar trends were seen in emergency ICU admissions (2% "after" versus 3% "before") and in mortality rates (42% "after" versus 52% "before"). A limitation of the study was not being able to control changes in disease patterns and admission patterns such as the increase in overall admission rate of 10.5% , and a decrease in overall emergency hospital admissions from 31% to 28%.

Laurens et al. [93] also carried out a before-and-after study and reported a positive impact of EWS-based intervention. The study was conducted in a 150-bed regional hospital in Australia between 2004 and 2008. They reported that the mortality rate decreased from 9.9 to 7.5 per 1,000 (corresponding to a relative risk reduction[5] of 24.2%), ICU admission rate from 22.4 to 17.6 per 1000 (with a corresponding relative risk reduction of 21.4%); and the relative risk reduction in cardiac arrest rates was 45.5%.

Another recent study was conducted by Harrison et al. [68]. They studied the impact of CCO teams in 52 acute care hospitals in the UK. 2,203 patients received CCO team visits before going into a critical care unit. There was no significant impact of the CCO team visit on the severity of illness. Of the 5,924 patients who were visited after being discharged from the critical care unit, a reduction in mortality rate was not found when compared with the control group. The study could not report conclusive results on the impact of CCOs.

With seemingly contradictory results, the studies described above illustrate, to some extent, the difficulty associated with assessment of clinical interventions. Most EWS systems are designed empirically based on expert opinion and without prospective evaluation, and have therefore shown variable performance [84, 149]. The main drawback is that most EWS systems are validated for patients in local institutions and hospitals but their performance is untested at scale. Therefore it is not known if these systems are adaptable for global populations [84].

[5]Relative risk reduction (RRR) is the reduction in risk in the intervention group relative to the reduction in risk in the control group.

Study (Year)	Description	Key findings
Hodgetts [72] (2002)	Review of 118 cardiac arrest to develop MEWS for MET calling criteria (UK).	Major indicators of cardiac arrest identified to be breathing rate, oxygen saturation, temperature, systolic blood pressure. MEWS had 52% sensitivity, 99% specificity.
Ball[3] (2003)	Before-and-after study of CCO on post-critical care patients, over a 2-year period (UK).	6.8% increase in survival-to-discharge with CCO. 6.4% decrease in re-admission to critical care unit after CCO intervention.
Kenward [85] (2004)	Before-and-after study of MET calls (UK).	No significant reduction in mortality or cardiac arrest. Reason could be MET activation failure.
Bellomo [8] (2004)	Before-and-after study of MET over an 8 month period (Australia) .	57.8% RRR in adverse events including respiratory failure, renal, stroke. 44.4% RRR in emergency ICU RRR, 36.6% RRR in post-operative death, decreases from 23.8 to 19.8 days in LOS.
Hillman [70] (2005)	MERIT. A multi-centre RCT of 12 MET and 11 control hospitals (Australia).	An increase of 3.1 (control) vs. 8.7 (MET) per 1000 admissions in emergency calls. No significant difference in control and MET groups in combined or individual cardiac arrest, emergency ICU, or unexpected death.
Gao [60] (2007)	Retrospective multi-centre time interrupted study of 79 CCOs (UK).	Between 1996 and 2004, reduction in CPR with CCO intervention, but no impact on mortality. Inconclusive evidence of CCO impact.
Chen [28] (2008)	Cluster-randomised trial of 24 hospitals using MET (Australia).	77% patients had at least one vital sign observation missing before event. MET improves documentation of vital sign especially of breathing rate and systolic blood pressure.
Chen [26] (2009)	Retrospective analysis of 3700 MET calls during the MERIT RCT (Australia).	A decrease of 2.2 per 10,000 cardiac arrests, and a decrease of 0.94 per 10,000 unexpected deaths with an increase of 10% in MET calls.
Harrison [68] (2010)	Cohort studyof 52 hospitals using CCO (UK).	Effect of CCO on 2203 patients before, and on 5924 patients after critical care inconclusive.

Study (Year)	Description	Key findings
Trinkle [164] (2011)	Retrospective study of an RRT using MET.	Failure to call MET in 22.8% cases. increased emergency ICU risk as result (34.4% v 22.5%).
Moon [108] (2011)	Retrospective before-and-after analysis of CCO using MEWS over 8 years in one hospital (UK).	Reduction in cardiac arrest (from 0.4% to 0.2%), emergency ICU (from 3% to 2%), in-hospital mortality (from 52% to 42%), during "before" and "after" phases.
Laursens [93] (2011)	Before-and-after study of MET over a 4-year period (Australia).	MET reduced mortality (RRR 24.2%), emergency ICU (RRR 21.4%), and cardiac arrest (RRR 45.%).
Prytherch [128] (2010)	Analysis of ViEWS (UK).	AUROC of 88.8% for prediction of in-hospital mortality in 24 hours of last vital sign observation is better than that of 33 EWS systems.
Opio [118] (2013)	Validation of ViEWS on 844 acute hospital patients (Uganda).	ViEWS has an AUROC of 88.6% for prediction of death within 24 hours of hospital admission for a resource-limited population.
Tarassenko [158] (2011)	Analysis of CEWS. 64,622 hours of data from 3 vital-sign datasets from US and UK.	Development of a centile-based EWS. Outcome-independent detection of deterioration based on vital signs. Clinical trial on-going.

Table 1.10: A summary of recent studies on rapid response systems (continued from previous page).

1.4.1 Evidence-Based EWS

The EWS systems discussed so far are empirical in nature. The ranges of vital signs and the corresponding scores are set heuristically. Most recently EWS systems are making the move from empirical to data-driven or evidence based systems. In 2007, NICE recommended that an EWS system should be used in all acute care hospitals in the UK for the timely detection of patient deterioration and escalation of care [21], and that such a system should monitor vital signs including the heart rate, breathing rate, systolic blood pressure, temperature, oxygen saturation, and the consciousness level. As a result EWS systems were deployed in many UK hospitals [79]. Since then the Royal College of Physicians has proposed the establishment of a standard EWS system to be used across the NHS, to be called NEWS.

An evidence-based EWS was recently developed by Prytherch et al. [128]. Here 33

existing EWS were compared and the system with the highest AUROC was selected and called ViEWS. Vital signs were weighted according to a data driven "trial and error" process, although the score ranges (0,1,2,3) were still selected empirically. An additional supplemental oxygen feature was added to the scoring system. If the patient was on oxygen support (via an oxygen mask or cannula), an additional score of 2 was added to the aggregate score. The performance of the 33 EWS systems was assessed by means of the patient status 24 hours after the event. Although this system has been claimed to perform better than 33 EWS systems (with an AUROC of 88%), a fundamental design limitation exists. The patient status 24 hours after the event such as dead/alive is a hard outcome. Where the event in question is a phenomenon as wide-ranging as deterioration, the use of a hard outcome can be an over-simplification. To elaborate, it assumes that patients have exactly "normal physiology" and likely to survive regardless of intervention, or have exactly "abnormal" physiology and likely to die regardless of intervention. It does not account for a third category of patients who can have "abnormal" physiology but can survive with intervention. It has been argued that it is this third category on which the design of an intervention should focus [158], as this contains the preventable events, and is most likely to benefit from the intervention. ViEWS has very recently been proposed as the first NEWS [148], and is being recommended by the Royal College of Physicians for implementation at the national level in the UK [117] . The system is reported to be adaptable for patient populations in the developing world [118].

Another recent evidence based system has been developed by Tarassenko et al. [158]. This system is called the centile-based EWS (CEWS) and its design overcomes the limitation of ViEWS by its design not being based on hard patient outcomes. Instead, the statistical distribution of the vital signs is used to define the ranges of the vital-sign values corresponding to each score. The underlying hypothesis is that an EWS score that generates an alarm (3, in most systems) should correspond to the 1^{st} and 99^{th} centile of the cdf of the corresponding vital sign. Similarly, a score of 2 should correspond to the vital sign values between the 1^{st} and 5^{th} and the 90^{th} and 95^{th} centile of the cumulative distribution of the corresponding vital sign. Considering the breathing rate, for example, a score of 3 occurs when breathing rate ≤ 7 rpm or when breathing rate ≥ 34 rpm. The

system was derived from a database comprising 64,622 hours of continuous monitoring data collected from three clinical trials conducted in the USA and UK between 2004 and 2008. CEWS is the current system in use in the Oxford University Hospitals in Oxford, UK.

1.4.2 Discussion

The impact of EWS is controversial [84, 68, 127, 149, 70, 53]. Further, the comparison of studies on EWS systems is very difficult due to variations in patient characteristics, data collection protocols, and methodologies [60].

The application of a global EWS based intervention has not been tested [84]. The performance of the afferent arm of an intervention system has been cited as one explanation for the inconclusive impact [164]. Trinkle et al. [164] report that in a total of 576 adverse events, the afferent arm failed to call the MET in 22.8% of cases. As a result, the chances of an emergency ICU admission were actually increased (34.4% vs 22.5% for those without afferent arm failure). Junior doctors and nurses can be reluctant to call senior doctors hence the under-utilisation [173] and inconclusive performance of the EWS [84]. Further, many hospitals in the UK are still operating without an EWS system [103], and those that are, deliver service in different ways. Some have CCO or MET teams on duty 24 hours 7 days a week, while others between 12 and 23 hours, less than 12 hours, and on selected days. Therefore it has been difficult to study the impact [60]. In a time interrupted analysis [60] of 350,000 admissions in 172 units between 1996 and 2004 in England, a decrease in CPR rates during 24 hours prior to admission was seen, but there was no significant impact on mortality. However the use of an EWS was considered to lead to earlier intervention and therefore reduced CPR rate.

At least three systematic reviews have been conducted on the topic. In a systematic review [53] of 23 studies between 1996 to 2004, 14 studies from the UK, 7 from Australia, and 11 from the USA, were reviewed. It was found that improvement in outcomes exists but due to the many differences in service delivery, the evidence was inconclusive. Further, most studies were of before-and-after design and of inconsistent and poor methodology. Single-institution studies have shown a drop in cardiac arrests [149], but there has only

been one multi-centre RCT [70] and one single-centre RCT [127], in which no significant differences were found.

Another systematic review [173] also found that "weak evidence" of reduction in cardiac arrests and mortality is present in the literature. Eight studies were short-listed as being of reasonable quality and relevance, of which six were retrospective or before-and-after, and two were RCTs. Most of these studies reported impact on the usual outcomes (a) in-hospital mortality, (b) emergency ICU admission, and (c) cardiac arrest.

In-hospital mortality was reported by seven studies reviewed in [173]. The total relative risk of the EWS-based intervention as reported by the retrospective or before-and-after studies was 0.87 in favour of the EWS-based intervention compared to the control groups. For the two RCTs the total relative risk was 0.76 in favour of the EWS system, but these results were not statistically significant. Only four of the 8 studies reported the impact on emergency ICU admissions, including 3 before-and-after and retrospective studies and the one multi-centre RCT. For the first group, the relative risk was 0.84 in favour of the EWS-based intervention. In-hospital cardiac arrest was reported by five studies. For the before-and-after and retrospective studies the relative risk was 0.70 in favour of the EWS systems, and for the one multi-centre RCT it was 0.94 in favour of the EWS system. However there was significant heterogeneity in the studies, due to which the evidence in favour of EWS systems was considered weak. Overall a decrease in mortality rates and the incidence of cardiac arrests was associated with EWS systems.

Gao et al. [61] identified 25 distinct track-and-trigger systems reported in 36 studies, which were very difficult to compare due to the heterogeneity of the studies. They also reported that evidence of "reliability, validity and utility" was weak. It was found that poor selection of trigger thresholds may be responsible for the low sensitivity of the EWS systems reviewed.

Despite mixed opinions in the literature, the Institute of Health's 100,000 lives campaign [12] recommended the inclusion of an EWS system in hospitals. It was further recommended by the Joint Commission's 2008 Patient Safety Goals [40]. A recent consensus conference of healthcare experts and researchers from around the world unanimously agreed that the vital signs should always be monitored in all hospitalised patients

because they can identify patient deterioration [48]. There was further consensus that the features of a patient monitoring system should have the following desirable features: evidence-based, automated, continuous, multi-channel, sensitive and specific.

Automated patient monitoring is the main focus of this thesis. Signal processing and machine learning methods are applied to the problem of detecting deterioration in patients outside the ICU. Our focus is on patients who are recovering from an initial medical or surgical ICU admission in a step-down unit (SDU), and on patients who are recovering after a major surgery. Both these patient populations are considered "high-risk" and have a high likelihood of being re-admitted to the ICU. Investigating deterioration patterns and constructing models for detecting deterioration in these two populations should contribute to the existing understanding of deterioration in vital signs, and the detection thererof.

1.5 Thesis Overview

In this chapter, the clinical need for intelligent patient monitoring systems has been established. Evidence from the literature suggests that premonitory signs of deterioration are present in vital signs, that detection of abnormal vital signs can help avoid adverse events, and that many of the adverse events occur on wards outside the ICU. It was shown that EWS-based interventions can enable timely detection of patient deterioration, and further that there is a need for automated, intelligent patient monitoring systems. The remainder of this thesis focuses on the detection of deterioration in high-risk patients recovering from ICU admission in the SDU and from surgery on the general ward, as shown in Figure 1.1.

In the next chapter, algorithms for automated patient monitoring systems are discussed. Non-invasive monitoring of vital signs is introduced, and the current state-of-the-art in automated patient monitoring is described. In Chapter 3, a dataset comprising continuously acquired vital-sign data from patients recovering in an SDU ward is analysed. First, a method is presented that allows for detection of deterioration in such patients without prior knowledge of their vital signs as being "normal"or "abnormal". Construction of a model of "normality" is described.

For this dataset, retrospective labelling of "abnormal" vital signs corresponding to

episodes of deterioration was performed by clinical experts. In Chapter 4, the knowledge of labels is incorporated into the model. First, statistical methods are investigated to estimate the "normal" and "abnormal" data. Then, a one-class classification approach is taken and vital-sign data are classified into "normal" and "abnormal" classes using a probabilistic thresholding technique.

In Chapter 5, the problem of detecting deterioration is assessed from the point of view of two-class classification in order to improve the classification of bradycardia and bradypnoea-related events. The performance of standard two-class classifiers is compared with that of one-class classifiers and with that of a proposed "hybrid" method. It is shown that this proposed method performs better than the standard methods.

In Chapter 6 the focus shifts towards explicit modelling of vital-sign "abnormality". A method is proposed which not only allows the detection of deterioration, but also indicates the type of vital-sign abnormality associated with the deterioration.

In Chapters 4, 5, and 6, the problem of detection of deterioration in vital signs is considered as a classification problem. Vital-sign data belonging to an adverse event, or a period of deterioration, is considered as an example of "abnormal" data. A "population-based" modelling approach is taken and methods are assessed according to their ability to assign vital-sign data to either the "normal" or the "abnormal" class.

In Chapter 7, the second dataset used in this thesis is introduced. This dataset comprises patients recovering from upper gastrointestinal cancer surgery. Both datasets comprise high-risk populations who are vulnerable to re-admission to the ICU. However, unlike the first dataset, in this case the vital-sign data are not collected continuously. Instead manual observations are made every 4 hours on average. A comparison of the two datasets is performed in order to assess the suitability of models developed in Chapters 5 and 6 to this population.

As the patient monitoring frequency is comparatively very low in the upper gastrointestinal ward, a limited amount of data is available per patient prior to the adverse event. In Chapter 8 the issue of limited per-patient data is addressed. It is argued that in wards where there are very little data recorded prior to an adverse event, a patient-based modelling approach is more feasible than a population-based approach.

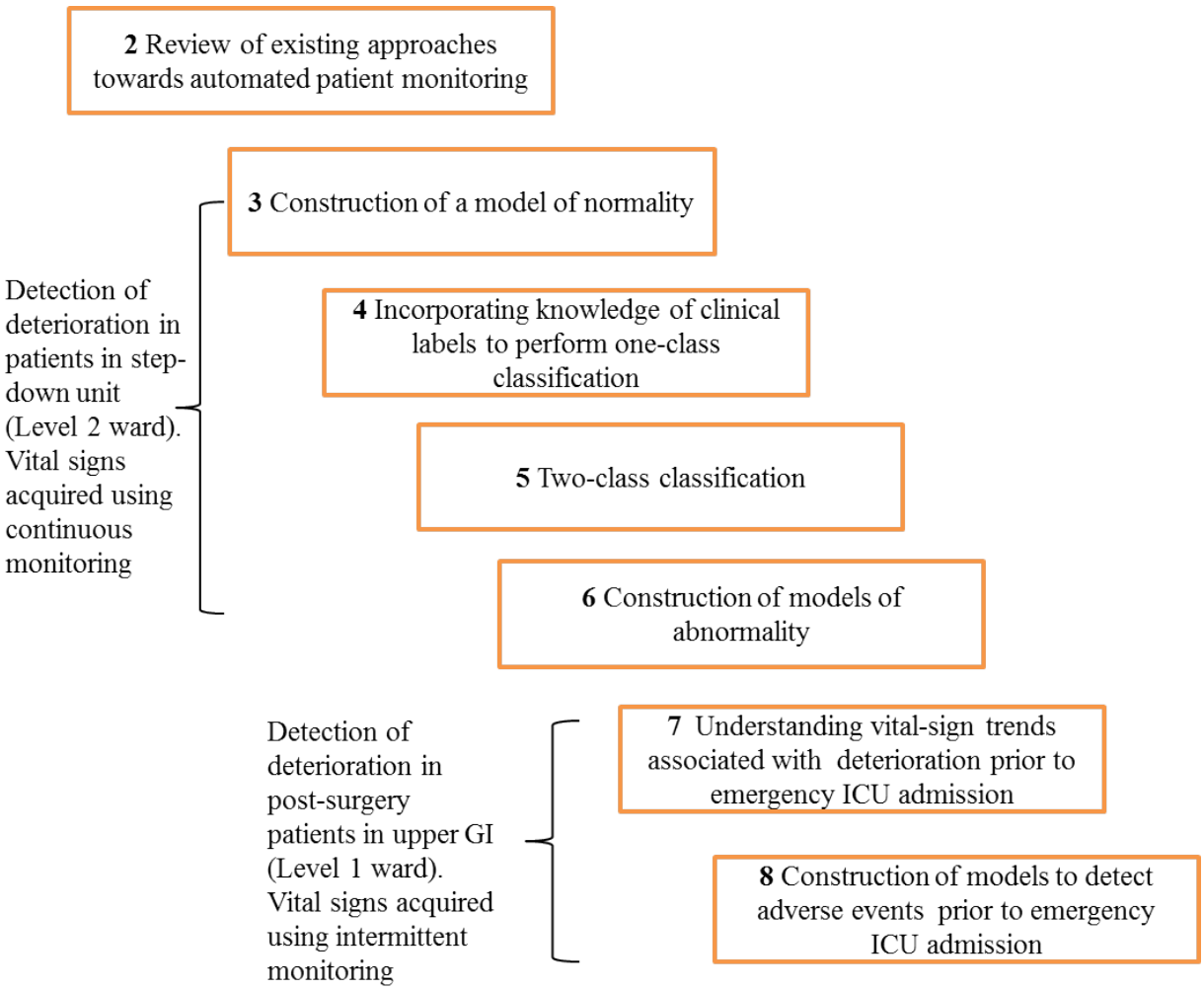


Figure 1.1: Overview of work presented in chapters 2 to 8 of this thesis.

In the final chapter the key findings of this thesis are summarized. Potential future work areas are identified including an ultimate move towards patient-specific modelling.

Chapter 2

Automated Detection of Deterioration using Non-invasive Vital-Sign Monitoring

As discussed in Chapter 1 there is consensus in the literature that the vital signs of patients in the hospital should be monitored [48]. Traditionally the vital signs are recorded by manual observations made by nurses. For instance the nurse counts the patient's pulse for a minute to estimate the heart rate, or counts chest movements to estimate the breathing rate. Since the late 1980s and early 1990s [152] this approach has been increasingly replaced by automated monitors (such as the one shown in Figure 2.1) which enable the continuous recording of vital signs. In this chapter the current techniques for monitoring of patient vital signs are introduced.

Secondly, the literature on alarm algorithms is discussed. Alarms are built-in in patient monitoring systems and are designed to discriminate between (some definition of) "normal" and "abnormal" vital signs. An alarm is generated if an "abnormality" in the vital signs is detected.

2.1 Non-invasive Monitoring of Vital Signs

Vital signs act as physiological indicators of patient condition, several of which are recorded routinely for the duration of a patient's hospital stay. Routinely monitored



Figure 2.1: A Phillips patient monitor performing continuous monitoring of the blood pressure and blood oxygen saturation of a patient in the upper GI ward of the Oxford Cancer Hospital (part of the Oxford University Hospitals NHS Trust Oxford, UK). The heart rate and breathing rate are not being recorded at the time.

vital signs include the heart rate, breathing rate, blood oxygen saturation, arterial blood pressure, and temperature. In the following section it is explained how these vital signs can be monitored non-invasively.

2.1.1 Heart Rate

Heart rate (HR) is an indicator of heart activity. It estimates the number of heart muscle contractions that have taken place per minute. Each heart beat comprises the systolic motion (contraction of the ventricular chambers of the heart), and the diastolic motion (relaxation of the ventricular chambers). The HR is recorded by the electrocardiogram (ECG), which is the electrical potential between two points on the chest, observed using electrodes, and the resultant waveform generated is used to derive HR. The QRS complex shown in Figure 2.2 corresponds to the ventricular depolarization. The R-peak can be detected using, for example, the Pan-Tompkins peak detection algorithm (or variations thereof) [121], and the time for one heart cycle is calculated as being the difference between two consecutive R-peaks. The reciprocal of this number is used to estimate the HR, which is recorded in beats per minute (bpm).

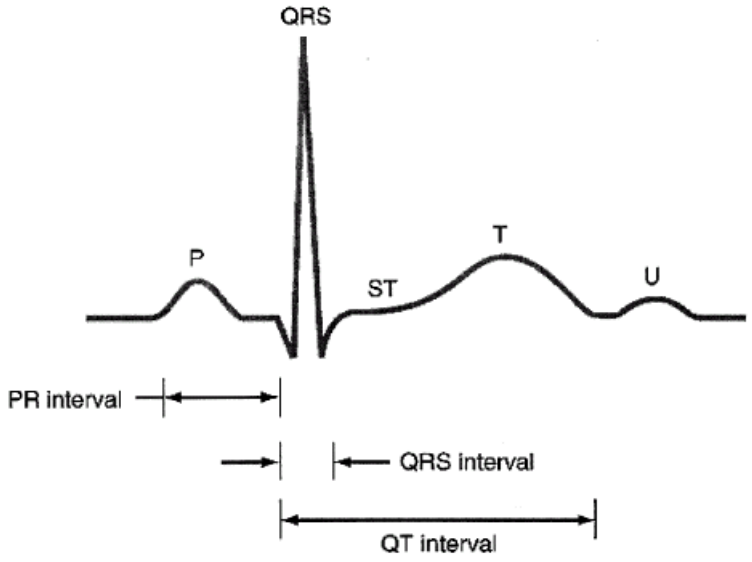


Figure 2.2: A typical ECG signal, recorded in seconds. The triangular QRS complex is shown.

2.1.2 Breathing Rate

Breathing Rate (BR) provides a measure of the number of breaths taken in one minute. As a person breathes in and out the chest volume increases and decreases, respectively. During respiration (or breathing), air moves into the lungs, and, being a poor conductor of electricity, it increases the electrical impedance of the chest cavity. This change in impedance is measured using impedance plethysmography (IP), which involves placing ECG electrodes on the chest and passing a small AC current (of 10-100 kHz frequency) through them. Plethysmography refers to measuring volume changes in a body [1]. As more air flows into the body with inspiration, the resistivity, and consequently, the impedance, change with changing thoracic volume. The voltage sensed via the electrodes is proportional to this changing impedance. The time difference between successive peaks in the resulting periodic signal is used to estimate the breathing rate, which is recorded in respirations per minute (rpm).

2.1.3 Blood Oxygen Saturation

Blood oxygen saturation is another vital sign routinely measured in hospital patients. Pulse oximetry is a standard non-invasive technique for measuring the saturation of oxygen (O_2) in the blood. It makes use of the differential light-absorbing properties of blood at

different wavelengths to estimate relative concentrations of oxygenated and deoxygenated haemoglobin (Hb) [35], where Hb is the oxygen-carrying component of blood. A standard pulse oximeter consists of an electronic probe attached to the finger or ear. Light pulsed at two different wavelengths is shone through the finger, and the amount of light received by a photodiode embedded inside the probe is measured. The assumption in this technique is that, while light is attenuated by arterial, venous, and tissue components, the *increase* in attenuation is caused by the pulsatile flow of blood in the arteries alone. Thus when the DC component of the received signal is removed, the remaining pulsatile signal, called the photoplethysmogram (PPG), corresponds to the cardiac-synchronous arterial blood flow, from which the oxygen saturation level is estimated [177]. Arterial oxygen saturation is recorded as a percentage and is abbreviated as SpO_2.

2.1.4 Blood Pressure

Flow of blood through the body is made possible by the pressure exerted on the walls of the arteries as the heart pumps blood into them. This pressure reaches a maximum (systolic blood pressure, SBP) during systole, and reaches a minimum (diastolic blood pressure, DBP) during diastole. It is typically measured non-invasively using a (manual or digital) sphygmomanometer. This device has an electric pump connected to an inflatable cuff which is placed around the patient's arm. The pressure in the cuff is raised to a level above SBP, at which point the artery is completely occluded. The pressure is then gradually released. As the pressure on the artery decreases, the arterial walls expand with the cyclical in-flow of blood, which causes oscillations in the cuff pressure. These oscillations continue until the cuff pressure drops below DBP. Thus by detecting the oscillations, using a pressure transducer built into the cuff, the SBP and DBP may be estimated. The point at which the oscillations reach their largest magnitude is regarded as the mean arterial pressure (MAP). This is found to be as $\frac{1}{3}(\text{SBP})+\frac{2}{3}(\text{DBP})$. The BP is calculated in units of mmHg. Other than the MAP, another measure of blood pressure is the arithmetic average of the SBP and DBP, called the systolic diastolic average (SDA). The SDA is computed as $\frac{1}{2}(\text{SBP})+\frac{1}{2}(\text{DBP})$.

2.2 Automatic Alarm Generation

Patient monitoring systems have built-in alarms which are designed to be activated when the alarm algorithm detects an abnormal change in patient condition [78]. However due to the high rate of false alarms, these systems are not considered particularly helpful by hospital staff [109]. In fact the high false alarm rate associated with existing patient monitors often leads to the "desensitisation" of the staff, i.e. staff learn to ignore the alarms which in turn increases the risk of missing true alarms. According to one study, an alarm was generated almost every half an hour in the ICU [23]. 3,188 alarms generated during a total of 1,971 hours of monitoring were studied. It was found that up to 72% of alarms were ignored, and that the false alarm rate was 42%. Furthermore, many (12.3%) of the false alarms were generated due to technical problems in the monitoring systems. Similar results were found in an earlier study [95]. Here 2,716 alarms generated in a total of 928 hours of monitoring inside the paediatric ICU were studied and a 68% false alarm rate was found. Tsien et al. reported an 86% false alarm rate in the ICU [166]. In a review of false alarm rates in the ICU it was found that while the existing monitoring systems are reasonably sensitive (i.e. able to detect clinically significant events), the majority of them had poor specificity (or alternatively, a high false alarm rate) [22]. A false alarm rate of up to 90% has been reported [78].

There can be many reasons for a false alarm, including technical problems and unsuitable alarm algorithms [78]. An example of a false alarm triggered due to a technical problem is the low SpO_2 alarm when the SpO_2 probe falls off, or a motion artefact due to movement of the patient. In fact SpO_2 dropouts due to poor contact or connection are considered to be a major source of false alarms [22, 95]. Unsuitable alarm algorithms can have poorly selected alarm thresholds that can also cause the monitoring system to alert unnecessarily.

Several models have been proposed in the literature to improve patient monitoring and reduce false alarm rates (as described below in Sections 2.2.1 and 2.3.2), however hardly any have been validated for clinical implementation and been commercialized. Many of these proposed models are based on alarm algorithms that exploit the statistical properties of vital signs. More recently the trend has shifted towards knowledge-based and

machine learning-based algorithms. Further, within these categories, algorithms are designed for either single-channel or multiple-channel alarms. Single-channel alarms have well-established drawbacks when compared with multiple-channel alarms. Extensive literature exists on recommendations for improvement in all categories, as is discussed in the following sections.

2.2.1 Statistical Methods

Some of the simplest alarm algorithms use filtering approaches such as a median filter [45] and a moving average filter to detect significant changes or outliers in the signal over a window of time. Here the choice of the time window is crucial and can be complicated. For instance, a vital sign time series may have both slow moving trends which can be captured with a large time window, and fast moving trends which would be missed out if the window size was too large. Makivirta et al. [100] described one way to overcome this issue by two moving median filters for ICU data - a slow moving filter with a window length of 15 minutes, and a fast moving one with a window length of 2.5 minutes. They reported a reduction in false alarm rates by two thirds. However, they found that some alarms were still missed out or detected too late.

Control charts such as cumulative sums (Cusums) which are used in control systems to detect process change points have also been adapted for detecting changes in vital sign time series [24, 178]. The main drawback of statistical process control techniques is that, unlike for a control process, it is very difficult to define "in-control" and "out-of-control" limits for vital sign data. As a result, while such methods are useful for pre-processing tasks such as noise removal, they are not ideally suited to the task of monitoring vital signs [78].

Other statistical methods have been proposed which aim to detect the pattern or the trend in vital-sign time series. Several trend based methods have been suggested [24, 25, 142, 178, 119]. Trend-based systems depend on a user-defined description of "trend". For instance, Schoeberg et al. [142] describe an algorithm in which "trend" is described as user-defined physiological change, such as a change in the average heart rate over a fixed time period. At each time interval, the change is assigned a score, and if this

score is above a certain threshold value, an alarm is generated. Charbonnier et al. [25] proposed another trend-based system. Here too, trend is defined as the rate of change in a physiological variable, i.e the slope of the variable. A linear function is fitted to segmented sections of the signal. If the linear approximation fails to describe the segment, a change is considered to have occurred, and a new linear function is fitted to the segment. The algorithm has several tuning parameters that must be pre-set by the user. Another trend-based change detection algorithm was proposed by Yang et al. [178]. Here nine different types of trends were identified in the heart rate signal of anaesthetised children, including a long or short, and abrupt or gradual, increase, decrease, or stable state.

Trend-based methods therefore depend on the definition of trend which is often variable [106] . Further most of these methods involve many variables and tuning parameters which are again user-defined and heuristically selected. An additional limitation is that these approaches impose pre-defined expected behaviour and "hard" thresholds which may not be appropriate for physiological data [58]. Furthermore these methods are uni-dimensional and are designed for single-channel alarms. In the case of vital-sign data, a deterioration may not necessarily manifest itself as a large change in a single vital sign, but rather manifest itself through smaller changes in multiple vital signs. Due to this, the utility of single-channel statistical methods for patient monitoring is limited [152].

Clifford et al. [32] developed an artefact-removing algorithm based on data from two channels. This method estimates an abnormality index from features extracted from the arterial blood pressure signal[1], and uses a set of thresholds and rules to classify ECG-based single-channel alarms. The algorithm was tested using 800 hours of data which had been retrospectively deemed to have 25% false alarms associated with them, and reported all false alarms to have been removed by their algorithm.

A multi-channel system based on robust sensor fusion was proposed by Feldman et al. [54]. Here information from the ECG channel, SpO_2, and an intra-arterial catheter was combined to perform heart rate estimation. The model was tested on 85 episodes of monitoring in the ICU and operating room, and it was found that the heart rate estimated from the combined sources was closer to the reference heart rate than that obtained from

[1]This data was collected from the ICU, where an intra-arterial catheter is used to measure the arterial blood pressure, which is different from the method described earlier in Section 2.1.4.

any individual source.

2.3 Data Fusion

In the context of patient monitoring, the combination of data from multiple channels to provide an aggregate patient status index may be referred to as the *data fusion* approach [66]. Data fusion for alarm generation can be split into two main topics: knowledge-based alarm systems, and machine learning systems. Knowledge-based systems combine the expert knowledge from clinical experts to define a set of rules or criteria according to which an alarm system is designed. Machine learning systems, on the other hand, avoid this dependency on rules. The classifier is trained to generate alarms (or alerts) by learning the characteristics of a set of example training data.

2.3.1 Knowledge-Based Methods

Knowledge-based, or expert-based algorithms emulate the human decision-making process by having recourse to a "knowledge base" of comprehensive information relevant to the medical problem [109] and a set of "if-then" rules. Several knowledge-based systems have been developed for patient monitoring systems, ranging from the purely rule-based, to those using fuzzy logic to implement "softer" rules [82, 142, 154]. For instance, a knowledge-based alarm algorithm for monitoring cardiac surgery patients was described by Koski et al. [88]. The system organised information from the knowledge base in to a decision tree and was reported to show encouraging results but was not developed further.

Laramee et al. [92] proposed an expert based system in which the rules were determined manually by clinical experts, and information from multiple vital signs was integrated to enhance the performance of the alarm algorithm. Results of testing this method on clinical data were not reported.

A knowledge-based system for monitoring post-operative cardiac surgery patients in the ICU was developed by Sukuvaara et al. [154]. Two clinical experts used the trends in a total of six vital signs to create a set of 87 "if-then" rules, such that the outcome of each rule would cause or prevent an alarm from being generated. This system was validated in a two-phase trial on ICU patients, and was reported to give 100 % sensitivity and 73%

specificity of alarms.

Both Oberli et al. [116], and Kannathal et al. [82] used fuzzy logic to fuse data from multiple vital-sign channels. Instead of assigning a rigid category (either "normal" or "abnormal") to data, fuzzy logic determines the degree of membership of given data to both categories. Clinical experts were asked to define rules based on the different combinations of scores assigned according to the fuzzy logic rules. The system proposed by Oberli et al. [116] integrated information from seven physiological variables and reported 60% true alarms of clinical relevance, compared to 45% of the standard monitor alarms. Kannathal et al. performed data fusion using the heart rate, breathing rate, blood pressure and SpO_2, and reported over 90% specificity and sensitivity on limited data; however no validation of the method against standard methods was reported [67].

In another study [174] the use of a fuzzy logic system was proposed to improve the monitoring of pre-term infants in the ICU. Information from eight vital signs was used in combination with fuzzy rules and the method was shown to be able to distinguish real alarms from movement artefact. Recently another fuzzy logic system was proposed for monitoring of coronary bypass surgery patients by Stevens et al. [152]. This system couples fuzzy logic with Bayesian theory to determine the probability of complications after coronary bypass surgery, as well as the associated risk factors. The model combines the heart rate, SpO_2, and blood pressure, as well as the medical history of the patient. The system reported a reduction of 52.1% in false alarm rates. However these results were based on a small dataset comprising 7 hours of data acquired from 4 patients.

Developing a comprehensive set of rules is a cumbersome task (and may even be intractable), and requires expert medical knowledge and experience. Further it is hard to decide how much information there should be in the knowledge base. Factors such as these limit the popularity of knowledge-based systems.

2.3.2 Machine Learning

Machine learning methods usually attempt to discriminate between "normal" and "abnormal" data. However, unlike knowledge-based approaches, in machine learning the model can be learned from the data itself. Machine-learning based approaches are broadly

grouped into two categories: supervised and unsupervised. Unsupervised methods learn from data without prior examples from the "normal" and "abnormal" classes. For instance clustering methods such as k-means and nearest-neighbour learn a pre-defined number of clusters from the training data, and optimise the model by reducing a cost function [51]. In contrast, in a typical supervised machine-learning approach, examples of "normal" and "abnormal" data are used to train a classifier, in a supervised learning manner. The training process requires sufficient examples from both classes. This poses a problem when dealing with critical-care data, since there are relatively few adverse-events, and so "abnormal" data are sparse, as compared to the number of "normal" data.

Many popular supervised classifiers including neural networks, support vector machines, and genetic algorithms have been used for classification of vital-sign data.

Neural Networks

A neural network can take different forms of architecture, but a simple neural network such as the multi-layer perceptron (MLP) comprises input and output neurons or nodes which are typically connected to each other via a layer of hidden nodes. The working of an MLP is discussed in detail in Chapter 5. There are several applications of neural networks for vital sign monitoring. Tsien et al. [165] used a neural network to perform a case study of monitoring vital signs in the paediatric ICU. The system was designed to detect true alarms that were clinically relevant. The heart rate, breathing rate, oxygen saturation, and blood pressure recordings were annotated in the pre-processing stage. Training, test, and validation sets were created which contained 70%, 21% and 9% of the data, respectively. The neural network achieved an AUROC[2] of 98.8% when applied to the case study data. A drawback of neural networks is that the relationship between the input and output data may not be explicit [56].

Tsein et al. [167], using data retrospectively annotated by clinical experts, applied supervised classification methods, such as logistic regression and decision trees, to classify artefactual data. They found that such classifiers performed better when the training data were multi-dimensional rather than uni-dimensional.

[2]Area under Receiver Operator Curve, described in Section 1.3.3.

Support Vector Machines

Support vector machines (SVM) seek to find an optimal hyperplane in high-dimensional space that is able to separate "normal" and "abnormal" data. Optimal values are sought for a "slack variable" ζ which controls the degree of over-fitting to training data, and C which controls how many misclassifications are allowed during training. SVMs are described in detail in Chapter 5. Clifton et al. [37] applied a one-class SVM to vital-sign data acquired from 476 patients in a large clinical trial in the emergency department. As the data were continuously acquired, an alert was considered to have been generated if the SVM estimated test data for a period of 4 out of 5 minutes to be "abnormal" with respect to the one-class model. The performance of the SVM was compared with that of standard early warning score systems used in the ward at the time of the trial, and it was shown that the SVM had a better "sensitivity" (85%) these methods, by detecting deterioration in a maximum of 29 out of the 34 cases of physiological deterioration. However, the SVM generated alerts in 58 of the 217 cases where no physiological deterioration was expected, resulting in a false-alarm rate of 27% for the SVM, which was higher than that of an early warning score (13%).

Bayesian Networks

Bayesian (belief) networks[3] (BN) are based on Bayes' theorem and are designed to calculate, for instance, the probability of adverse events in the hospital. They consist of nodes which represent variables and branches which define the relationships between variables. Laursen et al. [94] developed a BN for detection of cardiovascular events. Arterial blood pressure and central venous pressure were monitored. As a new vital-sign observation is made the prior belief of the network about the probability of cardiovascular event is updated. Crump et al. [42] proposed a BN for application in the ICU. Their model combined a rule-based algorithm with a BN to predict patient status in the ICU. The model was tested on data from 16 patients and was found to have a 38% false positive rate. It was also shown that a BN modelled using patient-specific norms could be different from one modelled using population-specific baselines. BNs are increasingly gaining ground in

[3]Bayesian networks are a branch of probabilistic graphical models, with directed links.

healthcare applications [96] including patient-specific modelling [91, 43, 135, 13]. Bayesian analysis can involve complicated integrals which are difficult to solve and the alternative approximation methods can be computation-intensive. As BNs have to learn the conditional probabilities defining the relationship between many variables in the network, large amounts of training data are needed [78].

A popular supervised approach is that of *novelty detection*, which has been used for applications such as fault detection in engines, hand-written digit recognition, target detection using radars, and for detecting masses in mammograms, among many others [67, 101, 102].

One-Class Classification (Novelty Detection)

Novelty detection can be considered to be a one-class classification technique, in which only the "normal" class is modelled. The terms *novelty detection*, *anomaly detection, outlier detection, and one-class classification* have been used interchangeably [134, 111]. The classifier evaluates test data with respect to a model constructed for the "normal" class, and if data deviate significantly from that model, they are deemed to be "abnormal". Typically, a threshold is introduced to define the boundary of "normality" [101]. A number of studies have investigated the application of novelty detection to patient monitoring in the ICU and step-down units (SDUs), and have reported encouraging results [130, 34, 77, 67, 69].

The approach used for novelty detection depends on the data in question. A simple statistical technique involves creating a "model of normality" by constructing the probability density function (pdf) for the "normal" data, and then calculating the probability that new test data belong to that distribution [101]. A threshold on the probability density can indicate how "novel" the test data are with respect to the model.

Markou et al. [101, 102] present a comprehensive review of different approaches used in novelty detection. Probabilistic approaches such as Gaussian Mixture Models (GMM) have been used for identification of masses in mammograms [155], epileptic seizure studies [136], engine fault detection [38, 33]. Non-parametric approaches such as those based on Parzen windows or nearest-neighbour are better suited to problems where no assumptions

may be made about the underlying distribution of the "normal" data. Parzen windows learn the form of the distribution from the training data. A Gaussian kernel is placed on each training sample, and the local estimates provided by each kernel are combined to estimate the pdf of the training data. In the nearest-neighbour approach, the Euclidean distance between a test sample and its nearest sample in the training set is calculated. Tax et al. [162] have shown that for sparse datasets, a Euclidean distance-based nearest neighbour approach performs better at identifying "novelty" than a probabilistic approach such as the GMM. Although there have been many studies using different novelty detection approaches, comparative analyses of different techniques, and resulting guidelines, are few [102]. It is therefore difficult to assess how each technique performs with different data types.

Much existing work [161, 159, 38, 69, 175] on vital-sign monitoring has used one-class classification to form a model of normal physiology, and then detected deviations away from that model as being "abnormal".

Two-Class Classification

In order to perform two-class classification, class labels must be available for both classes. Manual labelling of vital-sign data may be performed by clinical experts who would examine time-series plots of vital signs and determine which periods of data they deem to be sufficiently deviant from the normal range to be labelled as "abnormal". The periods of data not labelled "abnormal" may be considered as being "normal". As mentioned in the previous section, examples of "abnormal" data are often rare, and secondly, labelling of large datasets is a complicated and onerous task.

Given the difficulty in acquiring sufficiently large quantities of labelled patient data to adopt a two-class approach, there are very few examples of two-class classification approaches to the problem of patient vital-sign monitoring. Multi-class classification, in which each type of abnormality is explicitly modelled in addition to the model of normality, may also be performed in theory. However, this requires more data, and it could be challenging to scale up to datasets of higher dimensionality[4]; i.e., adding a new

[4]As more dimensions are added, the sparseness of samples in the data space increases. This is known as the curse of dimensionality [87].

vital sign would introduce further types of abnormality due to the extra vital sign that must be considered.

2.4 Discussion

In this chapter existing methods for automated detection of deterioration were described. Almost all of the systems discussed in sections 2.2.1 and 2.3.1, and most of the systems discussed in section 2.3.2 are designed for monitoring in the ICU setting, and have not been validated via clinical trials or been implemented at scale.

In the ICU, the monitoring regime is much more intensive and much more information about patient condition is recorded than is done in general wards or SDUs. Further, as the patients are non-ambulatory (and even sedated, in many cases) in the ICU, the recorded signal has less chance of containing artefacts or large periods of data dropouts.

In the remainder of this thesis, patient monitoring is considered in two environments where patients are recovering from a major event: (i) an SDU, and (ii) a post-surgery general ward. The SDU (or intermediate ward) of a hospital in the USA, from which vital-sign data was collected during a three-phase trial (explained in detail in the next chapter), is considered. This SDU is an intermediate level ward for patients coming out of the ICU, who require level 2 care (according to the definition in Table 1.1). As part of the trial, continuous monitoring was performed. The next three chapters focus on techniques to analyse continuously-acquired data from the SDU setting. It is investigated if the analysis proposed in this thesis can discriminate between "normal" data and that representing individual types of vital-sign abnormalities that are associated with deterioration in patients in the SDU, such as tachycardia, bradycardia, tachyapnea, apnoea, and hypoxia.

The upper gastrointestinal (upper GI) ward of a hospital in the UK, which is considered in the second half of this thesis, is a similar environment to the SDU of a US hospital in that, here too patients are recovering from a major event (which in the case of the upper GI ward, is cancer surgery). Further, as in the case of the SDU, patients in this ward are vulnerable to (re-)admission to the ICU under emergency conditions [39]. However, as this is a level 1 ward (see Table 1.1), which is a step lower than the level 2 SDU environment,

intermittent and manual patient monitoring is carried out, as opposed to the continuous, electronic monitoring carried out in the SDU. Nurses manually check and record the vital signs of a patient. Additionally, monitoring of the temperature using a tympanic thermometer is performed in the upper GI ward, whereas as continuous monitoring of temperature is very difficult [37], it was not performed in the SDU.

Therefore, while the patients in both SDU and the upper GI ward are at a high risk of deterioration leading to adverse events, the patient monitoring set-ups in the two settings is different. The aim of this thesis is to develop optimal alert-generation techniques based on the available vital-sign information acquired from these two environments.

Chapter 3

Constructing a Model of Normality

3.1 Introduction

In order to develop an automated system for monitoring patient vital-sign data, conventional multi-class classification techniques require sufficiently large quantities of data to represent both the class corresponding to "normal" patient physiology and the class corresponding to "abnormal" physiology. However the latter are typically difficult to obtain because examples of "abnormal" physiology are rare, relative to the amount of examples of "normal" physiology. This poses a limitation as adverse events are a rare occurrence outside the ICU. Further, there are many different modes of abnormality. Patients can be so different that a single model of abnormality may not be appropriate. This motivates the use of novelty detection, or one-class classification, in which a model of normality is constructed using "normal" data. Data lying within the model of normality are then considered "normal", and otherwise "abnormal".

In this chapter, a model of normality is constructed by estimation of the probability density function of the most "normal" vital-sign data. The model is trained and tested using separate datasets from a large dataset containing patient monitoring data. The dataset is introduced in section 3.2, where limitations of existing methods for labelling clinical datasets are also discussed. The use of the model of normality developed in this chapter to improve the class labels for one dataset is discussed in subsequent chapters.

3.2 The UPMC Study

A three-phase clinical trial was conducted at the Presbyterian Hospital, University of Pittsburgh Medical Center (UPMC), Indianapolis US, between November, 2006 and August, 2007. A population of 1,000 patients was monitored in a 24-bed step down unit (SDU) [75], which is a ward for patients coming out of the ICU, or those transferred from lower-acuity wards because they require higher-level care. The vital-sign data comprise measurements of heart rate (HR), breathing rate (BR), arterial oxygen saturation (SpO_2), and systolic-diastolic average blood pressure (SDA, which is the arithmetic mean of systolic and diastolic blood pressures. HR, BR, SpO_2 were acquired from Hewlett-Packard "Viridia 24" electronic bedside monitors, with a sampling rate of $T_s = 20$ seconds. A 3-lead ECG was used to estimate the HR, and impedance plethysmography (via the ECG electrodes) was used to estimate BR. SpO_2 was collected using pulse oximetry (using the M1191B Phillips Medical Systems, Germany). SDA was calculated from automated measurements made using a blood-pressure cuff at a minimum rate of approximately every 2 hours during the waking hours of the patient.

In Phase I, 18,692 hours of data were collected from 326 SDU patients, during a period of 8 weeks (see Table 3.1). The total number of hours of recording included data that may be recorded from a single vital sign or when data had been recorded for all vital signs [75]. Data were collected in a "blinded" fashion, and a panel of three clinical experts retrospectively reviewed the data. The Medical Emergency Team (MET) is the response team in hospitals in the USA, which is called when the limits shown in Table 3.2 are exceeded. Periods of data in Phase I that met the MET calling criteria were labelled as corresponding to adverse events. The labels in this dataset refer to the start-of-event and end-of-event labels annotated by the experts. A total of 112 adverse events were identified in a total of 59 patients. Details of the labelling process are presented in section 3.4.

In Phase II, the clinical staff was trained in the use of the Visensia (OBS Medical, Abingdon, UK) data-fusion monitor, and in Phase III, the monitors were used to collect a further 28,782 hours of data from 674 patients. The Phase III dataset was not retrospectively labelled by clinical experts.[1] The Phase I dataset is henceforth referred to as

[1]The reader is referred to [67] for further details of the nature and durations of the clinical trials.

	UPMC Phase 1 ($\mathcal{D}_1$)
No. patients	326
Age (mean±std)	58 ± 20
Sex (male)	58.7%
Discharge disposition	Home (65%), Lower-acuity ward (32.6%), Death (6%)
Data collected (hours)	18,692
No. labelled adverse events	112

Table 3.1: Patient demographics and data collected during of Phase 1 of the UPMC study.

	Lower Limit	Upper Limit
HR (bpm)	40	140
BR (rpm)	8	36
SBP (mmHg)	80	200
DBP(mmHg)	-	110
SpO_2(%)	85	-

Table 3.2: MET calling criteria used to identify periods of data to show clinical experts for labelling. Repeated from Table 1.7.

the $\mathcal{D}_1$ dataset.

3.3 Labelling of Vital-Sign Data

The success of classification depends on the accuracy of data labelling, but, for practical clinical datasets, this can be hard to achieve for three reasons: (1) The quantity of data acquired from continuous monitoring of a patient is typically so large that it is impractical for clinical experts to label data from all patients. In previous studies, clinical experts have labelled only those patients for whom known abnormalities occurred in practice [130, 160], or only those patients for whom simple uni-dimensional alerting criteria (e.g., heart rate $\geq$ 120 beats per minute) were identified during retrospective analysis [77, 36]. (2) Clinical labels are typically applied to intervals in the patient records, rather than to the (continuously-acquired) data that fall within those intervals. Thus, a period of "abnormality" may contain examples of "normality", and vice versa. (3) One cannot rely on the use of "hard" patient outcomes (e.g., alive or deceased at some time in the future, possibly after discharge) to provide class labels because patients who go on to have a negative outcome (such as an unexpected cardiac arrest, or an unplanned admission to the intensive care unit) could have periods of "normal" physiology, perhaps prior to their

Class	Clinical label	Description
C_2	"abnormal"	Contains data from dataset $\mathcal{D}_1$, belonging to periods considered by clinical experts to have "abnormal" patient condition.
C_1	"normal"	Contains data from dataset $\mathcal{D}_1$, except all data which belong to C_2.

Table 3.3: Description of classes C_1 and C_2 from dataset $\mathcal{D}_1$.

deterioration, and similarly periods of "abnormal" physiology prior to complete recovery.

3.4 Labelling of UPMC Vital-Sign Data

The $\mathcal{D}_1$ dataset was labelled by a panel of clinical experts [2] who determined retrospectively which periods of patient data exceeded standard "medical emergency team" (MET) calling criteria. The process of labelling was undertaken as follows: first an automatic check was made to select all data which exceeded the single-channel MET calling criteria (see Table 3.2). 564 such events were originally found in 18% (i.e., 59 out of 326) of the patients. It was noted that in some instances a series of alerts were generated due to vital-sign variability just above and below the MET threshold. Such instances were considered to be part of a single clinical event, and so were merged into a single event if the gap between two successive instances was of 30 minutes of less [67]. By doing so, individual "normal" data points occurring during the gap between "abnormal" instances also became part of the segment of data being considered to be "abnormal". Periods of patient data that exceeded the MET calling criteria for at least four minutes were shown to the panel of clinical experts, who subsequently determined which periods were due to artefact (such as a sensor becoming detached from the patient), and which appeared to be sufficiently abnormal to require actual MET intervention. Here, the latter are termed as having class labels C_2, corresponding to "abnormal" patient condition, and will refer to examples of "normal" patient condition as having class label C_1 (see Table 3.3). Validation of "normal" and "abnormal" events by clinical experts reduced the number of C_2 events to 112.

Figure 3.1 shows the distribution of vital signs during periods labelled C_2, and the

[2] The clinical experts were senior critical care medicine physicians (M.R.P).

distribution of each vital sign for patients that had no C_2 periods associated with them (and hence were assumed to be "normal" patients), which are here defined as being from class C_1. It may be seen that the distributions of data from periods labelled C_2 have heavier tails than the distributions of data from "normal" patients, because vital signs tend to take more extreme values during periods of abnormal physiology. However, it may be noted that there is significant overlap between the distributions of data from the two classes. This is because C_2 is not an exhaustive set of all "abnormal" patient condition in the data-set, because clinical experts were shown only those periods of data that exceeded MET calling criteria, from which they selected a subset as belonging to class C_2. Furthermore, a patient typically shows deterioration in fewer than the full number of vital signs, and so some class C_2 data will contain "normal" values for one vital sign even though another vital sign may be highly "abnormal" at the same time.

3.5 Constructing a Density-based Model of Normality

In this section a model of normality is developed using data from dataset $\mathcal{D}_1$, without incorporating knowledge of class labels. In subsequent chapters, the model of normality is used for the estimation and refinement of clinical labels.

In constructing the model of normality, it is assumed here that the data belong to an underlying distribution, and that the distribution may be estimated by density-based methods. A threshold on the probability density can indicate how "novel" the test data are with respect to the model. This is an example of a probabilistic approach to novelty detection, and can be divided into *parametric* and *non-parametric* approaches. Parametric approaches make assumptions about the form of the underlying distribution, and estimate distribution parameters from the data. For example, a simple Gaussian distribution may be assumed, with its mean and variance estimated using maximum likelihood. More complex distributions may be modelled using, for example, Gaussian mixture models. However, a large amount of training data is required to estimate the parameters accurately. [157] Non-parametric approaches such as kernel-based methods are better suited to problems where few or no assumptions may be made about the underlying distribution. Parzen windows is one such kernel estimator, which considers each datapoint as a

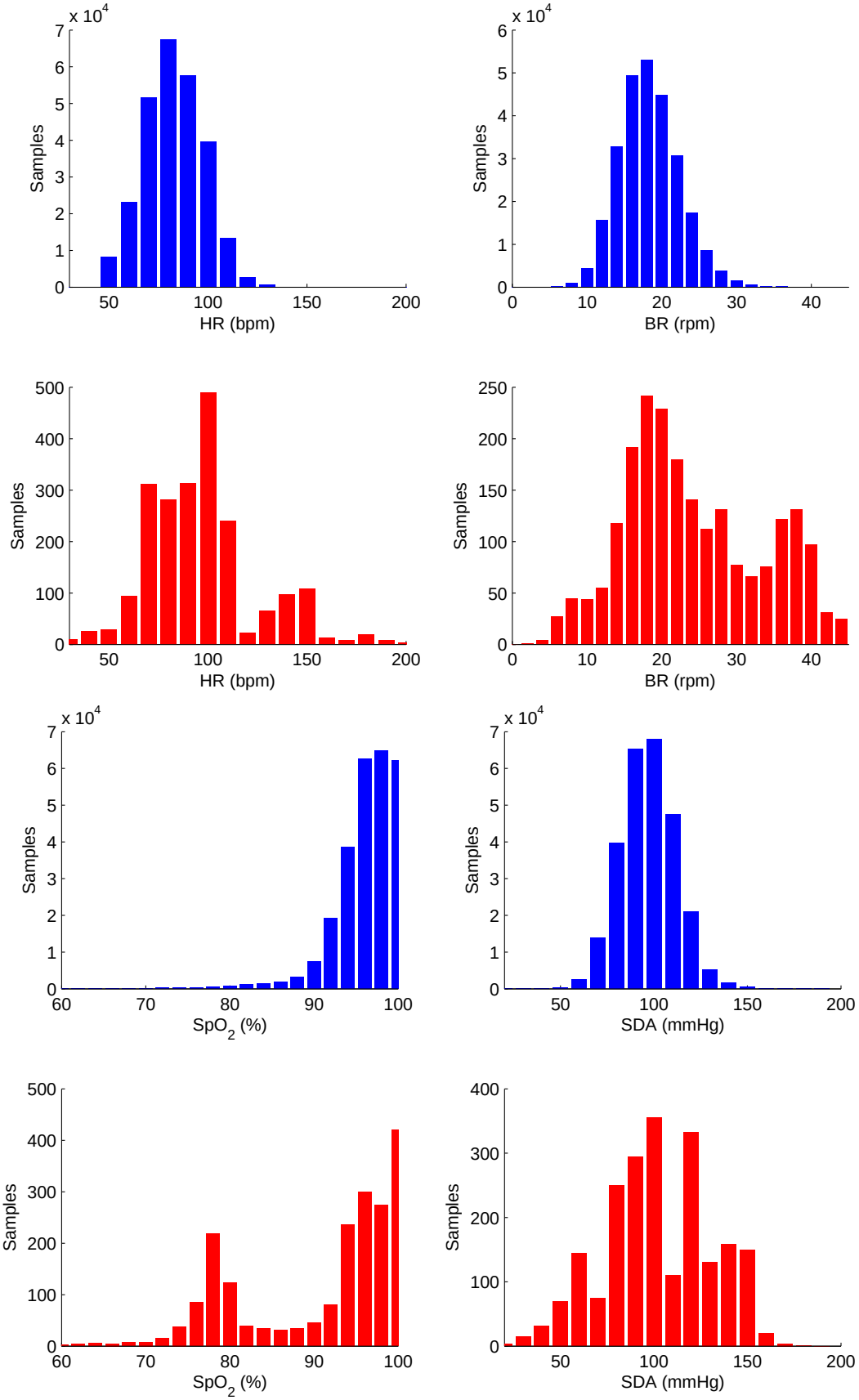


Figure 3.1: Frequency distribution of HR, BR, SpO_2, and SDA data. Data from periods labelled by the clinical experts as adverse events are shown in red, and data considered "normal" by the experts are shown in blue.

Gaussian kernel, and sums the contribution of each kernel. Other kernel-based methods include support vector machines but these are primarily used for two-class classification [157]. Making no assumption about the underlying form of the distribution of data in dataset $\mathcal{D}_1$, it was decided to use the Parzen windows approach for density estimation. Once the density distribution is estimated, a "novelty" score can be assigned to each new data point.

3.5.1 Data Pre-Processing

Prior to use of data in constructing a model of normality, pre-processing of vital-sign data must be performed. (1) Periods of data with gaps of 30 minutes or less in any one or more of the vital signs were removed by applying a zero-order hold (i.e. by holding values prior to the gap for the duration of the gap). For the SDA, zero-order hold was applied for up to 2 hours. (2) If there was a gap in any vital sign longer than the zero-order hold duration, the remaining vital signs belonging to that gap period (for which data may have been recorded) were also not used in constructing the model. (3) Data points lying outside the artefact rejection limits shown in Table 3.4 were removed [67]. (4) Due to the large number of data recorded from patients who did not have an adverse event, data were resampled to obtain one sample per minute, which is a reasonable sampling rate for physiological variables such as vital signs[3]. Application of this pre-processing to dataset $\mathcal{D}_1$ resulted in 2.6 $\times 10^5$ four-dimensional vectors (HR, BR, SpO_2, and SDA), as shown in Table 3.5. Of the 2.6 $\times 10^5$ data in dataset $\mathcal{D}_1$, 2,046 belong to class C_2 and 265,581 belong to class C_1 . Step (2) of this pre-processing particularly affected class C_2 because the number of periods with gaps in the data were much greater here than in class C_1, explaining the large difference in the number of data in C_1 and C_2. The highest number of data gaps were contributed by the SpO_2, corresponding to 30% of the total numbers of monitored hours, whereas HR and BR were not monitored 4.8% and 7.9% of the monitored hours, respectively [75].

[3]The sampling rate is considered reasonable for the analysis presented in this thesis. The same, however, may not be the case for instance in trend analysis, where it would be necessary to choose the sampling rate in view of the Nyquist criterion, in order not to incur aliasing.

	Lower Limit	Upper Limit
HR (bpm)	30	300
BR (rpm)	3	45
SpO_2 (%)	60	-
SDA (mmHg)	20	180

Table 3.4: Artefact rejection limits, decided by UPMC clinical experts. [67]

Class	Clinical label	# of data
C_2	"abnormal"	2,046
C_1	"normal"	265,581
	Total no. of data in $\mathcal{D}_1$	267,627

Table 3.5: Classes C_1 and C_2 in dataset $\mathcal{D}_1$ after pre-processing.

3.5.2 Normalisation and Clustering

The following zero-mean unit-variance transformation was applied to each measurement in dataset $\mathcal{D}_1$, in order to remove the dependency on the units of measurement for each vital sign, i.e.,

$$\hat{x} = \frac{x - \mu_g}{\sigma_g} \tag{3.1}$$

where μ_g and σ_g are the mean and standard deviation of the vital sign, and $\hat{x}$ is the normalised value of the original measurement x. From Figure 3.1, it may be seen that the distribution of SpO_2 is non-Gaussian. However, it may be noted here, that in applying Equation 3.1, the distribution of each vital sign is implicitely assumed to take Gaussian form. Table 3.6 shows the mean and standard deviations for each vital sign for dataset $\mathcal{D}_1$. Note that at this stage, class labels have been ignored.

For computational ease, the $\mathcal{D}_1$ dataset was reduced from R= 2.6 $\times 10^5$ vectors to k representative vectors generated using the k-means clustering algorithm [171]. The algorithm seeks to find k cluster centres that best represent the distribution of the data. $k = 500$ was recommended in [67] and [175] in order to reduce the number of training vectors with minimal loss of information, and is used here.

In order to visualise the resulting 500 vectors in a 2-D latent space, the NeuroScale mapping algorithm [163] was used, which provides a mapping from 4-D data space to 2-D latent space, using a radial basis function (RBF) neural network. Inter-point distances from the 4-D data space are presented in the 2-D latent space, with the aim of preserving

	μ_g	σ_g
HR (bpm)	82.87	16.83
BR (rpm)	18.42	5.06
SpO_2(%)	96.44	2.97
SDA (mmHg)	97.43	14.67

Table 3.6: Mean and standard deviation for each vital sign in dataset $\mathcal{D}_1$.

the inter-point distances as much as possible. The 2-D visualisation of the 500 prototype patterns obtained from k-means clustering is shown in Figure 3.2. From Table 3.5, we know that over 90% of data in dataset $\mathcal{D}_1$belong to the "normal" class C_1. However, from the discussion in section 3.4 it is also known that class C_1 is not exclusively representative of "normal" data. In light of this, as an initial approach for purposes of explaining the concept of model of normality, it is assumed (in this chapter only) that at least 80% of dataset $\mathcal{D}_1$ may be trusted as being "normal" [4]. Further, it is assumed that the most central point on the map (with coordinates [0,0]) represents the most "normal" vital signs, and define the "region of normality" to be the locus of the innermost 80% , i.e. the innermost 400 out of the 500 vectors. Here again it is noted that, strictly speaking, this assumption (of coordinates [0,0] representing the most "normal" vital signs) is valid where the vital-sign distributions contributing to the model of normality take Gaussian form in their individual 1-D space. As was previously mentioned in this section, although the distribution of SpO_2 is not Gaussian, as an initial approach, it is assumed to be so in this work. The remaining data, i.e. those lying outside of the region of normality, are considered to be "outliers" (see Figure 3.2).

Having defined the region of normality, we proceed to the final stage in the construction of a model of normality, where the density distribution of the "normal" data lying within the region of normality is estimated using Parzen windows.

3.5.3 Estimating Pdf using Parzen Windows

The Parzen windows [51] approach estimates the probability density function (pdf) by placing a kernel function on each of the N patterns in the multi-dimensional training data $\mathbf{x}_1, ..., \mathbf{x}_N$. Assuming there are k out of a total of N patterns in a certain window of data

[4]Methods for determining an optimal threshold for the model of normality are discussed in detail in Chapters 4 and 5.

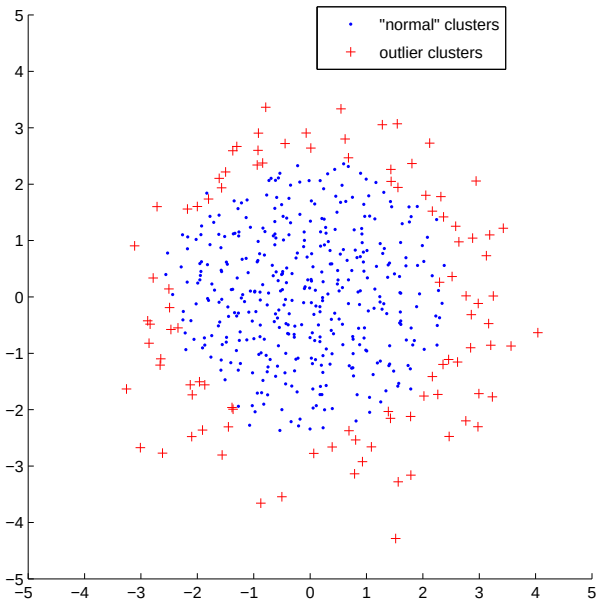


Figure 3.2: 2-D visualisation of the 500 vectors of dataset $\mathcal{D}_1$ obtained after k-means clustering. The 400 "normal" vectors, most close to the centroid, are shown in blue. The remaining 100 "outlier" or "abnormal" vectors are shown in red.

with volume V, the normalised probability density of patterns in volume V is $\frac{k/N}{V}$.

Further, for the sake of explanation, a uniform kernel function (or window function) $\varphi(\boldsymbol{\mu})$ centred at $\mathbf{x}$ may at first be assumed, such that

$$\varphi(\boldsymbol{\mu}) = \begin{cases} 1 & if\ \mathbf{x} - \mathbf{x}_i \leq h \\ 0 & otherwise \end{cases} \tag{3.2}$$

where h is distance from $\mathbf{x}$ to the edge of the window, giving the probability density $p(\mathbf{x})$ as a linear sum of kernels centred at each point $\mathbf{x}_i$ in the dataset,

$$p(\mathbf{x}) = \frac{k}{NV} = \frac{1}{NV} \sum_{i=1}^{N} \varphi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) \tag{3.3}$$

A popular kernel function is the Gaussian kernel function $\varphi(\boldsymbol{\mu})$, where for zero-mean, unit-variance,

$$\varphi(\boldsymbol{\mu}) = \frac{1}{(2\pi)^{d/2}} \exp\left\{-\frac{|\mathbf{x}|^2}{2}\right\} \tag{3.4}$$

Inserting Equation 3.4 in to Equation 3.3, the probability density $p(\mathbf{x})$ for a Gaussian kernel function may be expressed as

$$p(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sigma^d}\frac{1}{(2\pi)^{d/2}}\exp\left\{-\frac{|\mathbf{x}-\mathbf{x}_i|^2}{2\sigma^2}\right\} \tag{3.5}$$

where, for this work, N= 400, and number of data dimensions, d= 4. $p(\mathbf{x})$ is therefore the normalised sum of N Gaussians, each centred on $\mathbf{x}_i$, and each with covariance σ^2. The choice of the *bandwidth parameter* σ determines the smoothness of the resulting pdf estimate $p(\mathbf{x})$. Here, σ is selected using the method of Bishop [14], where for every pattern $\mathbf{x}_i$, a group Q_i, comprising of its m nearest neighbour datapoints is selected. N such groups result, one for each $\mathbf{x}$. The value of $m = 10$, used in earlier works [14, 67], is also adopted here. The bandwidth parameter σ is then calculated from the N intra-cluster variances as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{m}\sum_{j\in Q_i}|\mathbf{x}_i-\mathbf{x}_j|^2\right) \tag{3.6}$$

Inserting Equation 3.6 into Equation 3.5 yields the desired pdf $p(\mathbf{x})$.

3.5.4 Assigning a Novelty Score

The novelty score $z(\mathbf{x})$ for some vector $\underline{\mathbf{x}}$ is defined here as:

$$z(\underline{\mathbf{x}}) = \log_e \frac{1}{p(\underline{\mathbf{x}})} = -\log_e p(\underline{\mathbf{x}}) \tag{3.7}$$

It is expected that patterns lying at the centroid of the model of normality will be most "normal", and will therefore correspond to the peak of the probability density $p(\underline{\mathbf{x}})$, and reciprocally the lowest novelty score. $\log_e \frac{1}{p(\underline{\mathbf{x}})}$ is used since the more "abnormal" data, which one is interested to identify, will take a lower value of $p(\underline{\mathbf{x}})$ and a higher value of the novelty score $z(\underline{\mathbf{x}})$; in other words, $\log_e \frac{1}{p(\underline{\mathbf{x}})}$ emphasizes the more "abnormal" data. The novelty score of the mean of the 400 most "normal" patterns [-0.0380 0.1190 -0.1328 -0.0337] was calculated to be 4.1611. In order for the novelty score of the most "normal" pattern to be approximately 0, an offset of 4.0 was therefore subtracted from the novelty scores in all subsequent calculations.

3.6 Discussion

In this chapter, the construction of a model of normality for performing one-class novelty detection has been presented. The model of normality is based on "normal" data only without any knowledge of the "abnormal" class being incorporated. However, dataset $\mathcal{D}_1$ which was introduced, contains both "normal" and "abnormal" data, as annotated by clinical experts. The next chapter considers how the "abnormal" data may be incorporated into the modelling, despite its overwhelming under-representation in the dataset.

Chapter 4

One-Class Classification

4.1 Introduction

In the previous chapter, a purely one-class approach was described for the construction of a model of normality. To elaborate, no pre-existing knowledge of "class" was incorporated in to the modelling. Instead, as is the nature of one-class modelling, the technique used in the previous chapter relied on the density distribution of the data to inform which data should be considered to belong to some "normal" class. This chapter builds on this work by testing the strength of this one-class approach through incorporation of the knowledge of adverse event labels provided by the clinical experts.[1]

Dataset $\mathcal{D}_1$ was introduced in Section 3.4, which comprises class C_1, which contains vital-sign data labelled by clinical experts as "normal", and class C_2 which contains vital-sign data labelled by clinical experts as "abnormal". As explained in section 3.3, class C_2 is not an exhaustive set of all "abnormal" data. Furthermore, an adverse event need not necessarily manifest itself as an abnormality in all four vital signs, hence there is overlap in the distributions of vital signs between classes C_1 and C_2. Additionally, the various abnormal physiologies in the $\mathcal{D}_1$ patient population may not be equally represented.

Table 4.1 shows that there are far fewer adverse events associated with low HR (or bradycardia) and low BR (or bradypnoea) in $\mathcal{D}_1$ than there are of any other types of abnormality.

The following sections begin with a probabilistic method to discriminate between "normal" and "abnormal" data, based on the model of normality described in Chapter 3.

[1]Therefore, the title of this chapter is intended to be inferred as *the application of* classification methods (which require class labels) to test a model generated using a one-class approach (for which the knowledge of "class" is irrelevant).

Vital Sign	High	Low
HR	15	4
BR	20	6
SpO_2	–	39
SBP/DBP	13	15

Table 4.1: The number of adverse events in dataset $\mathcal{D}_1$ associated with high or low value of the corresponding vital sign, as annotated by the clinical experts.

Methods are presented to determine an optimal separation between the two classes. We also investigate if this approach results in the misclassification of the low HR and low BR data, in particular. For this reason, the HR and BR are used when a 2-D model is constructed in this and subsequent chapters. Lastly, results are presented for the 2-D dataset (including HR and BR) and for the full 4-D dataset (including HR, BR, SpO_2, and SDA).

4.2 Calculating the Probability of Novelty

A probability $\tilde{P}$ is defined, for each vector $\mathbf{x}$ in the data, such that $\tilde{P}$ represents with what probability the data vector $\mathbf{x}$ is likely to be observed with respect to the model of normality (described in Chapter 3), i.e. $\tilde{P}$ is an estimate of how "normal" each vector may be considered to be. Also defined here is a probability P, which is the probability for novelty or "abnormality" of vector $\mathbf{x}$ with respect to the model of normality (or more simply, P= 1 - $\tilde{P}$). From the model of normality, a highly "normal" pattern (i.e., one with a high value of $p(\mathbf{x})$, and correspondingly low novelty score), would be highly probable and thus have a high probability $\tilde{P}$ and low P. Conversely, a highly "abnormal" pattern would take a low value of $\tilde{P}$ and a high value of P. If the probability density $p(\mathbf{x})$ for a certain vector is above a certain threshold, it will be considered a "normal" vector, and "abnormal" otherwise.

The probability $\tilde{P}$ may then be calculated. For uni-dimensional x, the probability that x lies within a certain interval $[a, b]$ is given by :

$$\tilde{P}\left(x \in (a, b)\right) = \int_a^b p(x)\mathrm{d}x \tag{4.1}$$

which is the cumulative distribution (cdf) for pdf $p(x)$. However if the cdf is undefined

for multi-dimensional distributions, such as the 4-D pdf $p(\mathbf{x})$ described in the previous chapter, the probability of a vector lying within the region bounded by some threshold T_{pdf} on the probability density may be used, given by:

$$\tilde{P}\left(\{p(\mathbf{x}) \geq T_{pdf}\}\right) = \int_{\gamma}^{\tau} p(\mathbf{x}) \mathrm{d}\mathbf{x} \tag{4.2}$$

where $\gamma = \underset{x}{\operatorname{argmax}}\, p(\mathbf{x})$, which is the mode of the pdf, and where $\tau = \{\mathbf{x} : p(\mathbf{x}) = T_{pdf}\}$, which is the level set on the pdf at T_{pdf}. Thus $\tilde{P}\left(\{p(\mathbf{x}) \geq T_{pdf}\}\right)$ is the probability mass contained by integrating the pdf from its location of highest density γ to the probability density contour T_{pdf} at τ. The value of $\tilde{P}$ could be calculated by inserting the Parzen windows estimate of the pdf in Equation 3.5 into the above expression; however, it would not be possible to solve the resulting expression analytically. Instead the following numerical approach was used to estimate $\tilde{P}$.

First R random samples $\mathbf{x}_1, ...,\mathbf{x}_R$ were drawn from the pdf estimated using Parzen windows. This was performed by randomly selecting one kernel at a time from the 400 kernels in the Parzen windows model (obtained during construction of the model of normality in section 3.5.3), with each kernel having an equal probability of being selected. A sample was then drawn from the hyper-spherical Gaussian distribution corresponding to the selected kernel. The probability density for each sample was then calculated using the Parzen windows estimate of the pdf. As the R samples were randomly sampled, the resulting estimate of the pdf could be different. Therefore to ensure reproducibility of the results, each time the sampling was performed a sufficiently large number of samples must be drawn. In this work, a value of $R = 10^5$ was selected based on the suitability test recommended in [67] to ensure reproducibility.

Once the probability density $p(\mathbf{x})$ of all samples was known, the probability $\tilde{P}$ of a pattern $\mathbf{x}$ lying within the region bounded by τ could be estimated by the fraction of all samples that have a probability density greater than or equal to the threshold valueT_{pdf}:

$$\tilde{P}\left(\{\mathbf{x} : p(\mathbf{x}) \geq T_{pdf}\}\right) = \frac{S}{R} \tag{4.3}$$

where R is the total number of samples, and S is the number of samples with $p(\mathbf{x}) \geq T_{pdf}$. T_{pdf} is varied from 0.1 to 0.95 with a step size $\triangle T_{pdf} = 0.001$. The probability of

novelty $P(\mathbf{x})$ is then the fraction of samples that have a probability density $p(\mathbf{x}) \leq T_{pdf}$:

$$\tilde{P}\left(\{\mathbf{x} : p(\mathbf{x}) < T_{pdf}\}\right) = 1 - \frac{S}{R} \tag{4.4}$$

The aim of this procedure was to be able to produce a mapping between density values $p(\mathbf{x})$, and corresponding probability of novelty $P(\mathbf{x})$. This transformation could then be used to estimate $P(\mathbf{x})$ for any new vector $\mathbf{x}$, given only its probability density $p(\mathbf{x})$. For simplicity, let the probability of novelty $P\left(\{\mathbf{x} : p(\mathbf{x}) \leq T_{pdf}\}\right)$ now be expressed as $P(\kappa_{\mathbf{x}})$, where $\kappa_{\mathbf{x}} = p(\mathbf{x})$. Figure 4.1 displays the relationship between $P(\kappa_{\mathbf{x}})$ and $\kappa_{\mathbf{x}} = p(\mathbf{x})$ for the 4-D dataset. Data that lie closer to the model of normality, and thus have high probability density $p(\mathbf{x})$, take a correspondingly low probability of novelty $P(\kappa_{\mathbf{x}})$, whereas data that are further from the mode of normality have relatively lower $p(\mathbf{x})$ and correspondingly higher probability of novelty $P(\kappa_{\mathbf{x}})$. The figure also shows the relation between $P(\kappa_{\mathbf{x}})$ and novelty score $z(\mathbf{x})$. Both relationships shown mean that the "abnormality" of a vector increases with increasing values of $P(\kappa_{\mathbf{x}})$. As described in Section 3.5.4, the novelty score of the mean of the 400 most "normal" patterns [-0.0380 0.1190 -0.1328 -0.0337] was calculated to be 4.1611. In order for the novelty score of the most "normal" pattern to be approximately 0, an offset of $z(\mathbf{x}) = 4.0$ was therefore subtracted from the novelty scores in all subsequent calculations. The x-axis on the right-side plot in Figure 4.1 shows $z(\mathbf{x})$ after subtracting 4.0 from all $z(\mathbf{x})$, so that here $z(\mathbf{x}) = 0$ for the most-normal pattern.

Now a threshold T on $P(\kappa_{\mathbf{x}})$ is defined, and data for which data $P(\kappa_{\mathbf{x}}) \geq T$ are considered. Note that T is a threshold on the probability of novelty $P(\kappa_{\mathbf{x}})$, and must not be confused with T_{pdf} which is a threshold on the probability density $\kappa_{\mathbf{x}} = p(\mathbf{x})$. Data that lie furthest away from normality will take the largest value of $P(\kappa_{\mathbf{x}})$, and, as T is increased, the proportion of data with $P(\kappa_{\mathbf{x}}) \geq T$ will decrease. The latter will be the most "abnormal" data.

T can be used to estimate the clinical labels for "abnormal" data i.e. those data belonging to class C_2. Figure 4.2 shows the vital signs for patient 78 from dataset $\mathcal{D}_1$, for a period labelled as being an adverse event (i.e., a period of "abnormal" data) by the clinical experts. The top panel of the figure shows the corresponding values of $P(\kappa_{\mathbf{x}})$. It may be seen that one or more of the four vital signs show deterioration within the interval

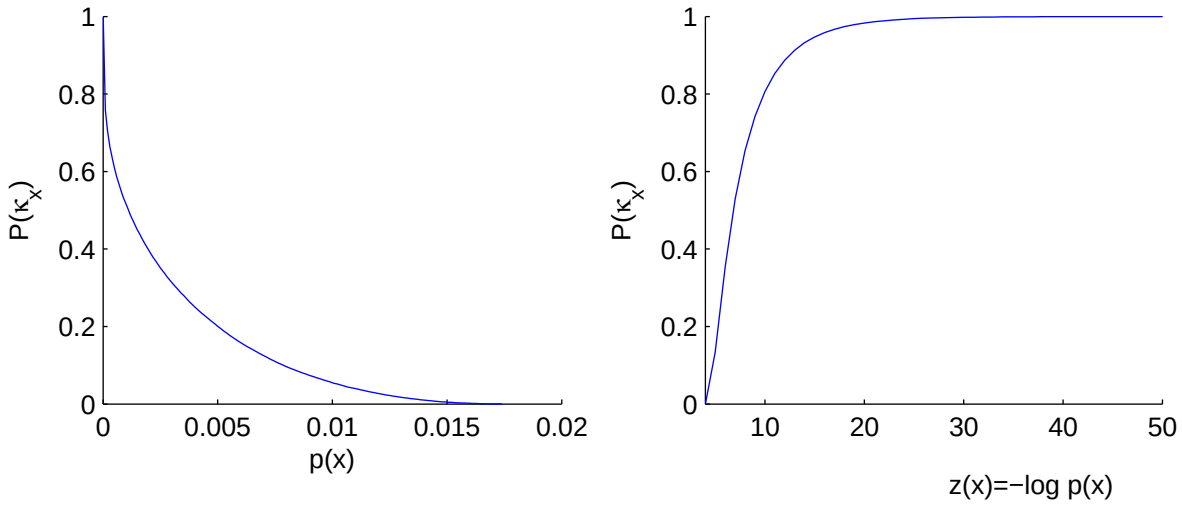


Figure 4.1: Probability $P(\kappa_{\mathbf{x}})$ against probability density, and against novelty score $z(\mathbf{x})$, in the left-hand and right-hand plots, respectively.

bounded by the start-of-event and end-of-event labels. For instance there is a significant decrease in SpO_2 starting at approximately $t = 4$, lasting until approximately $t = 11$ minutes. Similarly, it may be seen that for this same duration, the value of $P(\kappa_{\mathbf{x}})$ is significantly higher than the periods before and after the clinical expert-labelled adverse event. It may also be seen that the clinical experts' labels do not perfectly match the deterioration in vital signs; e.g., the onset of the oxygen de-saturation begins earlier than the start-of-event label, and the restoration to 100% saturation occurs much after the end-of-event label. If an example value of $T = 0.5$ is now selected, and all data lying in the period marked by $P(\kappa_{\mathbf{x}}) \geq T$ are considered as "abnormal", an estimate of the original clinical labels can be achieved, as shown by the red segments on the $P(\kappa_{\mathbf{x}})$ curve. This notion of imposing a threshold on $P(\kappa_{\mathbf{x}})$ forms the basis of the proposed method developed to estimate, and later refine the clinical labels (described in Chapter 5).

In the following sections, methods are discussed which are used to optimise the value of T so as to provide an estimation of the original clinical labels.

4.3 Optimisation of T Based on Data Distribution

A number of different approaches were investigated to yield a value of T that would best estimate the original clinical labels. Sections 4.3.1 and 4.3.2 below focus on estimating T

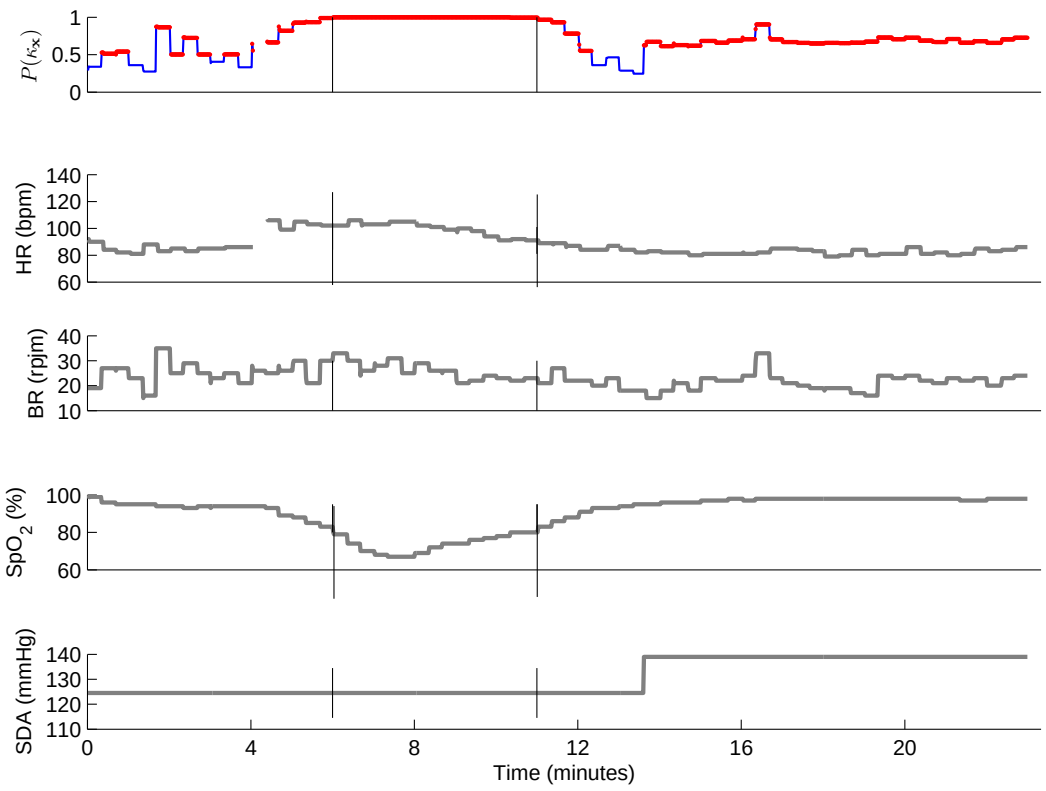


Figure 4.2: Estimating clinical labels. According to the original clinical labels, data from $t = 6$ to $t = 11$ minutes were labelled as corresponding to an adverse event, as marked by the vertical black lines. Periods of data estimated as having "abnormal" physiology according to $T = 0.5$ are shown in bold red on the uppermost plot, where T is a threshold on $P(\kappa_{\mathbf{x}})$.

Class	Clinical label	Description
C_2	"abnormal"	All data from dataset $\mathcal{D}_1$, belonging to periods considered by clinical experts to have "abnormal" patient condition.
C_1	"normal"	All data from dataset $\mathcal{D}_1$, except the data which belong to C_2.

Class	Estimated label	Description
$C_{2_{est}}$	"abnormal"	All data from dataset $\mathcal{D}_1$, for whom $P(\kappa_{\mathbf{x}}) > T$
$C_{1_{est}}$	"normal"	All data from dataset $\mathcal{D}_1$, for whom $P(\kappa_{\mathbf{x}}) \leq T$

Table 4.2: Description of classes C_1 and C_2 belonging to $\mathcal{D}_1$ dataset, and the estimated classes after introducing T.

by using statistical metrics for comparison of the uni-dimensional and multi-dimensional distributions of data from class C_2.

In section 4.2, it was suggested that if the threshold T is raised, data from $\mathcal{D}_1$ with $P(\kappa_{\mathbf{x}}) > T$ would increasingly resemble those data in class C_2. This is because in raising the value of the threshold, only increasingly "abnormal" data will remain above the threshold. In the following analysis, this resulting group of data with $P(\kappa_{\mathbf{x}}) > T$ will be referred to as class $C_{2_{est}}$. The remaining group of data from dataset $\mathcal{D}_1$ with $P(\kappa_{\mathbf{x}}) \leq T$ will be referred to as class $C_{1_{est}}$ (see Table 4.2).

4.3.1 Uni-dimensional Comparison

Considering one vital sign at a time, classes $C_{2_{est}}$ and C_2 were compared. The underlying premise is that for the optimal value of T, the distribution of $C_{2_{est}}$ would most closely resemble that of C_2. Such a value would be referred to as T_{opt}. Figures 4.3 and 4.4 show the uni-dimensional pdf and cdf distributions of two vital signs, HR and BR, for varying values of T. The pdf has been normalised such that the area under the curve is equal to one. It may be seen that as the value of T is increased from 0.1 to 0.9, both pdf and cdf distributions of $C_{2_{est}}$ increasingly resemble the corresponding pdf and cdf distributions of C_2.

The distance between the pdf distributions for C_2 and $C_{2_{est}}$ was calculated for a range of values of T. The expectation was that for a candidate value of T_{opt}, the distance between the two distributions would be at a minimum when T_{opt}. The distance metric used to compare the two distributions was the Kullback-Leibler divergence (KL), which

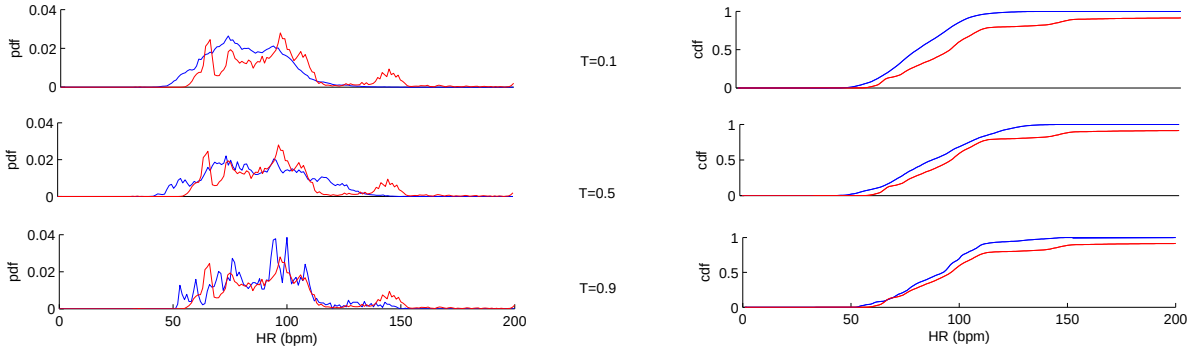


Figure 4.3: The pdf (left column) and cdf (right column) for HR for $T = 0.1$, 0.5, and 0.9, from top to bottom, respectively. The blue curve represents the distribution for class $C_{2_{est}}$ and the red curve represents the distribution for class C_2.

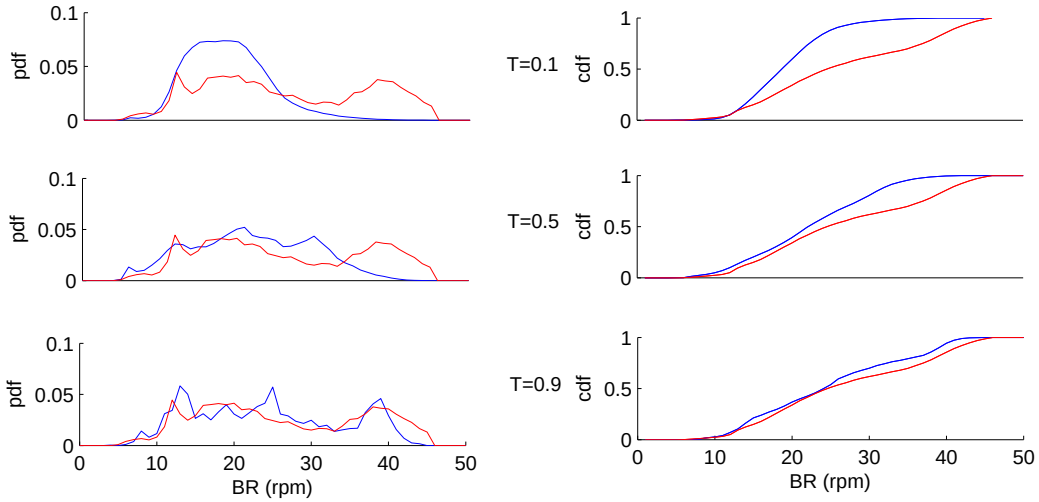


Figure 4.4: The pdf (left column) and cdf (right column) for BR at $T = 0.1$, 0.5, and 0.9, from top to bottom, respectively. The blue curves represent the distribution for class $C_{2_{est}}$ and the red curve represent the distribution for class C_2.

is a well-known metric for statistical comparison of distributions [14].

Kullback-Leibler Divergence

The Kullback-Leibler divergence $KL(r||q)$ between any two distributions, $r(x)$ and $q(x)$, computes their joint or relative entropy; i.e., how much information they contain in common [90], where $r(x)$ is the "true" distribution and $q(x)$ is some approximation of $r(x)$. The concept of entropy has its roots in information theory where, for a distribution of the variable x, the entropy $H(x) = \sum_x r(x) \log_2 r(x)$ describes the total number of extra bits required to represent each state of a variable x, where $r(x)$ is the probability of occurrence

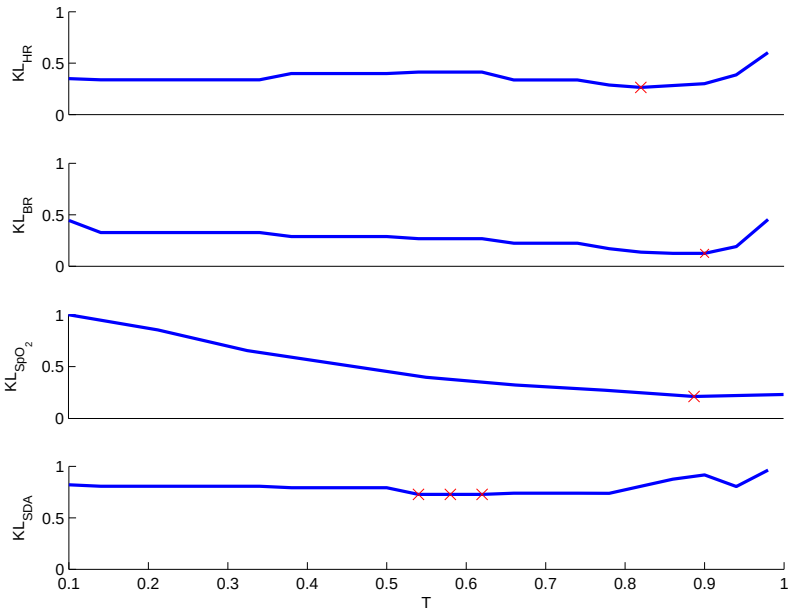


Figure 4.5: Kullback-Leibler divergence (y-axes) calculated for the pdf distributions of the four vital signs HR, BR, SpO_2, and SDA. The value of T at which each divergence is calculated is shown on the x-axes. The red crosses mark the minimum value of KL.

of each state of x. The KL metric must be a non-negative value, and is given by

$$KL(r||q) = \int r(x) \ln \left\{ \frac{q(x)}{r(x)} \right\} dx \tag{4.5}$$

where $KL(r||q) = 0$, if $r = q$. A low value of KL may therefore be interpreted as indicative of a closer match between $r(x)$ and $q(x)$ than a relatively higher value of KL. For the purpose of the uni-dimensional investigation, $r(x)$ is the uni-dimensional pdf distribution of class C_2, and $q(x)$ is the uni-dimensional pdf distribution of the same vital sign from class $C_{2_{est}}$. Figure 4.5 shows the KL divergence calculated for each of the four vital signs considered individually. From these results, it may be seen that while the values of T for which KL_{HR}, KL_{BR} and KL_{SDA} and KL_{SpO_2} are at a minimum is in the range $T = 0.5$ to $T = 0.9$, there is no single common minimum on T_{KL} for the four vital signs. In order to find a single value for T'_{opt}, a comparison of multi-dimensional distributions was considered instead.

4.3.2 Multi-dimensional Comparison

In order to compare distributions that were representative of the combined behaviour of all four vital signs, the distributions of the $P(\kappa_{\mathbf{x}})$ values for data from classes C_2

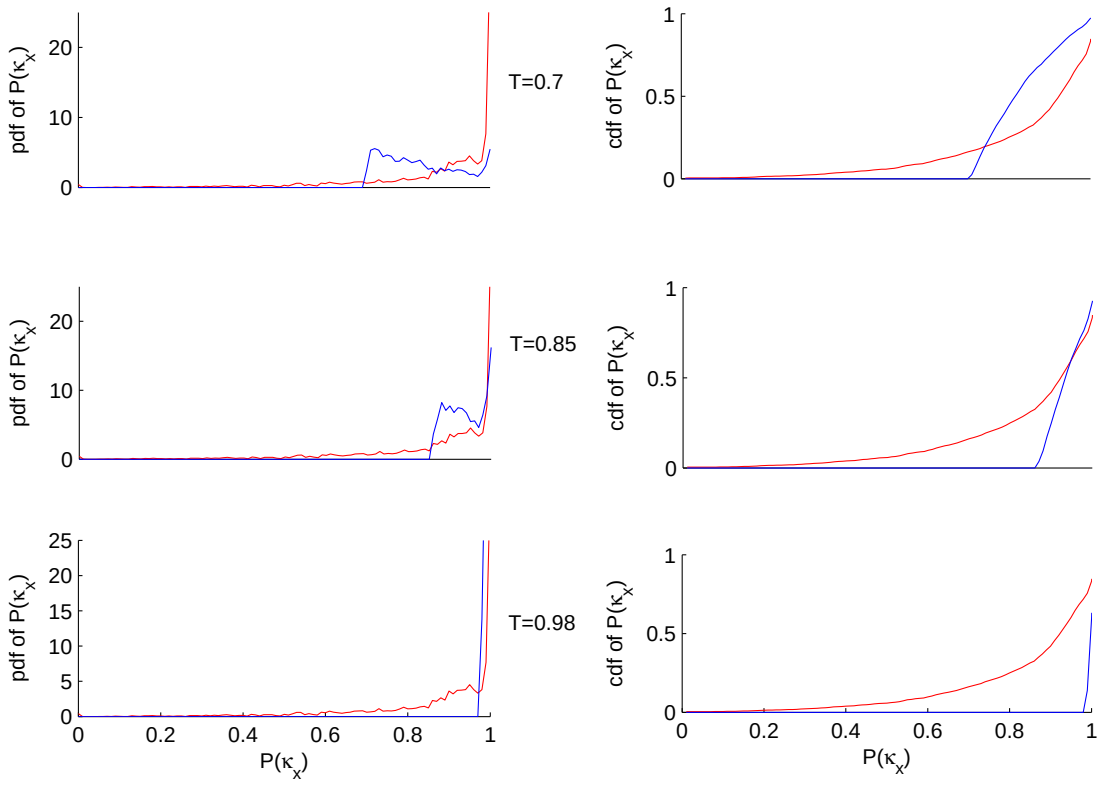


Figure 4.6: The pdf (left column) and cdf (right column) distributions for $P(\kappa_{\mathbf{x}})$ at $T = 0.7$, 0.85, and 0.98, from top to bottom, respectively. Blue curves represent the distributions for class $C_{2_{est}}$ and red curves represent the distributions for class C_2.

and $C_{2_{est}}$ were compared. Figure 4.6 shows the pdf and cdf distributions of $P(\kappa_{\mathbf{x}})$ for different values of threshold T. The distribution for $C_{2_{est}}$ changes shape according to T. From visual inspection of these results it is not obvious which value of T provides the largest overlap in the C_2 (red) and $C_{2_{est}}$ (blue) distributions. The KL distance was calculated for these distributions over the range $0.1 \leq T \leq 0.99$. Figure 4.7 shows the resulting KL distance over a smaller range of values of T (i.e., $0.1 \leq T \leq 0.99$), in order to emphasize the region containing the minimum KL distance. The KL divergence decreases as T increases, indicating that the overlap between C_2 and $C_{2_{est}}$ increases with T. The distributions converge to a minimum at $T_{KL_{min}} = 0.86$ (with the corresponding KL distance being 46.8 units), which is therefore considered here as a candidate optimal threshold, T'_{opt}.

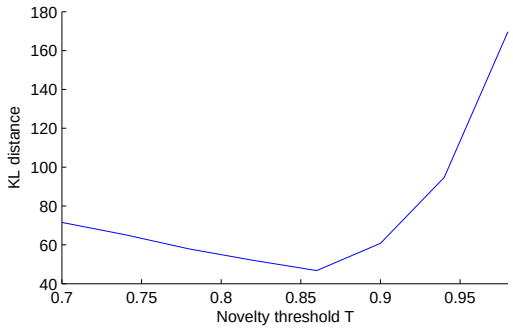


Figure 4.7: Kullback-Leibler divergence (y-axis) calculated for distributions of pdf of $P(\kappa_{\mathbf{x}})$. The value of T at which the divergence is calculated is shown on the x-axis. KL is at a minimum at $T = 0.86$.

4.4 Optimal Value of T for Identifying Abnormal Data

It is now investigated if $T = 0.86$ is an appropriate choice for the optimum value of threshold T to estimate the clinical labels. The thresholding technique discussed in section 4.3 is first investigated using a 2-D subset for purposes of convenient visualisation, and later applied to the full 4-D dataset.

To apply the proposed thresholding concept to 2-D data, it was necessary to calculate $P(\kappa_{\mathbf{x}})$ in two dimensions, rather than the full four dimensions considered previously. This could be achieved by either re-training the model of normality in 2-D, or, more simply, by reducing the number of dimensions in the data, by marginalising out the unused dimensions from the model, as done in [67]. Let the 2-D dataset comprising HR and BR be considered first, where SpO_2 and SDA are removed using the "marginalisation" technique found in [67].

Figure 4.8 illustrates the effect of applying the proposed method to data from dataset $\mathcal{D}_1$, which is here refered to as the threshold-sweeping approach. The grey cluster shows the resulting set of data $C_{2_{est}}$ for the corresponding value of threshold T. The superimposed black clusters depict data from class C_2. Note that here, for purposes of presenting a purely 2D approach, only data in C_2 associated with only HR or BR or both is included. This was possible to do since the labels in dataset $\mathcal{D}_1$ also identified which vital sign(s) the clinical experts considered to have been associated with the adverse event. As can be seen from the plots from left to right, as the value of T is increased, only those data

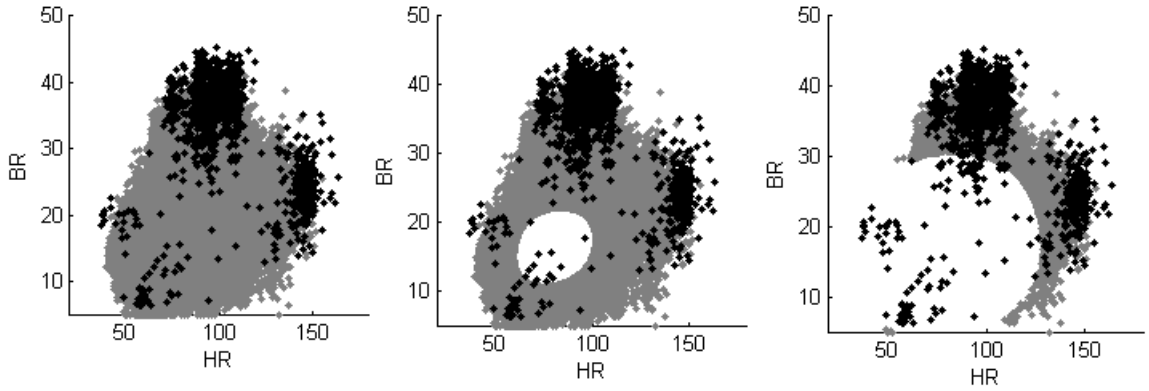


Figure 4.8: The effect of T. Data from class $C_{2_{est}}$ are shown in grey. From left to right, the plots show $C_{2_{est}}$ for $T = 0$, $T = 0.5$, and $T = 0.86$, respectively. Data from class C_2 are superimposed on each plot, shown in black. The central white region in each plot represents $C_{1_{est}}$ for the corresponding value of T.

with $P(\kappa_{\mathbf{x}}) > T$ are retained in $C_{2_{est}}$, and the centre-most (i.e. the most "normal") data are removed from class $C_{2_{est}}$, forming $C_{1_{est}}$. One might think that at T'_{opt}, grey clusters would coincide with black clusters, as then they would both be representing the same "abnormal" data. However, recalling the discussion in section 4.1, data in class C_2 are not an exhaustive set of all "abnormal" data, and so an exact match of the grey and black clusters is not to be expected.

Consider now the candidate value of $T'_{opt} = 0.86$. It may be seen in the right-most plot of Figure 4.8 that at this value of T, the grey data are quite similar in shape and position to the two black clusters representing "abnormal" data resulting from high-HR and high-BR events (tachycardia and tachypnoea, respectively). However, there is no grey cluster approximating the low-HR and low-BR events (bradycardia and bradypnoea, respectively). Thus at $T = 0.86$, two significant types of "abnormal" clusters have been erroneously estimated as being "normal". This is because the approach to calculate T'_{opt} has so far been entirely based on a density-based model of normality, which does not take into account the prior knowledge that data corresponding to bradycardia and bradypnoea events are abnormal, but of much lower density that the tachycardia and tachypnoea data.

In Figure 4.8, for a purely 2-D presentation of the approach, C_2 data which were labelled by clinicians as being associated with only HR or BR are shown. In Figure 4.9, the effect of this "threshold-sweeping" approach is shown when applied to the complete set of data in C_2, i.e. to data that are labelled by clinicians as being associated with

an abnormality in either the HR, BR, SpO_2, and SDA, not only HR or BR. Figure 4.10 shows the resulting $C_{1_{est}}$ (light blue) and $C_{2_{est}}$ (pink) obtained using this method. Note that in addition to the effects discussed for Figure 4.8, there is a much larger overlap between classes C_1 and C_2. This larger overlap is as expected - it is caused by data from C_2 where HR or BR or both are "abnormal" but the other vital signs are not, as would be in the case of an adverse event where e.g. the HR and BR deteriorated but the SDA and the SpO_2 remained within the normal range, as is illustrated in Figure 4.11, which shows the effect of threshold sweeping on other 2-D vital-sign pairs, and in Figure 4.12, which shows a 3-D view of the dataset.

Next, the effect of threshold-sweeping when applied to the 3-D and 4-D datasets is considered. Here of course $P(\kappa_{\mathbf{x}})$ has been calculated using three and four vital-sign dimensions, respectively. For the 3-D dataset, from Figure 4.12(a) it can be seen that with the inclusion of the SpO_2 dimension, the overlap between the classes C_1 and C_2 is already reduced, even at $T = 0.1$.

The overlap is further reduced in 4-D. From Figure 4.13, which shows different 3-D views of the 4-D dataset, it can be seen that while there is a large overlap in the classes contributed by the HR, BR and SDA, the overlap is reduced when the SpO_2 dimension is added. This is because the low SpO_2 cluster in class C_2 is distinctly separated from the region covered by class C_1 and thus it is expected that any classifier will result in fewer misclassification cases for the 4-D dataset than for the 2-D or 3-D datasets.

4.5 Optimal Value of T for Minimising Classification Error

In the previous section, the use of an optimal T_{opt} for estimating class C_2 was demonstrated. The optimal threshold was selected by minimising the KL divergence between

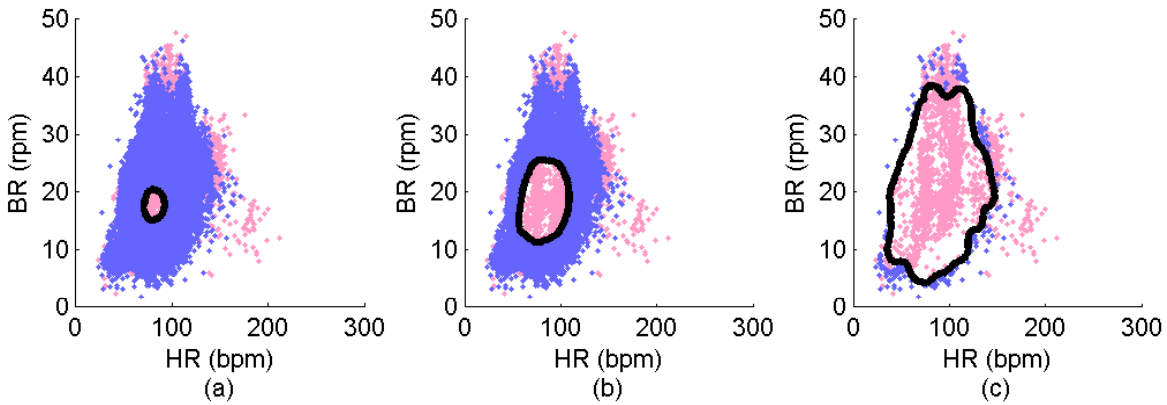


Figure 4.9: The effect of T on the 2-D dataset. Data from class $C_{2_{est}}$ are shown in light blue. From left to right, the plots show $C_{2_{est}}$ for (a) $T = 0.1$, (b) $T = 0.5$, and (c) $T = 0.86$, respectively. Data from class C_2 is superimposed on each plot, shown in pink. The area bounded by the black line in each plot represents $C_{1_{est}}$ for the corresponding value of T.

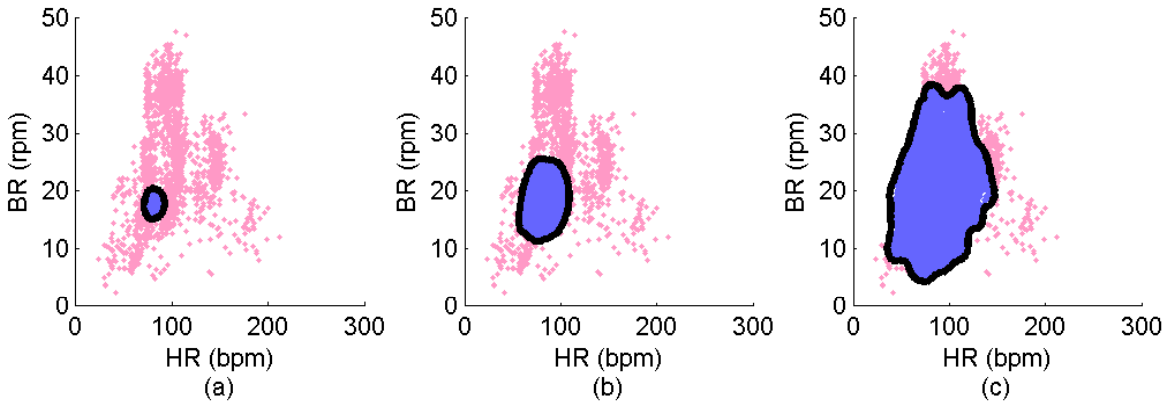


Figure 4.10: The effect of T on the 2-D dataset. Data from class $C_{1_{est}}$ are shown in light blue. From left to right, the plots show $C_{1_{est}}$ for (a) $T = 0.1$, (b) $T = 0.5$, and (c) $T = 0.86$, respectively. Data from class C_2 is superimposed on each plot, shown in pink. The area bounded by the black line in each plot represents $C_{1_{est}}$ for the corresponding value of T.

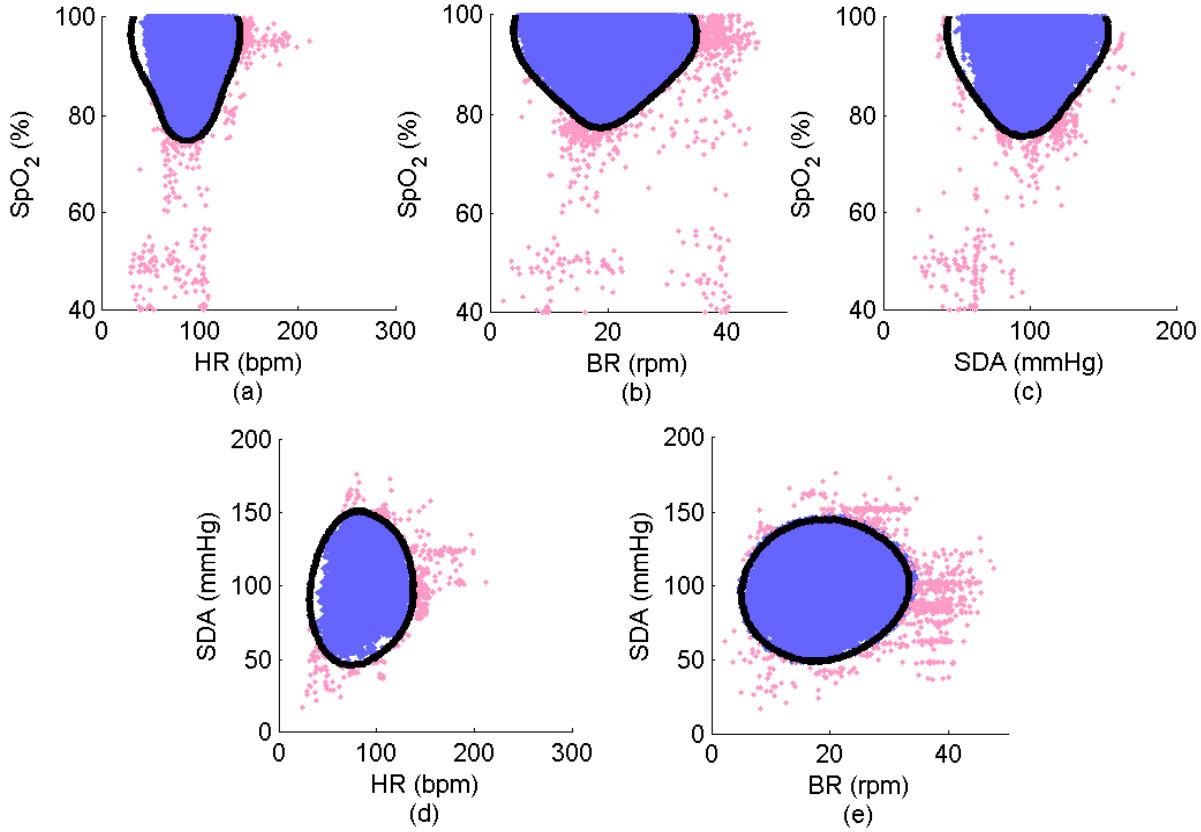


Figure 4.11: The effect of the candidate threshold value of $T'_{opt} = 0.86$ on the 2-D dataset comprising the following combination of vital signs: (a) HR and SpO_2, (b) BR and SpO_2, (c) SDA and SpO_2, (d) HR and SDA, and (e) BR and SDA. In each plot, data from class C_2 are super-imposed on each plot (in the background), shown in pink. Data from class $C_{1_{est}}$, shown in light blue, lie inside the area bounded by the black line.

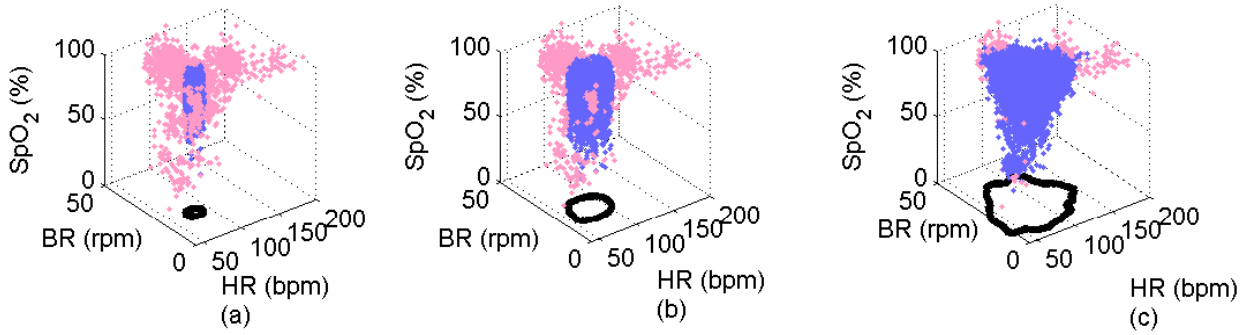


Figure 4.12: The effect of T on the 3-D dataset. Data from class $C_{1_{est}}$ are shown in light blue. From left to right, the plots show $C_{1_{est}}$ for (a) $T = 0.1$, (b) $T = 0.5$, and (c) $T = 0.86$, respectively. Data from class C_2 is superimposed on each plot, shown in pink. The area bounded by the black line in each plot represents $C_{1_{est}}$ for the corresponding value of T, by projecting on to the HR-BR plane.

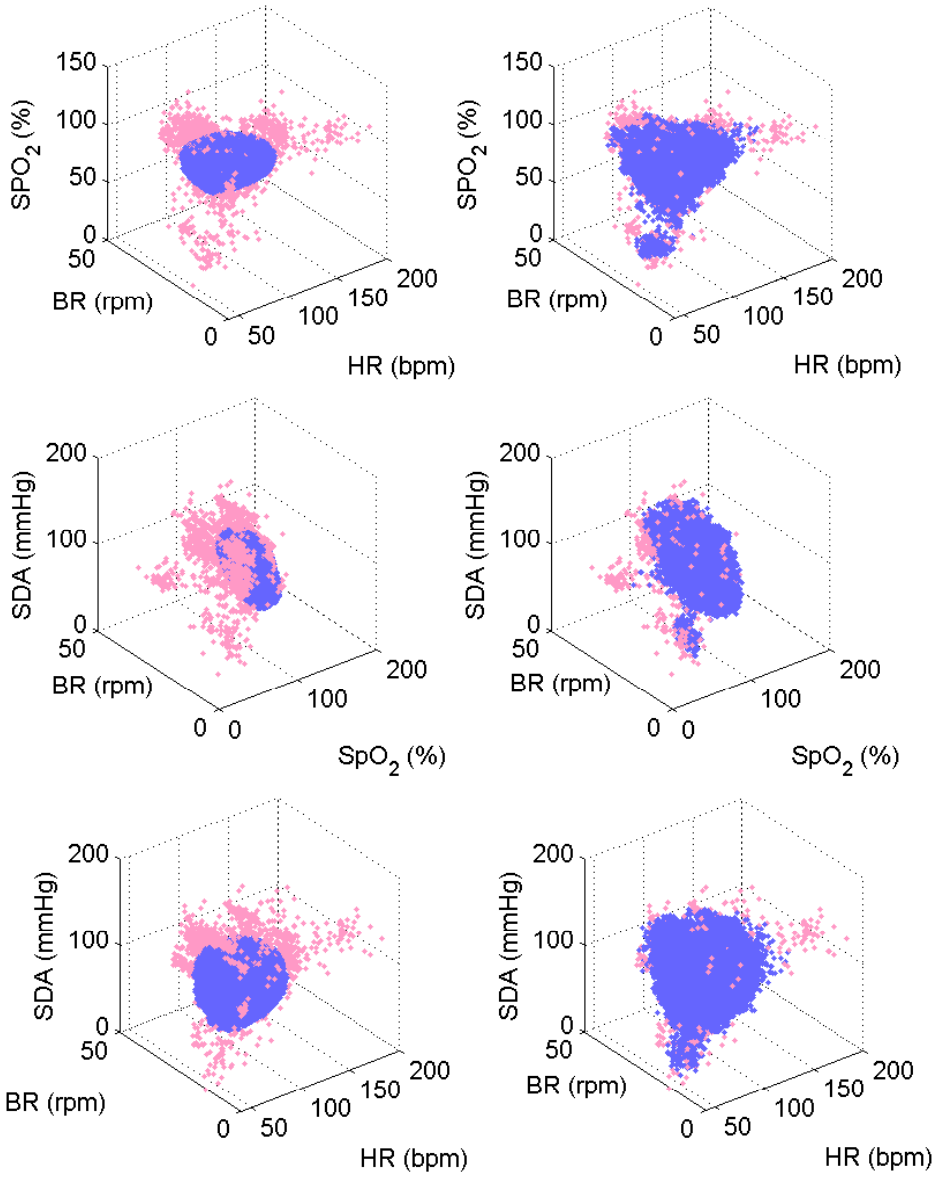


Figure 4.13: The effect of T on the 4-D dataset. Data from class $C_{1_{est}}$ are shown in light blue. Data from class $C_{2_{est}}$ is not shown for ease of visualisation. The three rows show 3-D plots of different combinations of vital signs are shown. In each row, from left to right, the plots show $C_{1_{est}}$ for $T = 0.1$, $T = 0.5$, and $T = 0.86$, respectively. Data from class C_2 is superimposed on each plot, shown in pink.

the distribution of class C_2 and the estimated class $C_{2_{est}}$, with the estimation technique being based on novelty detection (or one-class classification). In the following section, the same paradigm (i.e. one-class classification) is continued with, but the aim is to optimise T so as to minimise the error in classification of "normal" and "abnormal" data (instead of the KL divergence). In order to do so, a training set is selected from dataset $\mathcal{D}_1$ to optimise T, and the performance of the candidate T, T_{opt}, is assessed on an independent held-out test set, also selected from dataset $\mathcal{D}_1$. Note that the training set and test set selected as described in the following section will continue to be used for training and testing other classification methods in the next chapter.

4.5.1 Training and Test Sets for Classification

Recalling from Table 3.5 that the total number of data in C_2 are 2046, which is far fewer than the number of data in C_1 (265,581), the following approach was taken to define training and test sets containing equal number of data from C_1 and C_2.

50% of randomly selected data from C_2 were selected to form C_2^{test} (where 50% of 2,046=1,073), and were placed in the test set, henceforth called $\mathcal{D}_1^{test}$. The remaining 50% of data from C_2 (i.e. $2,046 - 1,073 = 1,073$) were selected to form C_2^{train} and placed in a training set, henceforth called dataset $\mathcal{D}_1^{train}$. Randomly selected data from C_1 , equal in number to data in C_2^{test} , were selected to form C_1^{test} and placed in $\mathcal{D}_1^{test}$. Similarly, randomly selected data from C_1, equal in number to data in C_2^{train}, were selected to form C_1^{train} and placed in $\mathcal{D}_1^{train}$. Thus $\mathcal{D}_1^{test}$ and $\mathcal{D}_1^{train}$are sub-sets of dataset $\mathcal{D}_1$, containing balanced number of labelled data from classes C_1 and C_2. Table 4.3 summarizes the contents of datasets $\mathcal{D}_1^{test}$ and $\mathcal{D}_1^{train}$ and the description of their corresponding class labels.

Also shown in the table is the description of class labels as estimated according to the one-class methods considered in this chapter. Data from $\mathcal{D}_1^{test}$ for whom $P(\kappa_{\mathbf{x}}) > T$ are estimated as belonging to class $C_{2_{est}}^{test}$. Data from $\mathcal{D}_1^{test}$ for whom $P(\kappa_{\mathbf{x}}) \leq T$ are estimated as belonging to class $C_{1_{est}}^{test}$. A data-point from C_2^{test} placed into $C_{2_{est}}^{test}$ is considered to be a True Positive (TP), while a datapoint from C_1^{test} placed into $C_{1_{est}}^{test}$ is considered to be a True Negative (TN). A data-point belonging to C_2^{test}, but estimated as $C_{1_{est}}^{test}$ is considered

Class	Clinical label	Description	Number of data
C_2^{train}	"abnormal"	Contains data from dataset $\mathcal{D}_1^{train}$, belonging to class C_2.	1,073
C_1^{train}	"normal"	Contains data from dataset $\mathcal{D}_1^{train}$, belonging to class C_1, and equal in number to ${C_2}^{train}$.	1,073
		Total number of data in dataset $\mathcal{D}_1^{train}$	2,046

Class	Clinical label	Description	Number of data
C_2^{test}	"abnormal"	Contains data from dataset $\mathcal{D}_1^{test}$, belonging to class C_2.	1,073
C_1^{test}	"normal"	Contains data from dataset $\mathcal{D}_1^{test}$, belonging to class C_1, and equal in number to ${C_2}^{test}$.	1,073
		Total number of data in dataset $\mathcal{D}_1^{test}$	2,046

Class	Estimated label	Description
$C_{2_{est}}^{test}$	"abnormal"	Contains data from dataset $\mathcal{D}_1^{test}$, for which $P(\kappa_{\mathbf{x}}) > T$.

$C_{1_{est}}^{test}$	"normal"	Contains data from dataset $\mathcal{D}_1^{test}$, for which $P(\kappa_{\mathbf{x}}) \leq T$.

Table 4.3: Description of classes within datasets $\mathcal{D}_1^{train}$ (top) and $\mathcal{D}_1^{test}$ (middle), and estimated class labels (bottom) according to the one-class label estimation method.

to be a False Negative (FN), and a data-point belonging to C_1^{test} placed into $C_{2_{est}}^{test}$ is considered to be a false Positive (FP). Classification performance is assessed using the criteria that the best classification (or the lowest classification error) will be achieved when the lowest values of both FP and FN are obtained. In other words, lowest classification error is achieved when the total number of FP and FN is its lowest, and when FP and FN are individually at their lowest. The value of T resulting in this best classification will be considered to be T_{opt}. To summarize:

- True Positive (TP): a correctly classified C_2^{test} point;
- True Negative (TN): a correctly classified C_1^{test} point;
- False Positive (FP): a C_1^{test} point misclassified as belonging to class C_2^{test} ;
- False Negative (FN): a C_2^{test} point misclassified as belonging to class C_1^{test}.

Further, FN_{brady} is defined as the percentage of bradycardia- or bradypnoea-related FN, and FN_{tachy} as the percentage of tachycardia- or tachypnoea-related FN.

4.5.2 Results

In this section, first the results for applying the classification to the 2-D dataset are discussed. Later the same is presented for the 3-D and4-D datasets.

4.5.2.1 2-D Dataset

Figure 4.14 shows the misclassified data for the 2-D dataset $\mathcal{D}_1^{test}$, for a range of values of T. From the figure it may be seen that for a low value of T, such as $T = 0.1$, the class $C_{1_{est}}^{test}$ includes a very small number of data from C_1^{test}, and so the number of FP is correspondingly high (78% C_1^{test} are FPs), as shown in Figure 4.15(a). As T is increased, the region representing $C_{1_{est}}^{test}$ (bounded by the black-coloured boundary) widens and the number of FPs correspondingly decrease. At the same time, the increasing T causes more data from C_2^{test} to be included in $C_{1_{est}}^{test}$, causing the number of FNs to rise. For instance, the percentage of FN cases increases from 8.4% at $T = 0.1$ to 66.5% at $T = 0.9$.

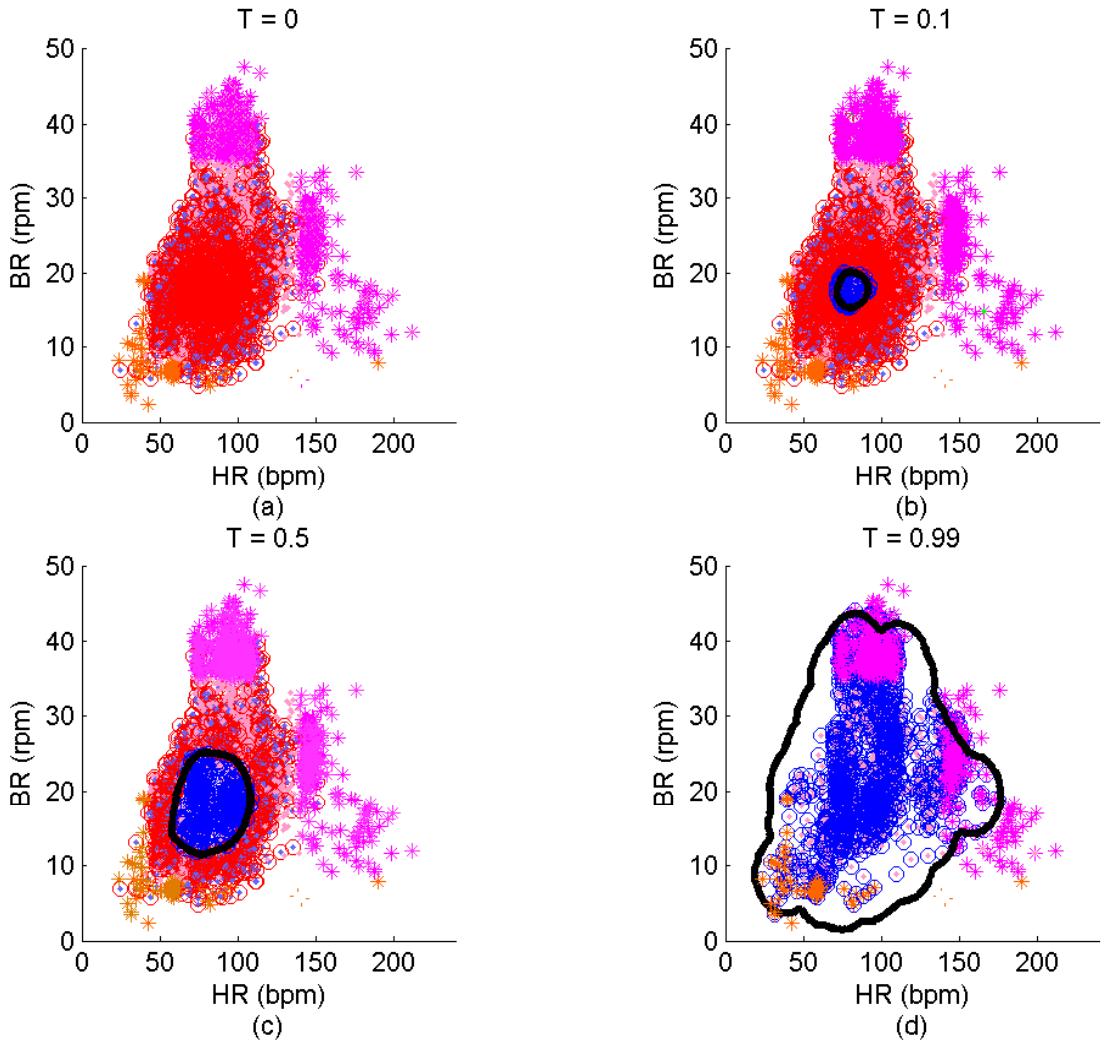


Figure 4.14: One-class classification. Results for (a) $T = 0$, (b) $T = 0.1$, (c) $T = 0.5$, and (d)$T = 0.99$ from top left to bottom right, respectively. Light Pink and light blue dots represent classes C_2^{test} and C_1^{test} of the test set $\mathcal{D}_1^{test}$ (as in previous figures). Tachycardia- and tachypnoea-related C_1^{test} data are shown by bright pink stars. Bradycardia- and bradypnoea-related C_1^{test} data are shown by orange stars. The black line bounds the region considered by the one-class method to belong to class $C_{1_{est}}^{test}$. FP data are encircled in red, FN data are encircled in dark blue.

According to the definition of the classification error described earlier, in Figure 4.15(a) , the classification error will be the lowest at the intersection of the FP and FN curves. Accordingly, it may be seen that the smallest classification error is obtained at $T = 0.4$, where the sum of FN and FP is the smallest. Thus according to this method, the optimal value of T, at which the clinical labels are estimated with the smallest classification error, is $T_{opt} = 0.4$.

4.5.2.2 2-D False Negatives

In section 4.4 it was seen that a limitation of the one-class method is that it can lead to misclassification of the two physiologically important clusters of "abnormal" data, namely the clusters representing bradycardia and bradypnoea. To test this, the bradycardia and bradypnoea as defined according to the upper and lower limits used by the MET calling criteria (from Table 3.2) were considered.

Figure 4.15 (b) shows the effect of raising T on the number of FNs. It may be seen that as T is raised, the contribution of the bradycardia and bradypnoea data to the number of FNs increases. At $T = 0.99$, 90% of the bradycardia and 60% of the bradypnoea data are misclassified. Table 4.4 summarizes the classification results for the candidate optimal threshold, $T_{opt} = 0.4$.

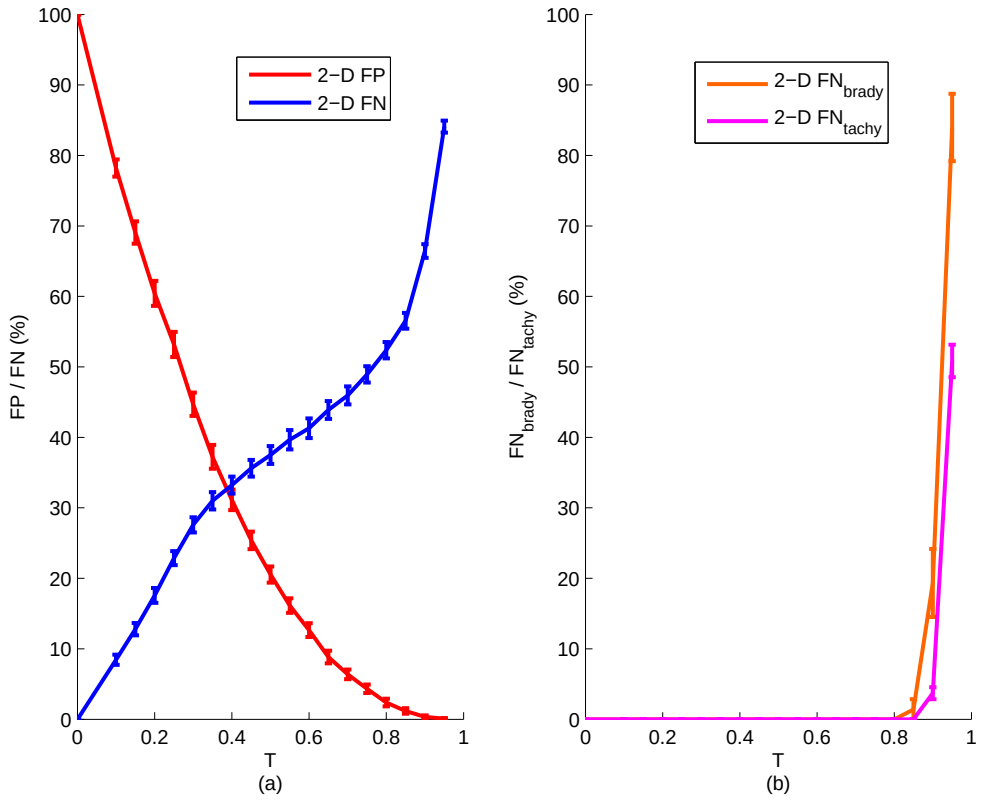


Figure 4.15: Misclassifications over range of T, for the 2-D dataset $\mathcal{D}_1^{test}$. (a) The percentage of false positive outcomes is shown in red, and the percentage of false negatives is shown in blue. (b) The percentages of misclassified tachycardia or tachypnoea data ($\mathrm{F}N_{tachy}$) are shown in bright pink, and the percentages of misclassified bradycardia or bradypnoea data ($\mathrm{F}N_{brady}$) are shown in orange at $T_{opt} = 0.4$.

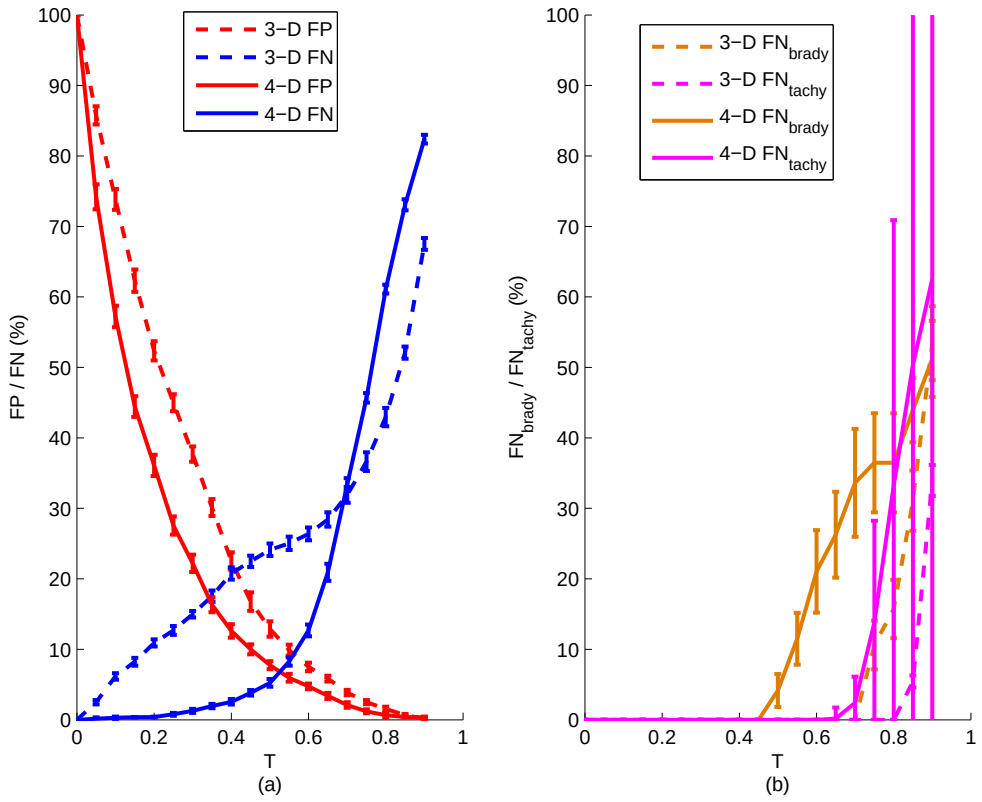


Figure 4.16: Misclassifications over range of T, for the 3-D and 4-D datasets. (a) The percentage of false positives is shown in red, and the percentage of false negatives is shown in blue. 3-D and 4-D datasets are represented by broken and solid line, respectively. (b) The percentages of misclassified tachycardia or tachypnoea data (FN_{tachy}) are shown in bright pink, and the percentages of misclassified bradycardia or bradypnoea data (FN_{brady}) are shown in orange at $T_{opt} = 0.4$. 3-D and 4-D datasets are represented by broken and solid line, respectively.

4.5.2.3 3-D and 4-D Datasets

The method was repeated for the 3-D dataset $\mathcal{D}_1^{test}$, including SDA as the third dimension, and then for the 4-D dataset $\mathcal{D}_1^{test}$, including SpO_2 as the fourth dimension. The 3-D and 4-D results are shown in Figure 4.16. From Figure 4.16(a) it may be noted that the FP and FN percentages at T_{opt} are lower for the 3-D dataset (20.8%), and even lower for the 4-D dataset (8.3%), than they are for the 2-D dataset (33.2%). This is because the full 4-D dataset has much less overlap than the 2-D dataset, and as discussed earlier in section 4.4, this result is to be expected due to the fact that inclusion of the SpO_2 dimension drastically reduces the overlap between the two classes; consequently, the classifier misclassifies fewer data.

One-class Classification	FP (%)	FN (%)	FP_{tachy}(%)	FP_{brady}(%)
2-D dataset, at T_{opt}= 0.4	33.2±1.2	33.2±1.5	0	0
3-D dataset, at T_{opt}= 0.4	20.8±1.2	20.8±0.9	0	0
4-D dataset, at T_{opt}= 0.5	8.3±0.5	8.3±0.6	0	11.5±3.7

Table 4.4: The number of correctly classified and misclassified data from dataset $\mathcal{D}_1^{test}$ for $T_{opt} = 0.4$ for the 2-D and 3-D datasets, and at $T_{opt} = 0.5$ for the 4-D dataset. False negatives corresponding to the tachycardia and tachypnoea region are called FP_{tachy}, and those corresponding to the bradycardia and bradypnoea region are called FP_{brady}. As data in $\mathcal{D}_1^{test}$ are selected randomly (see section 4.5.1), each experiment in which $\mathcal{D}_1^{test}$ is used is repeated 50 times, and the mean ±1 standard deviation error of the results obtained over the 50 experiments is shown in each column of this table.

4.5.2.4 3-D and 4-D False Negatives

Figure 4.16(b) shows the break-up of the FNs for the 3-D and 4-D datasets. It may be seen that that for the 4-D dataset, there is generally a larger percentage of misclassified tachycardia and tachypnoea data, over the range of values of T, than there is in the 3-D dataset. For the 3-D dataset, as with the 2-D dataset, the percentage of false negatives rises sharply as the threshold is raised above $T_{opt} = 0.4$.

The results of the 2D, 3-D and 4D datasets, at their respective T_{opt}, are summarised in Table 4.4. Even with the reduced overlap between classes in 4-D, at $T_{opt} = 0.5$, 11.5% bradycardia and bradypnoea data are misclassified (on average). These results suggest that the one-class method adopted in this chapter manages to classify nearly 91.7% C_1 and C_2 data correctly due to the reduced overlap between classes, but that it does so at the expense of misclassifying the low-density (bradycardia and bradypnoea) data from class C_2.

4.6 Discussion and Conclusions

In this chapter, a thresholding technique was introduced to allow an estimation of the clinical expert-labelled classes C_1 and C_2. First, a probability of novelty was defined based on the model of normality developed in Chapter 3. Uni-dimensional and multi-dimensional analyses were presented to optimise the value of a threshold based on the probability of novelty, such that at the optimal value of this threshold, the estimated class $C_{2_{est}}$ would most closely resemble the original class C_2. It was seen that this method

suggested an optimal value of T, which resulted in significant overlap between $C_{2_{est}}$ and C_2, particularly in the highly dense regions corresponding to tachycardia and tachyapnoea events. The candidate optimal value of T ($T'_{opt} = 0.86$) was however unable to identify bradycardia or bradypnoea correctly, as the data density of the latter is much lower than that for tachycardia and tachypnoea.

The density-based model of normality was then optimised using a threshold T based on knowledge of the class labels ("normal"/ "abnormal"). The optimal threshold of $T_{opt} = 0.4$ was obtained for the 2-D and 3-D datasets, and $T_{opt} = 0.5$ for the 4-D dataset, at which the lowest classification error was obtained, as shown in Table 4.4. The misclassified cases for the 3-D and 4-D datasets were fewer than those for the 2-D dataset. For the 4-D dataset, in particular, while the overlap between classes C_1 and C_2 decreased to a large extent with the inclusion of the SpO_2 dimension, the misclassification rate of the bradycardia and bradypnoea events was still 11.5%. The nature of density-based modelling is such that data that are rare (such as the bradycardia and bradypnoea events) are much more likely to be misclassified.

As the one-class method does misclassify less-represented data such as bradycardia and bradypnoea, even when the threshold T on $P(\kappa_{\mathbf{x}})$ is set at an optimal value, it is important to explore other methods such as two-class classification. In the next chapter, the performance of classifiers based on a two-class classification approach is compared.

Chapter 5

Two-Class Classification

5.1 Introduction

In this chapter, firstly standard two-class classification is used, and the results are compared with the one-class method of Chapter 4. A range of classifiers is considered, including linear and non-linear methods such as logistic regression, multi-layer perceptron, and support vector machines. Also proposed is a "hybrid" method such that (i) the overlap between the "normal" and "abnormal" classes is reduced, and (ii) misclassification of data from the less-represented events (such as bradycardia and bradypnoea) is reduced. The methods are first discussed for the 2-D dataset, and later for the 3-D and 4-D datasets.

Classifiers were trained using dataset $\mathcal{D}_1^{train}$, and their performance was tested on the held-out dataset $\mathcal{D}_1^{test}$ (see Table 4.3 for description of the two datasets). As in Chapter 4, classifications of data from classes C_2^{test} and C_1^{test} are considered to be positive and negative outcomes, respectively. Additionally, the following terms are defined:

$Sensitivity = \frac{TP}{TP+FN}$,

and

$Specificity = \frac{\mathrm{TN}}{\mathrm{TN+FP}}$,

where FP, FN, TP, TN and FN_{tachy} and FN_{brady} are as defined earlier in Section 4.5.1.

Sections 5.2 and 5.3 introduce the methods compared in this chapter. Subsequent sections discuss the performance of different classifiers for the two methods, followed by a discussion of the results.

Class	Estimated class	Description
$C_{2_{est}}^{test}$	"abnormal"	Contains data estimated by classifier as belonging to class C_2^{test}.
$C_{1_{est}}^{test}$	"normal"	Contains data estimated by classifier as belonging to class C_1^{test}.

Table 5.1: Description of classes as estimated by the classifier.

Algorithm 5.1 $\widehat{K}$-fold cross validation

Partition data into $\widehat{K}$ equal sets, where $\widehat{K}$=10
for iteration $\widehat{K}$, **do**
Validation set ← set $\widehat{K}$, training set← all sets except $\widehat{K}$
Train classifier, obtain results for classification using validation set
end for
combine results of $\widehat{K}$ folds (by averaging or otherwise)

5.2 Standard Two-class Classification

Classifiers were trained using dataset $\mathcal{D}_1^{train}$, and their performance was tested on the held-out dataset $\mathcal{D}_1^{test}$. The description of the estimated classes is presented in Table 5.1. Note that this description is different from the description presented in Table 4.3, as the classification methodology is different.

For any classifier that required the setting of architecture-dependent free variables, the dataset $\mathcal{D}_1^{train}$ was divided into training and validation datasets, and $\widehat{K}$-fold cross validation approach (see Algorithm 5.1) was used. $\widehat{K}$-fold cross-validation is a technique which can be used when the dataset is not sufficiently large to set aside separate training and validation sets, and so the available data are divided in $\widehat{K}$ folds (or non-overlapping subsets) of equal size. The algorithm is performed for $\widehat{K}$ iterations, and in each iteration, the $\widehat{K}^{\text{th}}$ fold is used as the validation set, while the remaining $\widehat{K}-1$ folds are used as the training set. As a result, all data are separately used both as training and validation data.

5.3 Hybrid Two-class Classification

Here a method is proposed to refine the class labels by optimising T, which is a threshold on $P(\kappa_{\mathbf{x}})$ (see section 4.2). The method involves training a classifier using a T-dependent subset of $\mathcal{D}_1^{train}$. Such a T-dependent training set is here refered to as $\tilde{\mathcal{D}}_1^{train}$. It contains classes $\tilde{C}_2^{train}$ and $\tilde{C}_1^{train}$, as defined in Table 5.2. To illustrate how the method works,

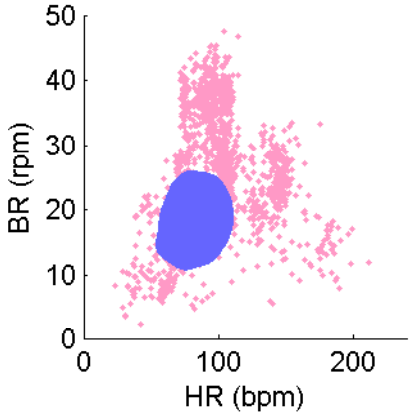


Figure 5.1: For $T = 0.5$, data from classes $\tilde{C}_2^{train}$ and $\tilde{C}_1^{train}$, shown in pink and light blue, respectively.

let the case of T=0 be considered first. In this case, classes $\tilde{C}_2^{train}$ and $\tilde{C}_1^{train}$ will be the same as classes C_2^{train} and C_1^{train} , or more simply $\tilde{\mathcal{D}}_1^{train}$ is the same as $\mathcal{D}_1^{train}$. As T is increased there will be fewer C_2^{train} data left in $\tilde{C}_2^{train}$, and similarly fewer C_1^{train} data left in $\tilde{C}_1^{train}$. Figure 5.1 presents a 2-D example of classes $\tilde{C}_2^{train}$ and $\tilde{C}_1^{train}$, for which $T = 0.5$. In effect, the separation between classes $\tilde{C}_2^{train}$ and $\tilde{C}_1^{train}$ will increase with the corresponding increase in T. For each value of T, the classifier will be trained using $\tilde{C}_2^{train}$ and $\tilde{C}_1^{train}$. As is the case in Chapter 4 and throughout this chapter, the trained classifier will be tested using the independent test set $\mathcal{D}_1^{test}$.

It is expected that the decision boundary generated by the classifier when trained with $T = 0.1$, for instance, will lead to a very large number of false positives and a very small number of false negatives; whereas the boundary generated by the classifier when trained with $T = 0.9$, for instance, will lead to a much larger number of false negatives and a much smaller number of false positives. Here it is postulated that (i) at an optimal value of T, T_{opt}, the classification error ϵ, will be the smallest, and that (ii) at this optimal value of T, the labels will be optimally refined.

It may be argued that this proposed hybrid method combines the one-class approach of Chapter 4 with the standard two-class classification methodology described in section 5.2 above. The hybrid method exploits the prior knowledge of class labels, while imposing the additional T-related constraint to try to refine those labels, given the knowledge that the labelling undertaken by the clinical experts was far from ideal.

Class	Clinical label	Description
$\tilde{C}_2^{train}$	"abnormal"	Contains those data from C_2^{train} for which $P(\kappa_{\mathbf{x}}) > T$.
$\tilde{C}_1^{train}$	"normal"	Contains those data from C_1^{train} for which $P(\kappa_{\mathbf{x}}) \leq T$, equal in number to $\tilde{C}_2^{train}$.

Class	Estimated label	Description
$C_{2_{est}}^{test}$	"abnormal"	Contains data estimated by classifier as belonging to class C_2^{test}.
$C_{1_{est}}^{test}$	"normal"	Contains data estimated by classifier as belonging to class C_1^{test}.

Table 5.2: The modified description of the training set $\tilde{\mathcal{D}}_1^{train}$. Below: The description of the (test set) classes estimated by the classifiers remain the same as in Table 5.1.

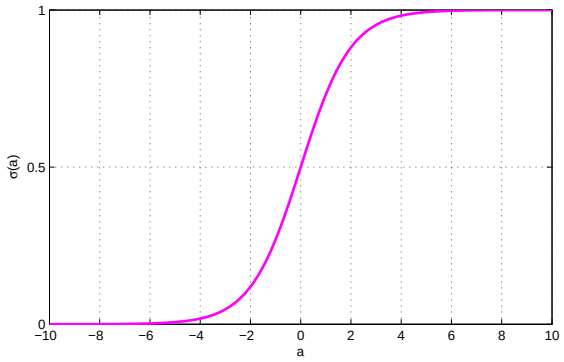


Figure 5.2: The sigmoid function $\sigma(a)$, against a. The figure is adapted from Figure 4.9 of [14].

5.4 Logistic Regression Model

The logistic regression model is considered as an initial standard classifier to establish the benchmark performance of a linear classifier, with which the performance of more complex classifiers may be compared. The logistic regression model uses the logistic sigmoid function [14]:

$$\sigma(a) = \frac{1}{1 + \exp(-a)} \tag{5.1}$$

where a is also known as the *logit function.* Figure 5.2 shows the sigmoid function which resembles an "S", and is also called a *squashing function.* For uni-dimensional data vector x, a is defined as the log of the ratio of the posterior probability density, $p(C_1|x)$, and the posterior probability density, $p(C_2|x)$:

$$a = \ln \frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)} = \ln \frac{p(C_1|x)}{p(C_2|x)} \tag{5.2}$$

where $p(C_1)$ and $p(C_2)$ are the priors for each class.

The posterior probability for each class, $p(C_l|x)$, is defined according to Bayes' theorem, so that for class C_1, according to Bayes' theorem:

$$p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1) + p(x|C_2)p(C_2)} \tag{5.3}$$

and the posterior probability for C_2, $p(C_2|x)$, can be defined similarly.

According to [14], it is possible that the class conditional densities are Gaussian, and all classes can have the same covariance function Σ. Thus, assuming this to be the case, the class conditional density for class C_l, $p(x|C_l)$, can be expressed as:

$$p(x|C_l) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}\left(x-\mu_l\right)^{\mathrm{T}}\Sigma^{-1}\left(x-\mu_l\right)\right\}. \tag{5.4}$$

The above equations can be manipulated to express the class probability as a function of the logistic sigmoid $\sigma(.)$:

$$p(C_1|x) = \sigma(\mathrm{w}^{\mathrm{T}}x + \mathrm{w}_0) \tag{5.5}$$

with the parameters inside the sigmoid function defined as:

$$\mathrm{w} = \Sigma^{-1}(\mu_1 - \mu_2) \tag{5.6}$$

and

$$\mathrm{w}_0 = -\frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \ln\frac{p(C_1)}{p(C_2)} \tag{5.7}$$

where μ_1 and μ_2 are the means of classes C_1 and C_2, respectively. Thus it has been possible to express the posterior class probability as a linear function of x, within the sigmoid σ. Finally, generalising the expression for a multi-dimensional vector $\mathbf{x}$, the two classes can be separated based on their class probabilities: $p(C_1|\mathbf{x}) = \sigma(\mathrm{w}^{\mathrm{T}}\mathbf{x})$ and $p(C_2|\mathbf{x}) = 1-p(C_1|\mathbf{x})$. A vector $\mathbf{x}$ with class probability $p(C_1|\mathbf{x}) = 0.5$ has equal probability of being assigned to class C_1. For a class probability $p(C_1|\mathbf{x}) \geq 0.5$, the vector $\mathbf{x}$ is assigned to class C_1; otherwise, it is assigned to class C_2.

Figure 5.3 shows the results of classification when using the logistic regression model. The left plot shows classification boundary for the 2-D training set $\mathcal{D}_1^{train}$, and the right

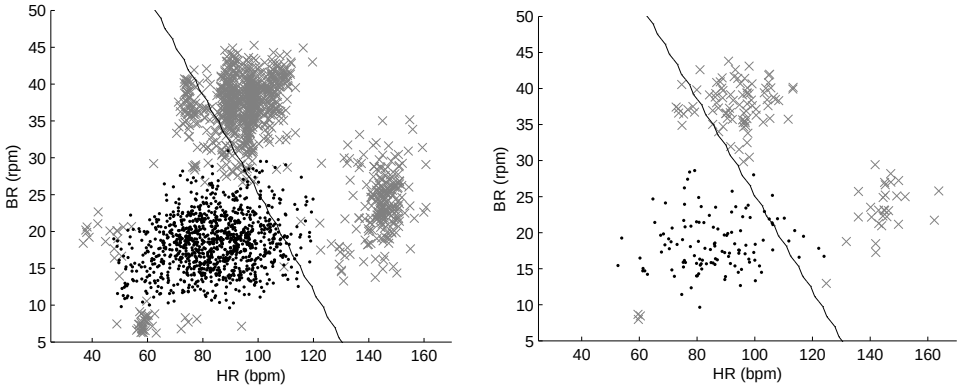


Figure 5.3: Training set $\mathcal{D}_1^{train}$, (left) and test set $\mathcal{D}_1^{test}$, (right) . The decision boundary generated by the logistic regression classifier is shown by the straight black line. Data from classes C_2^{train} and C_1^{train} (left plot), and from classes C_2^{test} and C_1^{test} (right plot) are shown by grey crosses and black dots, respectively.

plot shows the classification boundary for the 2-D test set $\mathcal{D}_1^{test}$. From Figure 5.3 observed that the classifier fails to classify correctly the bradycardia (low HR) and bradypnoea (low BR) points nearer the origin. The logistic regression classifier can only generate linear decision boundaries, and thus cannot correctly classify data that are not separable by a simple linear boundary. The logistic regression classifier was therefore not used further in the investigation.

5.5 Multi-Layer Perceptron Model

The multi-layer perceptron (MLP) model was introduced in the 1960s [156]. Figure 5.4 shows the schematic diagram for an *I-J-K* (or two-layer) model, where I is the number of inputs, J is the number of *hidden units*, and K is the number of output units. For the 2-D classification problem, $I = 2$ (corresponding to the dimensions HR and BR), and $K = 1$, while the selection of J will be optimised during the process of selecting the network architecture appropriate for the dataset.

The output of the first layer can be expressed as:

$$\vartheta_j = \sum_{i=1}^{d} x_i w_{ij} + x_0 w_{0,j} \tag{5.8}$$

where $i = 1, ..., d$ is the index of dimensions in the input vector, $j = 1, ..., J$ is the index

of the hidden unit, and w_{ij} represents the i-to-j weight for the hidden unit j.

ϑ_j passes through an *activation function* $f(.)$ such as the *sigmoid* function [51] (shown in Figure 5.2), such that the output of the hidden units is given by:

$$y_j = f(\vartheta_j) \tag{5.9}$$

For each output unit:

$$\vartheta_k = \sum_{j=1}^{J} y_j w_{jk} + y_0 w_{0,k} \tag{5.10}$$

where w_{jk} represents the j-to-k weights for the output unit k and $y_0 = 1$, allowing a bias to be added. Finally:

$$z_k = f(\vartheta_k) \tag{5.11}$$

and the overall expression for the two-layer i-$j-k$ feedforward operation can thus be summarised as:

$$z_{k=} f\left[\left(\sum_{j=1}^{J} w_{jk} f\left(\sum_{i=1}^{d} x_i w_{ij} + x_0 w_{0,j}\right) + y_0 w_{0,k}\right)\right] \tag{5.12}$$

The weights in Equation 5.12 are initialised randomly and are then learned by using *error back-propagation*, in which the error E between actual and target outputs, over all the dataset, is back-propagated to update the network weights, until the error E is minimised [156]. This is typically performed using *gradient descent*;

$$\triangle w = -\eta \frac{\partial E}{\partial w} \tag{5.13}$$

where η is *learning rate*.

5.5.1 Choice of Network Architecture

A two-layer network (with sigmoid activation functions) has been shown to perform with sufficient accuracy for most problems [156], and hence such a model, with a single layer of hidden units, is selected for use in this work. This leaves the number of hidden units, J, as the main parameter to select when defining the architecture. In order to determine

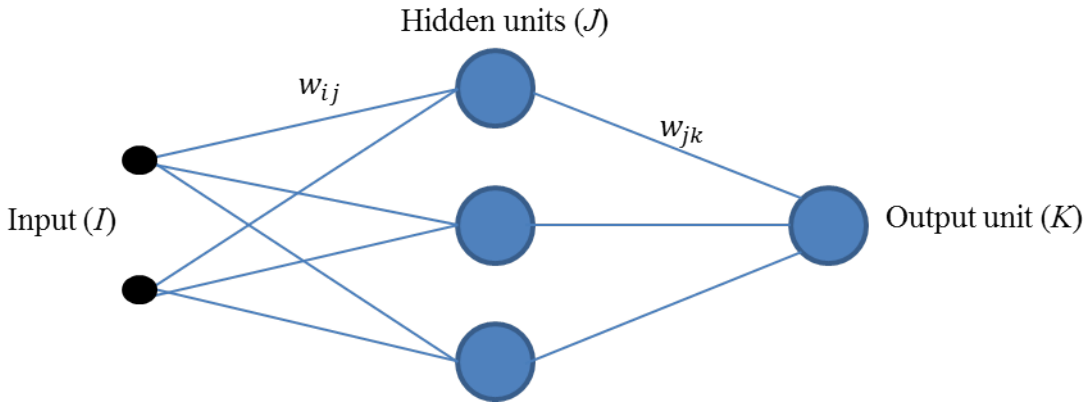


Figure 5.4: A 2-3-1 multi-layer perceptron for two-class classification. Bias weights are not shown.

J, the $\widehat{K}$-fold cross validation approach described in section 5.2 is used here. In order to avoid confusion, $\widehat{K}$ is used here to denote the number of folds described in algorithm 5.1, as K is being used in the context of the MLP to denote the output unit of the MLP. As the choice of J is typically associated with the complexity of the training data, the value of J obtained using the training and validation subsets of $\mathcal{D}_1^{train}$ should also be valid when used later with $\tilde{\mathcal{D}}_1^{train}$ for the hybrid method, as the data complexity does not differ between the two datasets.

To learn the network weights w_{jk} and w_{ij}, it is recommended to train the network at least ten times using different values of J [156]. Therefore, a 10-fold cross validation approach was performed for a range of values of J, using $\mathcal{D}_1^{train}$. The initial weights were initialised to take values in the range ± 0.01, so that the initial values of all y_j and y_k would lie near the centre of the sigmoid function. Additionally, a value of $\eta = 0.1$ is considered reasonable in the literature [156] and was used in this work. For each value of J, the classification error across ten validation sets was averaged. The results are shown in Figure 5.5, which shows that the lowest mean classification error ϵ, as well as the smallest ± 1 std occurs for $J = 4$. The value of $J = 4$ was therefore selected as the optimal number of hidden units.

Once the network architecture had been fixed, the MLP was trained independently first as a standard classifier, and then using the hybrid method. As before, the resulting classifier in each case was tested using the independent dataset $\mathcal{D}_1^{test}$.

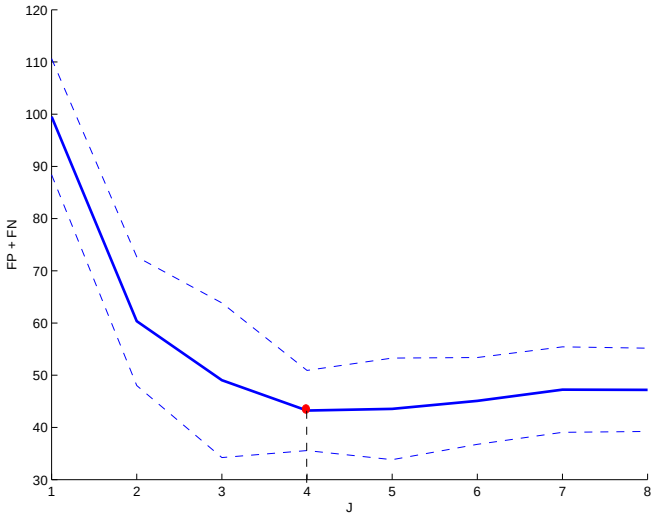


Figure 5.5: Average classification error for a range of values of J, obtained using 10-fold cross-validation. The mean and ± 1 std error are shown in solid and dashed blue lines, respectively. The lowest mean classification error occurs for $J = 4$.

5.5.2 MLP Trained with Hybrid Method

As discussed in section 5.3, the hybrid method was designed to identify that value of T for which a classifier would provide the "optimal" boundary between the two classes C_1^{test} and C_2^{test}, such that the classification error ϵ was minimised. Figure 5.6 shows the results of applying the hybrid method using the MLP. The experiment was carried out for $T = 0.1$ to $T = 0.9$. The upper bound on the range of T was not extended to $T = 1$, because use of this threshold would leave set $\tilde{C}_2^{train}$ empty. For each value of T, the MLP's performance was evaluated using the same test set $\mathcal{D}_1^{test}$, and the classification error ϵ was recorded. As the initialisation of the MLP is a random process and can affect results, the experiment was repeated 50 times and the average results are shown in Figure 5.6. It can be seen that the initialisation process does not cause significant deviations in the FP and FN outcomes, over the range of T.

Referring back to Figure 4.8, as T increases, the decision boundary widens in the data space. It is expected that this widening boundary will begin to include many data from the class C_1. Figure 5.6(a) illustrates how the increasing value of T (and the correspondingly widening decision boundary) influences the FP and FN outcomes. As the boundary expands to increase the coverage of C_1, the number of FN increases, while the number of FP decreases. Thus an optimal value of T_{opt} is one at which FP and FN are equal;

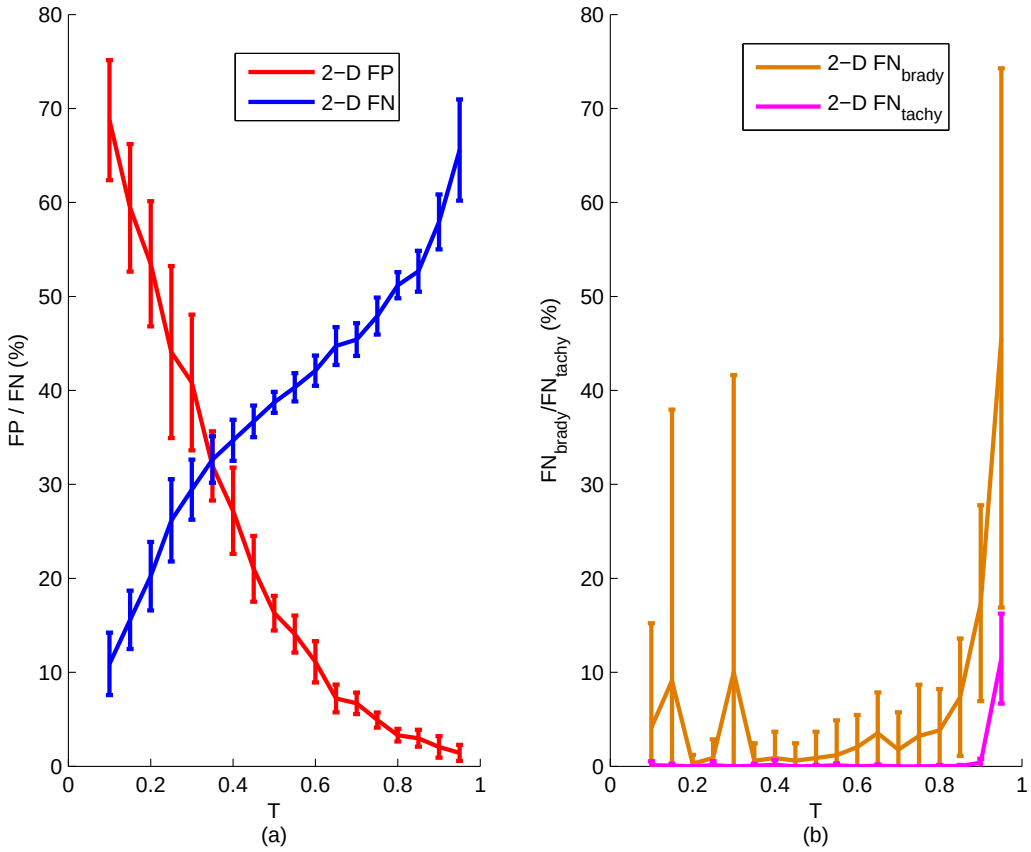


Figure 5.6: Misclassification performance of the hybrid method for the 2-D dataset $\mathcal{D}_1^{test}$, using the MLP model. (a) The mean percentage of false positives (FP) averaged over 50 experiments is shown by the red line, and a confidence interval at ±1std FP is shown by the vertical red bars. The mean percentage of false negatives (FN) averaged over 50 experiments is shown by the blue line, and a confidence interval at ±1 std FN is shown by the vertical blue bars. (b) The percentage of FN_{brady}, and the percentage of FN_{tachy} averaged over 50 experiments, are shown in orange and bright pink, respectively. T_{opt} =0.35.

i.e., the value of T at which the classifier estimated $C_{1_{est}}^{test}$ and $C_{2_{est}}^{test}$ classes would most closely match the original labels C_1^{test} and C_2^{test}. From Figure 5.6, $T_{opt} = 0.35$ is obtained. Using a value of $T \leq T_{opt}$ tends towards misclassification of data from class C_2^{test}; conversely when $T \geq T_{opt}$, the classifier tends to misclassify data from C_1^{test}. Figure 5.6(b) shows the corresponding break down of FNs into FNs corresponding to bradycardia- and bradypnoea-related (FN_{brady}), and tachycardia- and tachypnoea-related ($\text{F}N_{tachy}$) outcomes.

The 2-4-1 MLP was trained both as a standard classifier and using the hybrid method, and in each case the resulting classifier was tested using the independent dataset $\mathcal{D}_1^{test}$. Figure 5.7(a) shows the results obtained for the standard classifier, and Figure 5.7(b) shows the result using the hybrid method at $T_{opt} = 0.35$. Note that the FNs include many

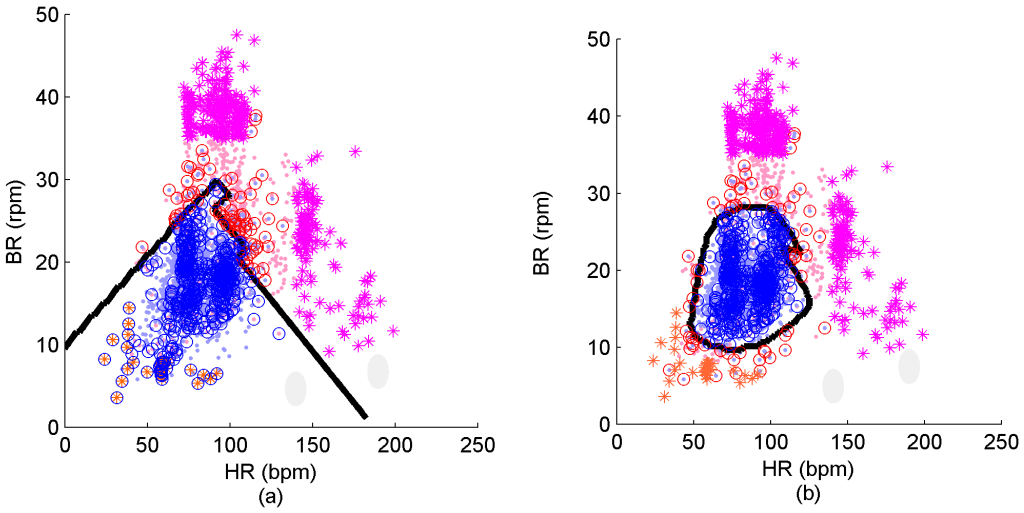


Figure 5.7: Classification results for the 2-D dataset $\mathcal{D}_1^{test}$ (a) using a standard MLP classifier. (b) trained using the hybrid method, at $T_{opt} = 0.35$. For the test set $\mathcal{D}_1^{test}$, light blue and light pink dots show data from classes C_1^{test} and C_2^{test}, respectively. C_2^{test} data corresponding to tachycardia and tachypnoea, and C_2^{test} data corresponding to bradycardia and bradypnoea according to Table 3.2 are shown by bright pink and orange stars, respectively. FNs are encircled in blue and FPs are encircled in red. The decision boundary is plotted in black.

of the data from C_2^{test} which lie completely overlapped with the data from C_2^{test}, in the central region of the figure. The C_2^{test} data that lie in this central region correspond to a low SpO_2 "abnormal" cluster, which will be explored in subsequent sections which discuss the full 4-D dataset. Therefore this "central region" overlap exists in the 2-D dataset, but not in the 4-D dataset where the SpO_2dimension is included, and accounts for a much higher FN rate for the 2-D dataset as compared to that for the 3-D and 4-D datasets which will be discussed in subsequent sections.

While the MLP classifier qualitatively appears to perform better classification than that achieved by the LR decision boundary (see Figure 5.3), the under-represented data from C_2^{test} are still misclassified when using the standard classifier. In contrast with this, the use of a threshold T gives a decision boundary that better classifies those data from each of the under-represented physiological abnormalities, as shown in Figure 5.7(b). A comparison of the classification methods discussed in this section will be presented in section 5.8.

5.6 Support Vector Machines

Support Vector Machines (SVM) were proposed in the 1990s by Vapnik et al. [41], and they have been used extensively for novelty detection problems, such as the one discussed in this thesis. An SVM defines a hyperplane to separate data from different classes [102]. Using a kernel function, SVMs are well-suited to dealing with non-linearly separable data as the kernel function allows for projection into higher dimensions such that the projected data become separable. The introduction of "slack variables" allows a proportion of data to fall on the "wrong" side of the decision boundary and this increases the SVM model's ability to generalise to previously unseen data. A brief description of SVM theory is introduced here; for a more detailed study the reader is referred to [14, 20].

SVMs are often called "maximum margin classifiers", using a two-class model of the form [14] :

$$y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b \tag{5.14}$$

where $w^T\phi(\mathbf{x})$ represents the weighted transformation of the input vector $\mathbf{x}$, and b is a bias parameter. This defines $y(\mathbf{x})$, which is the estimate of the class for each point. In section 5.5 it was seen that the MLP model is typically initialised in a random manner and the results of training depend on this random initialisation. This necessitates the training process to be repeated several times to determine the parameter values which result in the best performance. The SVM model overcomes this limitation by replacing the training process with quadratic optimisation. The SVM margin, which training seeks to maximise, is shown in Figure 5.8, and refers to the minimum Euclidean distance between the decision boundary and the closest points in the training set. The SVM learning process aims to maximise this margin, so as to obtain a decision boundary, or separating hyperplane, that best separates points from the two classes. Points that lie on hyperplanes with the maximum margin from the decision boundary are known as the support vectors [14, 20].

For the hyperplanes S, A, and B in the figure, it can be seen that the $y|_A = \mathbf{w}.\underline{\mathbf{x}}+b = 1$, $y|_B = \mathbf{w}.\underline{\mathbf{x}} + b = -1$, and $y|_S = \mathbf{w}.\underline{\mathbf{x}} + b = 0$. Thus for at least one combination of $\mathbf{w}$ and b , the constraint $|y_n(\mathbf{w}.\underline{\mathbf{x}}_n + b)| > 0$ is satisfied for all n in training data. Considering support vectors $\underline{\mathbf{x}}_1$ and $\underline{\mathbf{x}}_2$, the margin can be estimated by calculating the projection of the distance between $\underline{\mathbf{x}}_1$ and $\underline{\mathbf{x}}_2$ onto the normal vector of the separating hyperplane,

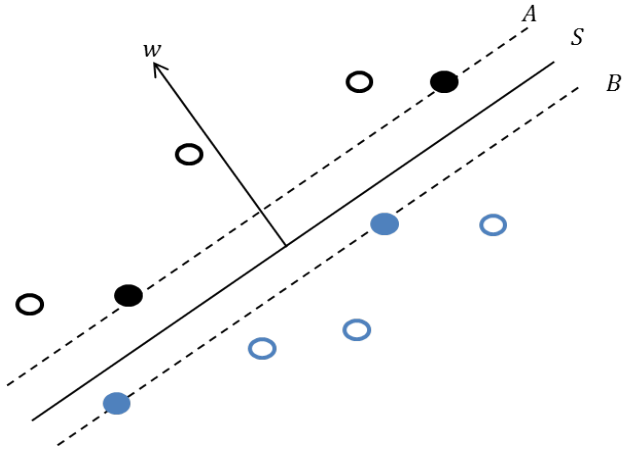


Figure 5.8: Hyperplanes A, B and S illustrating the SVM model. The support vectors are represented by the solid circles lying on hyperplanes A and B. $\mathbf{w}$ is normal to the hyperplane S. $y|_A = 1$, $y|_B = -1$, and $y|_S = 0$.

computed as [20]:

$$\left(\frac{\mathbf{w}^\top}{\|\mathbf{w}\|}(\underline{\mathbf{x}}_1 - \underline{\mathbf{x}}_2)\right) = \frac{2}{\|\mathbf{w}\|} \tag{5.15}$$

From Equation 5.15 it may be seen that smaller the norm of the weight vector $\mathbf{w}$, the larger is the margin. The task of maximising the margin therefore reduces to one of minimising $||\mathbf{w}||$, subject to the constraint

$$|y_n(\mathbf{w}.\underline{\mathbf{x}}_n + b)| \geq 1 \tag{5.16}$$

This problem can be solved using Lagrange multipliers $a_n \geq 0$, which are a set of variables introduced to minimise (or maximise) a function according to given constraints. A detailed description of Lagrange multipliers can be found in Appendix E in [14]. The resulting Lagrangian function,

$$L(\mathbf{w}, b, a) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{n=1}^{N} a_n \left\{|\mathbf{w}^T \phi\left(\underline{\mathbf{x}}_n\right) + b| - 1\right\} \tag{5.17}$$

can then be solved to yield

$$y(\underline{\mathbf{x}}) = \sum_{n=1}^{N} a_n |k(\underline{\mathbf{x}}, \underline{\mathbf{x}}_n) + b| \tag{5.18}$$

where $k(\underline{\mathbf{x}}, \underline{\mathbf{x}}^{'}) = \phi(\mathbf{x})^T \phi(\underline{\mathbf{x}}^{'})$ is a kernel function which is introduced to map the training

data into a higher-dimensional feature space where the classes can be separated linearly. The sign of the resulting $y(\underline{\mathbf{x}})$ determines to which class the point ($\underline{\mathbf{x}}$) is assigned.

5.6.1 Soft Margins

One of the advantages of the SVM over other classifiers is that it allows for some misclassifications of the training data which helps the model to avoid over-fitting the training data. A slack variable ζ_n (where $n = 1, ..., m$) is introduced into Equation 5.16, to give:

$$y_n(\mathbf{w}.\underline{\mathbf{x}}_n + b) \geq 1 - \zeta_n \tag{5.19}$$

This replaces the constraint $a_n \geq 0$ by $0 \leq a_n \leq C$. The additional variable C cannot be directly associated with a tangible characteristic of the model, and has to be optimised during the optimisation process [20]. Scholkopf et al. [143], Bishop [14], and Clifton [34, 38] recommend the use of parameter ν, related to C by

$$\nu = \frac{1}{Cm} \tag{5.20}$$

instead of C, as it can be understood intuitively to serve as an upper bound on the number of training data lying on the wrong side of the decision boundary, as well as a lower bound on the number of support vectors [14].

5.6.2 Optimising the SVM Model

In addition to the choice of ν, another factor to be considered while optimising the SVM architecture is the choice of kernel function used to map the training data into the higher-dimensional feature space associated with the kernel. Tax and Duin (1999) [162] investigated the Gaussian kernel function and a polynomial kernel function and found that the former performs best for many practical data sets. A Gaussian kernel function was therefore selected for the model. The width of the kernel function σ, is typically determined by experimentation using validation sets [20]. This presented two parameters, ν and σ to optimise in the SVM model used here. While ν determines the number of support vectors, σ controls the complexity of the decision boundary [38]; the larger the value of

σ, the more complex the contour described by the decision boundary. Hsu et al. [76] recommend using a "grid search" to find optimal values of ν and σ.

For the optimisation process, a 10-fold cross-validation approach, similar to that used in section 5.5.1, was taken, by dividing dataset $\mathcal{D}_1^{train}$ into training and validation folds. The grid search was repeated for each of the 10 folds, over the ranges $0 \leq \sigma \leq 5$ (with a step size of $\sigma = 0.5$), and $0.01 \leq \nu \leq 0.1$ (with a step-size of $\nu = 0.01$). The different combinations of ν and σ were validated using the validation subset of $\mathcal{D}_1^{train}$, in order to determine which pair resulted in the smallest classification error (for that validation subset). This process yielded the optimal values of $\nu = 0.01$, and $\sigma = 1$. It was observed that other combinations of ν and σ resulted in an unnecessarily large number of support vectors, while others significantly over-fitted the data, thus giving poor generalisation.

Once the model parameters had been determined, the SVM model was tested using the independent dataset $\mathcal{D}_1^{test}$, for the standard SVM classifier, and one trained with the hybrid method. With the latter, the same threshold optimising experiment as described in section 5.5.2 was carried out, resulting in the candidate T_{opt} value occurring at the point of intersection of the FP and FN curves as shown in Figure 5.9(a). In this instance, the T_{opt} value was 0.35, which is the same as the T_{opt} value of 0.35 obtained using the MLP.

Figure 5.10 shows that while the standard SVM classifier results in fewer FPs, it also causes a large number of FNs, particularly those related to the bradycardia and bradypnoea region. In contrast the misclassification of the bradycardia and bradypnoea data appears reduced with the hybrid method trained at $T_{opt} = 0.35$. A quantitative comparison of the standard and hybrid classifiers for the 2-D dataset is presented in Section 5.8.

5.7 Application to 3-D and 4-D Datasets

Figure 5.11 (a) shows a 3-D view of dataset set $\mathcal{D}_1^{test}$, where the HR, BR and SDA are shown. In Figure 5.11 (b), for the same dataset, the HR, BR, and SpO_2 are shown. As mentioned previously in Section 4.3, the overlap between classes C_1^{test} and C_2^{test} is reduced

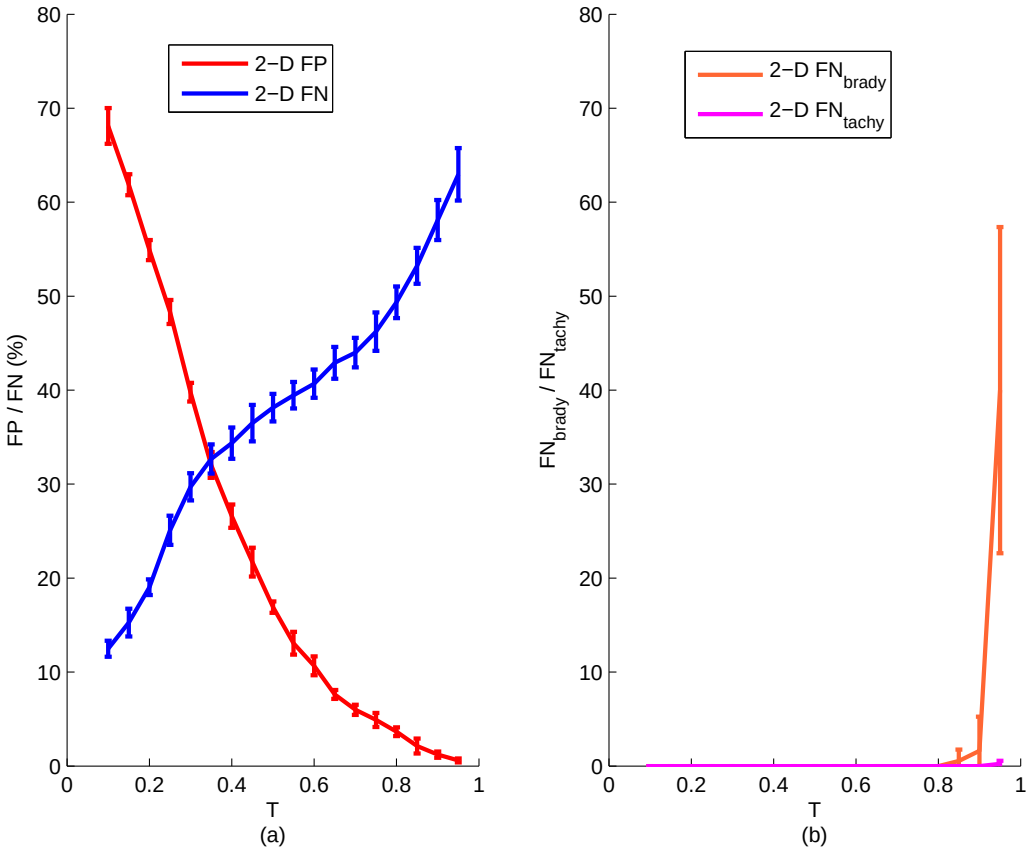


Figure 5.9: Misclassification performance of the hybrid method for the 2-D dataset $\mathcal{D}_1^{test}$, using the SVM model. (a) The mean percentage of false positives (FP) averaged over 50 experiments is shown by the red line, and a confidence interval at $\pm$1std FP is shown by the vertical red bars. The mean percentage of false negatives (FN) averaged over 50 experiments is shown by the blue line, and a confidence interval at $\pm$1 std FN is shown by the vertical blue bars. (b) percentage of FN$_{brady}$, and the percentage of FN$_{tachy}$, averaged over 50 experiments are shown in orange and bright pink, respectively. T_{opt} =0.35.

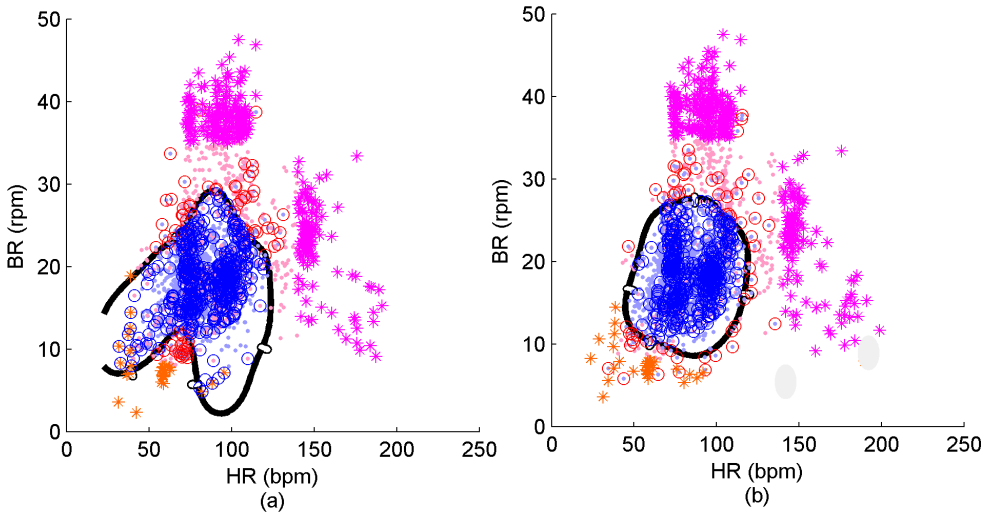


Figure 5.10: Classification results for the 2-D dataset $\mathcal{D}_1^{test}$ (a) using a standard SVM classifier. (b) trained using the hybrid method, at $T_{opt} = 0.35$. For the test set $\mathcal{D}_1^{test}$, light blue and light pink dots show data from classes C_1^{test} and C_2^{test}, respectively. C_2^{test} data corresponding to tachycardia and tachypnoea, and C_2^{test} data corresponding to bradycardia and bradypnoea according to Table 3.2 are shown by bright pink and orange stars, respectively. FNs are encircled in blue and FPs are encircled in red. The decision boundary is plotted in black.

in higher dimensions as compared to the overlap in two dimensions. This reduction in the overlap is easier to see in Figure 5.12 which shows dataset $\mathcal{D}_1^{train}$ at $T = 0.5$ (which is shown just to demonstrate the overlap). Here it may be seen that while there is still some overlap between the "normal" cluster and the "abnormal" cluster representing the low SDA (in Figure 5.12(a)), there is a clear separation between the "normal" class and the abnormal cluster representing low SpO_2 below 80% (in Figure 5.12(b)). In fact, the pink "abnormal" clusters appear mostly to be positioned on the circumference of the central blue "normal" cluster. From these 3-D views it is therefore expected that the classifiers will perform better for the 3-D and 4-D datasets than for the 2-D datasets.

The standard MLP and SVM classifiers were applied to the 3-D dataset and the 4-D dataset. The MLP and SVM classifiers were also applied to the 3-D and 4-D datasets using the hybrid method. For the 3-D dataset, HR, BR, and SDA were included. The SpO_2 was added in the full 4-D treatment. The architectures of the MLP and the SVM were set in the same manner, by repeating the optimisation process described in sections 5.5 and 5.6, respectively, giving $J = 4$ for the MLP, and $\nu = 0.01$ and $\sigma = 1$ for the SVM.

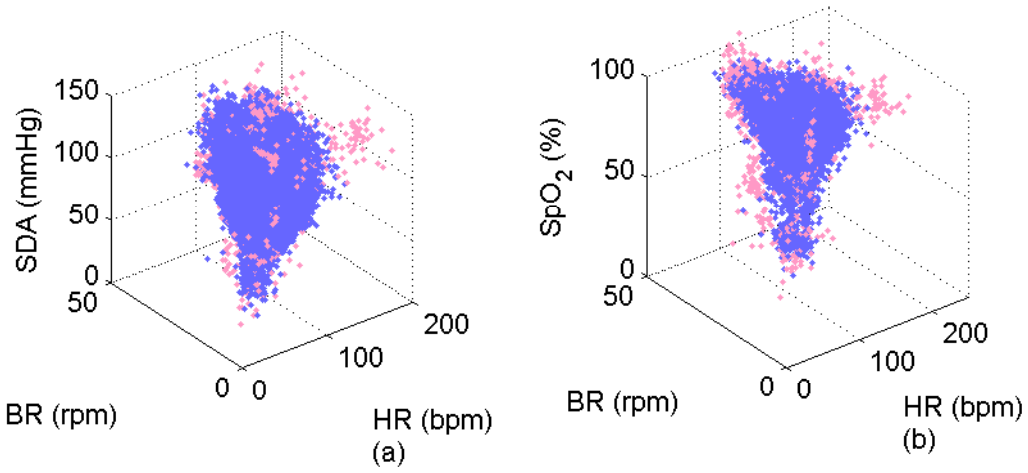


Figure 5.11: A 3-D view of dataset $\mathcal{D}_1^{test}$. (a) HR (x-axis), BR(y-axis) and SDA (z-axis) are shown. (b) HR (x-axis), BR (y-axis) and SpO_2 (z-axis) are shown. Classes C_1^{test} and C_2^{test} are shown in light blue and pink, respectively.

Figures 5.13 and 5.14 show the classification errors obtained using the hybrid MLP and SVM classifiers, respectively. As expected, for both MLP and SVM, the FP and FN percentages at the respective values of T_{opt} are lower for the 4-D dataset than they are for the 3-D dataset. However, the percentage of bradycardia- or bradypnoea-related false negatives, FN_{brady}, is higher for the 4-D dataset. These results are discussed in detail in the next section.

5.8 Results

Classification performance of the standard classifiers and the hybrid method for the 2-D, 3-D and 4-D datasets is discussed. Results for the one-class method from Table 4.4 are also included for comparison. Table 5.3 summarises the results of applying the standard two-class, hybrid, and one-class classifiers to the 2-D dataset. For the hybrid method, the results at T_{opt} are shown. The average sensitivity and specificity (as defined in Section

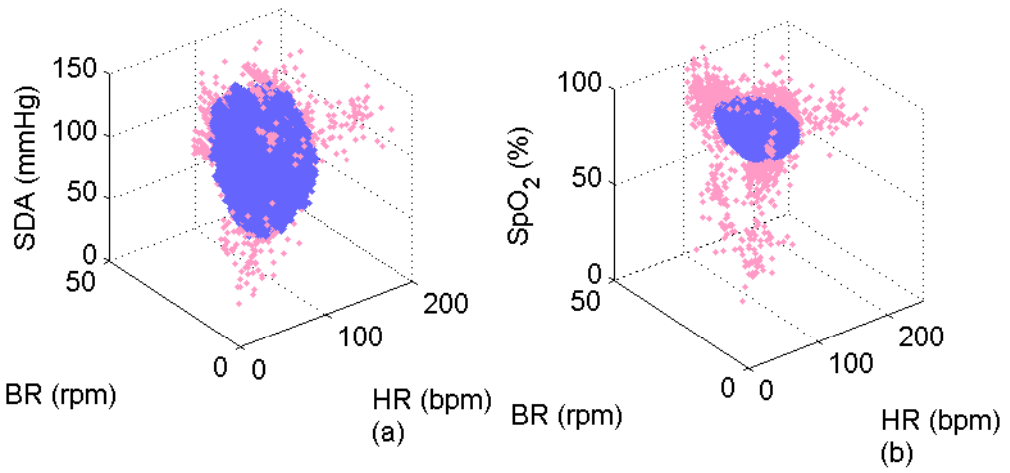


Figure 5.12: A 3-D view of dataset $\mathcal{D}_1^{train}$ at $T = 0.5$. (a) HR (x-axis), BR (y-axis) and SDA (z-axis) are shown. (b) HR (x-axis), BR (y-axis) and SpO_2 (z-axis) are shown, for classes C_1^{train} and C_2^{train} in light blue and pink, respectively.

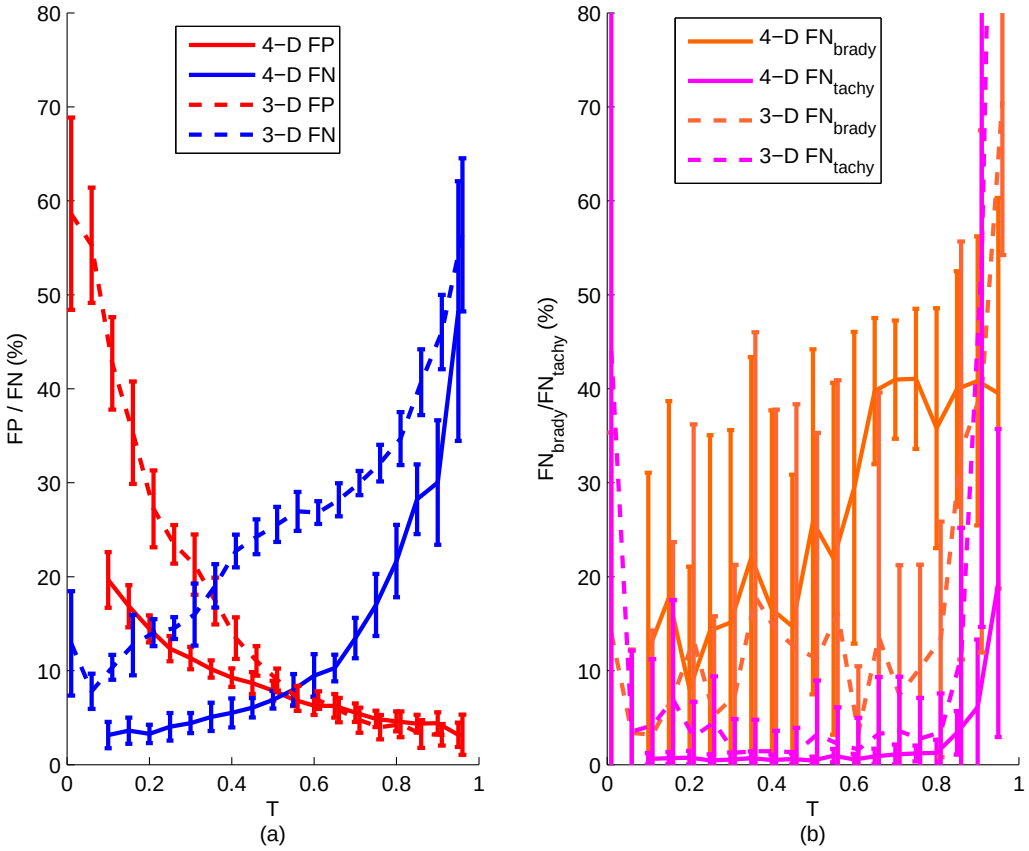


Figure 5.13: Misclassification performance of the hybrid method for the 3-D (dashed lines) and 4-D (solid lines) dataset $\mathcal{D}_1^{test}$, using the MLP model. (a) The mean percentage of false positives (FP) averaged over 50 experiments is shown by red line, and a confidence interval at ±1std FP is shown by the vertical red bars. The mean percentage of false negatives (FN) averaged over 50 experiments is shown by blue line, and a confidence interval at ±1 std FN is shown by the vertical blue bars. (b) The percentage of FN_{brady}, and the percentage of FN_{tachy} averaged over 50 experiments, are shown in orange and bright pink, respectively. For 3-D , T_{opt} =0.36, and for 4-D, T_{opt} =0.5.

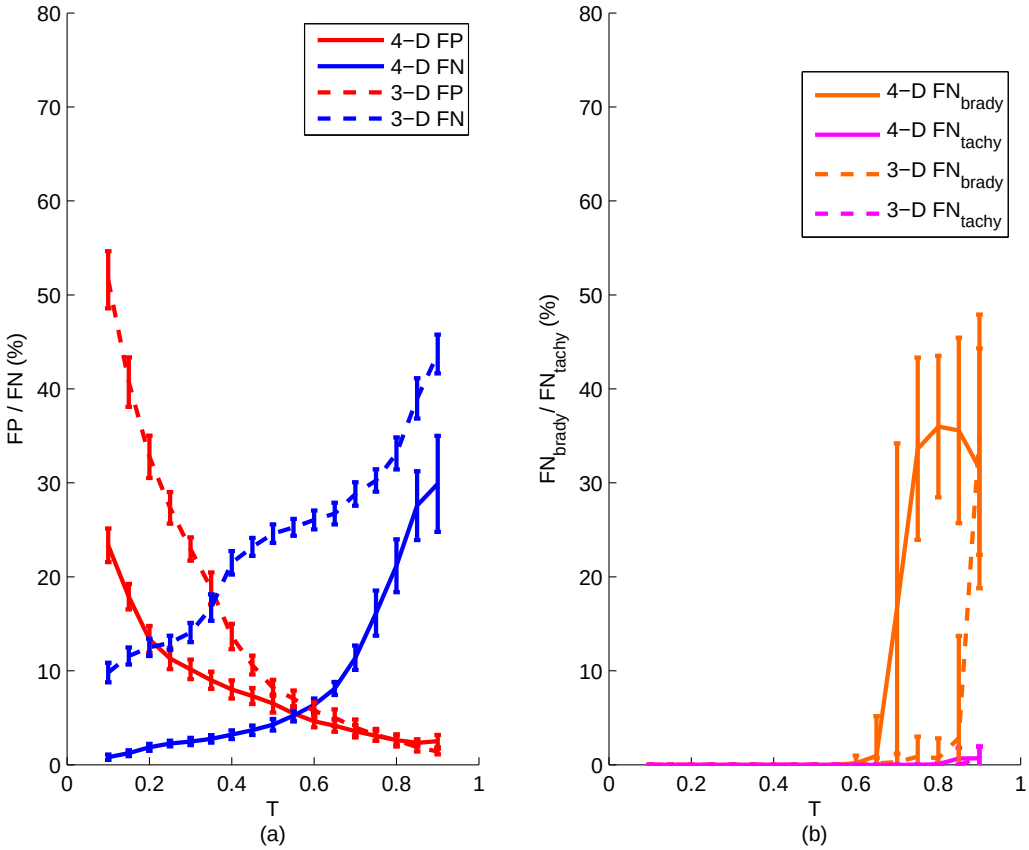


Figure 5.14: Misclassification performance of the hybrid method for the 3-D (dashed lines) and 4-D (solid lines) dataset $\mathcal{D}_1^{test}$, using the SVM model. (a) The mean percentage of false positives (FP) averaged over 50 experiments is shown by red line, and a confidence interval at $\pm$1std FP is shown by the vertical red bars. The mean percentage of false negatives (FN) averaged over 50 experiments is shown by blue line, and a confidence interval at $\pm$1 std FN is shown by the vertical blue bars. (b) The percentage of FN_{brady}, and the percentage of FN_{tachy} averaged over 50 experiments, are shown in orange and bright pink, respectively. For 3-D , T_{opt} =0.34, and for 4-D, T_{opt} =0.55.

5.1), obtained over 50 experiments, are shown. As the ±1std errors of the sensitivity and specificity over 50 experiments was negligible and and similar across classifiers (1) - (5) shown in the table, they are not shown. The FN are further categorised: FN_{brady} refer to the percentage of bradycardia- and bradypnoea-related false negative cases. FN_{tachy} refer to the percentage of tachycardia- and tachypnoea-related false negative cases. The mean ±1std errors, obtained from the average of 50 experiments are also shown.

Tables 5.4 and 5.5 present the same summary of results as described above, for the 3-D and 4-D datasets, respectively. From these results, the following key observations can be made.

	Classifier	Sensitivity (%)	Specificity (%)	Classification error (%)			
				FP	FN	FN_{tachy}	FN_{brady}
1	Two-class MLP	56.7	91.6	8.4 ± 3.2	43.3±2.7	0.1±0.1	2.2 ± 2.5
2	Two-class SVM	57.5	93.1	6.9 ± 2.6	42.5±3.3	0	4.3 ± 6.2
3	Two-class MLP ($T_{opt} = 0.35$)	67.4	67.4	32.6±2.5	32.6±3.7	0	0.6 ± 1.9
4	Two-class SVM ($T_{opt} = 0.35$)	68.0	68.0	32.0±1.6	32.0±1.4	0	0
5	One-class classification ($T_{opt} = 0.4$)	68.0	66.8	33.2±1.2	33.2±1.5	0	0

Table 5.3: Classification performance for the 2-D dataset $\mathcal{D}_1^{test}$ using the standard two-class (1) MLP and (2) SVM, and the hybrid (3) MLP and (4) SVM, and (5) the one-class method reproduced from Table 4.4.

	Classifier	Sensitivity (%)	Specificity (%)	Classification error (%)			
				FP	FN	FN_{tachy}	FN_{brady}
1	Two-class MLP	79.6	89.7	10.3±1.3	20.4±1.0	0.2 ± 0.2	39.7 ± 30.8
2	Two-class SVM	82.2	92.5	7.5 ± 0.8	17.8±1.3	0	1.2 ± 2.1
3	Two-class MLP ($T_{opt} = 0.36$)	80.6	80.7	19.3±1.8	19.3±3.1	1.6 ± 3.6	28.2 ± 34.4
4	Two-class SVM ($T_{opt} = 0.34$)	80.3	81.6	18.4±1.3	19.7±1.1	0	0
5	One-class classification $T_{opt} = 0.4$	79.2	79.2	20.8±1.2	20.8±0.9	0	0

Table 5.4: Classification performance for the 3-D dataset $\mathcal{D}_1^{test}$ using the standard two-class (1) MLP and (2) SVM, and the hybrid (3) MLP and (4) SVM, and (5) the one-class method reproduced from Table 4.4.

	Classifier	Sensitivity (%)	Specificity (%)	Classification error (%)			
				FP	FN	FN_{tachy}	FN_{brady}
1	Two-class MLP	93.0	91.9	8.1 ± 1.3	6.9 ± 1.1	4.1 ± 4.3	12.8 ± 7
2	Two-class SVM	94.2	95.6	4.4 ± 1.1	27.7±3.7	0	0.4 ± 0.3
3	Two-class MLP ($T_{opt} = 0.5$)	92.1	93.1	6.9 ± 0.9	7.8 ± 1.0	0.4 ± 0.4	25.8 ± 18.3
4	Two-class SVM ($T_{opt} = 0.55$)	94.6	94.8	5.2 ± 0.6	5.4 ± 0.1	0	0
5	One-class classification ($T_{opt} = 0.5$)	91.7	91.7	8.3±0.5	8.3±0.6	0	11.5±3.7

Table 5.5: Classification performance for the 4-D dataset $\mathcal{D}_1^{test}$ using the standard two-class (1) MLP and (2) SVM, and the hybrid (3) MLP and (4) SVM, and (5) the one-class method reproduced from Table 4.4.

2-D Dataset

- Both the MLP and the SVM trained using the hybrid method result in a smaller percentage of FNs (FN=32.0 % and FN=32.0 % for the MLP and the SVM, respectively) than the standard MLP (FN=43.4%) and the standard SVM (FN=42.5%). Consequently the sensitivity of the hybrid method (which is 67.4% for the MLP and 68.0% for the SVM) is higher than that of the standard MLP (56.7%) and the standard SVM (57.5%).

- For either the MLP or the SVM, both the hybrid method and the one-class method result in fewer FN_{brady} than the standard classifiers.

- The FP and FN results for the MLP and the SVM using the hybrid method, and the results for the one-class method are comparable.

3-D Dataset

- Overall, the percentages of FP and FN are lower for the 3-D dataset than for the 2-D dataset. This is reflected in the sensitivity which for the 3-D dataset using the standard MLP is 79.6% and using the SVM is 82.2%. The sensitivity of the same two classifiers using the hybrid method is comparable (80.6% for the MLP, and 80.3% for the SVM), as well as the sensitivity of the one-class method (which is 79.2%).

- Using either the standard or hybrid method, the MLP results in more FN_{brady} than it does for the 2-D dataset.

- As with the 2-D dataset, the hybrid method results in fewer FN_{brady} than the standard classifiers.

4-D Dataset

- There are fewer FP and FN for the 4-D dataset than in either the 2-D or the 3-D datasets. This is can be seen by considering the sensitivity, which is greater than 90% for all methods used.

- While the one-class method, the hybrid MLP, and the standard MLP result in more than 11% FN_{brady}, the hybrid SVM results in 0% FN_{brady}.

- While the standard SVM also results in very few FN_{brady}, it also results in a much higher percentage of FN (27.73%) than the hybrid SVM does (5.46%). Therefore the hybrid 4-D SVM at the optimal threshold T_{opt} =0.55 is considered to be the best model.

Overall, these results suggest that

- The SVM performs as well as, or better than, the MLP, a result which is not contrary to expectation. Intuitively one expects that the SVM may perform better than the MLP since it is designed to project data into a higher dimension space where they are more easily separable. Such a technique is expected to work better with a dataset such as the one used in this work: even in the original data space, one can observe that the classes become more separable as the dimensionality is increased from 2-D to 4-D. As shown earlier in Figures 5.11 and 5.12, much of this increased separation between the two classes is due to the low SpO_2 cluster which removes the overlap seen in Figure 4.10 for the 2-D dataset without the SpO_2 information. The decrease in overlap between classes with the increase in data dimensionality, going from the 2-D dataset to the 4-D dataset, also explains the corresponding decrease in the number of FP and FN occurrences and the consequent increase in the sensitivity of the classifiers.

- The hybrid method performs at least as well as, or better than the standard and one-class methods. This is not an unexpected result - the hybrid method is a combination of the one-class method and a standard classifier. It correctly classifies the low-density abnormal clusters representing bradycardia and apnoea. As a result, there are fewer FN_{brady} occurrences using hybrid method, whereas both the one-class and the standard classifiers fail to identify these abnormal clusters in the 4-D dataset.

- It is recognised, however, that the hybrid method generally results in higher percentages of FP than the standard classifiers do. There is a trade-off in the design of the hybrid method, with optimality defined as the point at which the FP and FN percentages are both minimised. This is based on the assumption that reductions in the percentage of FNs and FPs are equally important in the clinical setting. It may be noted that, this assumption notwithstanding, the hybrid method does have a similar sensitivity, if not higher, than that of the other two methods.

- Finally, it can be seen that the one-class method of Chapter 4 results in more FP and FN occurrences than either the standard or hybrid SVM, which confirms the

intuition that the two-class approach results in better classification than does the one-class approach.

5.9 Conclusions

The performance of different types of classifiers was investigated in this chapter. It was seen that the logistic regression model performed the worst, since it is unable to classify non-linearly separable data, and was not considered further. The MLP and SVM models are both able to generate non-linear classification boundaries and so are more suited to the vital-sign classification problem. The SVM model has the advantage of being able to project the data into kernel space where, with the help of "soft" margins and slack variables, good separation was achieved.

The results of Section 5.8 show that classifier performance is better when the classifier is trained with the hybrid method rather than with standard methods. The percentage of false negatives is reduced, as well as the instances of misclassification of under-represented data, with the hybrid method, when compared with standard training procedures. This is seen for all datasets regardless of dimensionality. As is discussed in the next chapter, the separation in classes achieved with the hybrid method at T_{opt} can be exploited for modelling individual types of vital-sign abnormalities.

Chapter 6

Constructing Models of Abnormality

6.1 Introduction

Novelty detection approaches to modelling patient physiology have so far relied on models constructed using only "normal" data. In previous chapters of this thesis, a "hybrid" novelty detection approach has been proposed, which defines an "optimal" boundary (based on a one-class model of "normal" data) between the "normal" data and any "abnormal" examples that may occur. Due to the shortage of vital-sign data representing abnormal physiology, previous methods have not explicitly modelled regions of abnormal physiology. In this chapter, a multi-class approach in which both "normal" and "abnormal" classes are modelled explicitly, is described.

In Chapter 5 the multi-dimensional clusters representing abnormal physiology in dataset $\mathcal{D}_1$ were identified. The method proposed in this chapter will seek to model these abnormal clusters explicitly, and will generate a multi-class output. A multi-class output will indicate (i) if the test data are "normal" or "abnormal" (as with novelty detection), and (ii) classify the type of abnormality. This last feature could be a useful addition to existing patient monitoring systems, as it can aid clinical decision-making by identifying the type of physiological deterioration taking place.

In the following sections, first the construction of a model of abnormality is described. The model is fitted to "abnormal" clusters identified in the the 2-D dataset and the 4-D dataset. Finally results for the multi-class approach are presented. As in previous chapters, dataset $\mathcal{D}_1^{train}$ is used for the training of models and $\mathcal{D}_1^{test}$ is used for testing the

Vital Sign	High	Low
HR	15	4
BR	20	6
SpO_2	–	39
SBP	6	15
DBP	7	–

Table 6.1: Break-down of the 112 adverse events in $\mathcal{D}_1$, along with the corresponding vital sign in which deterioration was considered by the panel of clinical experts to have occurred.

method.

6.2 Description of "Abnormal" Clusters

Figure 6.1 shows the location of "abnormal" data from class C_2 in dataset $\mathcal{D}_1^{test}$, representing adverse events (in pink). The "normal" data corresponding to class C_1 is shown in blue. Figure 6.1 (a) shows the HR and BR data for C_1 and C_2, 6.1 (b) the BR and SpO_2, and 6.1 (c) the BR and SDA for the two classes.

In order to construct explicit models of abnormality type, labels of abnormality type associated with the data from class C_2 are required. From Figure 6.1 individual clusters representing abnormality types may be discerned within class C_2. For example, in Figure 6.1 (b), a separation can be seen between the "normal" C_1 cluster (in blue) and a *potential* "abnormal" cluster corresponding to low SpO_2, which is shown by the pink dots encircled by the hypothetical boundary marked by the circle in dark pink. Similarly, in Figure 6.1 (a), a potential high-BR cluster, a high-HR cluster, a low-BR cluster, and a low-HR can be considered to lie within the hypothetical boundaries shown by circles in purple, orange, light purple, and brown, respectively.

As described in Section 3.4, in labelling dataset $\mathcal{D}_1$, clinical experts used the MET calling criteria (see Table 3.2) to label which abnormality type an adverse event was associated with. Table 6.1 shows the number of adverse events labelled by clinical experts, and the "abnormality" type considered by the experts to be associated with the adverse events, in accordance with the MET calling criteria. To construct models for individual type of abnormality, one would therefore have to apply these labels to the training set, $\mathcal{D}_1^{train}$ (which contains classes C_1^{train} and C_2^{train}) and the test set, $\mathcal{D}_1^{test}$ (which contains

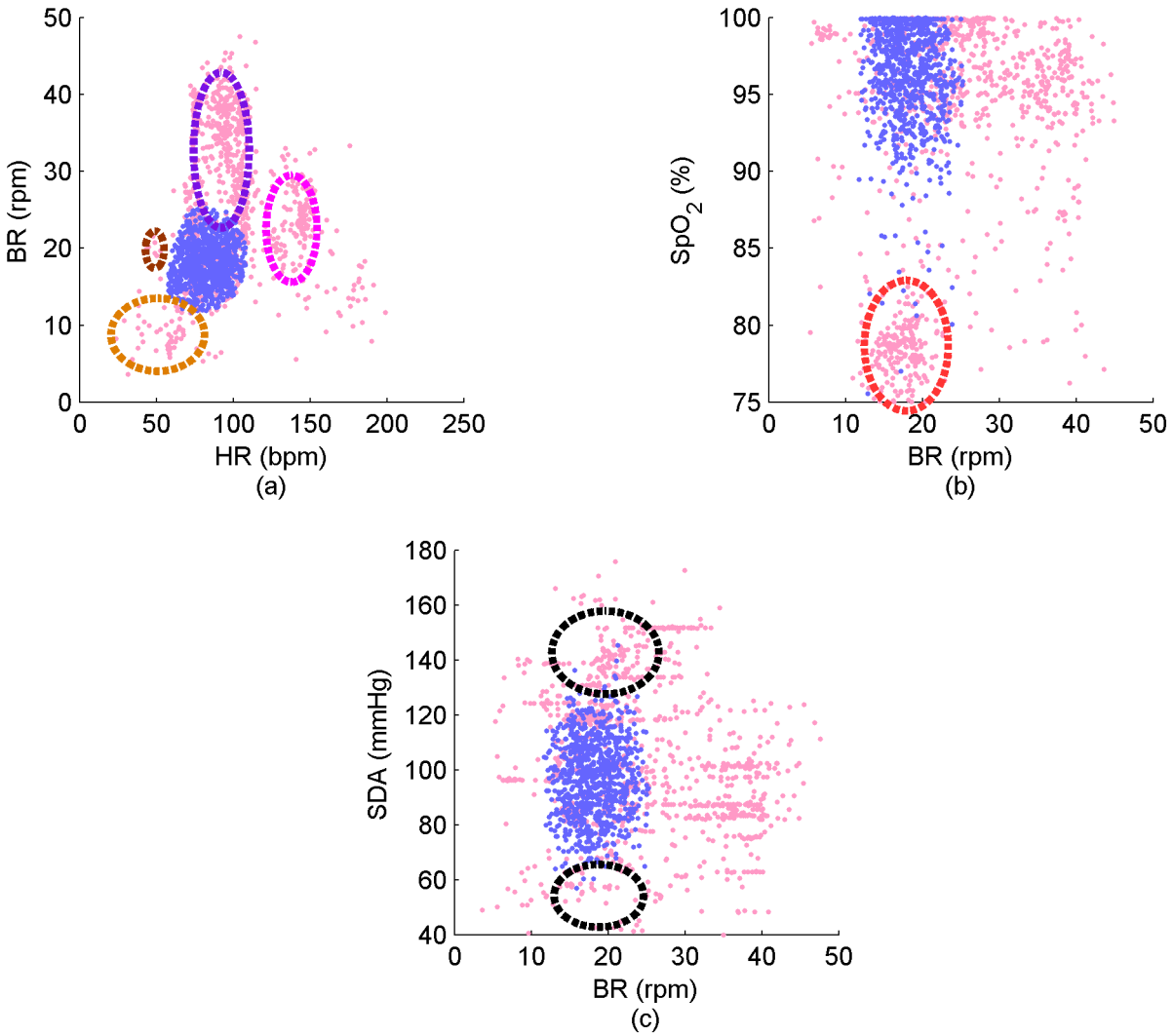


Figure 6.1: Two-dimensional scatter plots of dataset $\mathcal{D}_1^{test}$. Data from classes C_1^{test} and C_2^{test} are shown in light blue and pink, respectively. (a) BR v. HR, (b) SpO_2 v. BR, and (c) SDA v BR, are shown. Hypothetical boundaries are shown in (a) by the circles in purple, orange, light purple, and brown, for potential "abnormal" classes representing abnormality types high-BR, low-BR, high HR, and low HR, respectively, and in (b) in dark pink for for potential class representing low-SpO_2, and in (c) in black for for potential hypothetical classes representing high and low SDA clusters.

classes C_1^{test} and C_2^{test}). Figure 6.2 shows the individual abnormal clusters in class C_2^{test}, labelled using the MET calling criteria. The clusters labelled in the Figure 6.1 (a) as C_{2A}, C_{2B}, C_{2C}, C_{2D} represent the high-BR, low-BR, high-HR, and low-HR classes respectively. In 6.1 (b) the cluster labelled as C_{2E} represents the low-SpO_2 class respectively. To elaborate, any C_2^{test} data (i.e., data shown in pink) for which, following the MET calling criteria, the corresponding BR $>$ 36 rpm, is labelled as belonging to the high-BR class C_{2A}. Similarly, C_2^{test} data (i.e., data shown in pink) for which, following the MET calling criteria, the corresponding HR $>$ 36 rpm, is labelled as belonging to the high-HR class C_{2C}. Classes C_{2B}, and C_{2D} were labelled in the same manner, using the MET limits for low HR and low BR. For C_{2D}, an exception was made. According to the MET limits, the cluster encircled in brown in Figure 6.1 (a) would be missed out as an "abnormal" cluster, therefore in this case only, the MET limits are ignored and an artificial rule is applied such that data for which 15 rpm$\leq$BR$\leq$22 rpm, and for which 15 bpm$\leq$HR$\leq$55 bpm, would be included in the class C_{2D}. As expected from Table 6.1 it is seen that the C_{2B} and C_{2D} clusters are more sparse as shown in Figure 6.2 (a) than the C_{2A} and C_{2C} clusters since there are only 4 bradycardia (low HR) and 6 bradypnoea (low BR) events in the dataset.

From Figure 6.2 (c) a high-SDA cluster and a low-SDA cluster may be discerned. While it is possible to model these latter two potential clusters, it is not possible to map them into SBP or DBP clusters (as SDA is the arithmetic average of the two), and as such it would not be possible to derive the underlying SBP and DBP physiology that such clusters might represent. Therefore only classes C_{2A}, C_{2B}, C_{2C} , C_{2D} and C_{2E} are considered in this work. For these five MET-labelled abnormal clusters, Table 6.2 shows the mean and ± 1 standard deviation of each cluster in $\mathcal{D}_1^{test}$, where it is assumed that each cluster (including the one corresponding to low SpO_2) contains Gaussian-distributed data. The corresponding vital-sign distributions and the number of data in $\mathcal{D}_1^{train}$ are similar to those shown in this table for $\mathcal{D}_1^{test}$. According to Table 4.3 there are 1,073 data in each of C_1^{train} and C_1^{test}, and an equal number of data in C_{12}^{train} and C_2^{test}. Note from the last column of Table 6.2 that labelling abnormal clusters according to the MET criteria reduces the total number of data in C_2^{test} from 1,073 to 395 (which is distributed as: 212

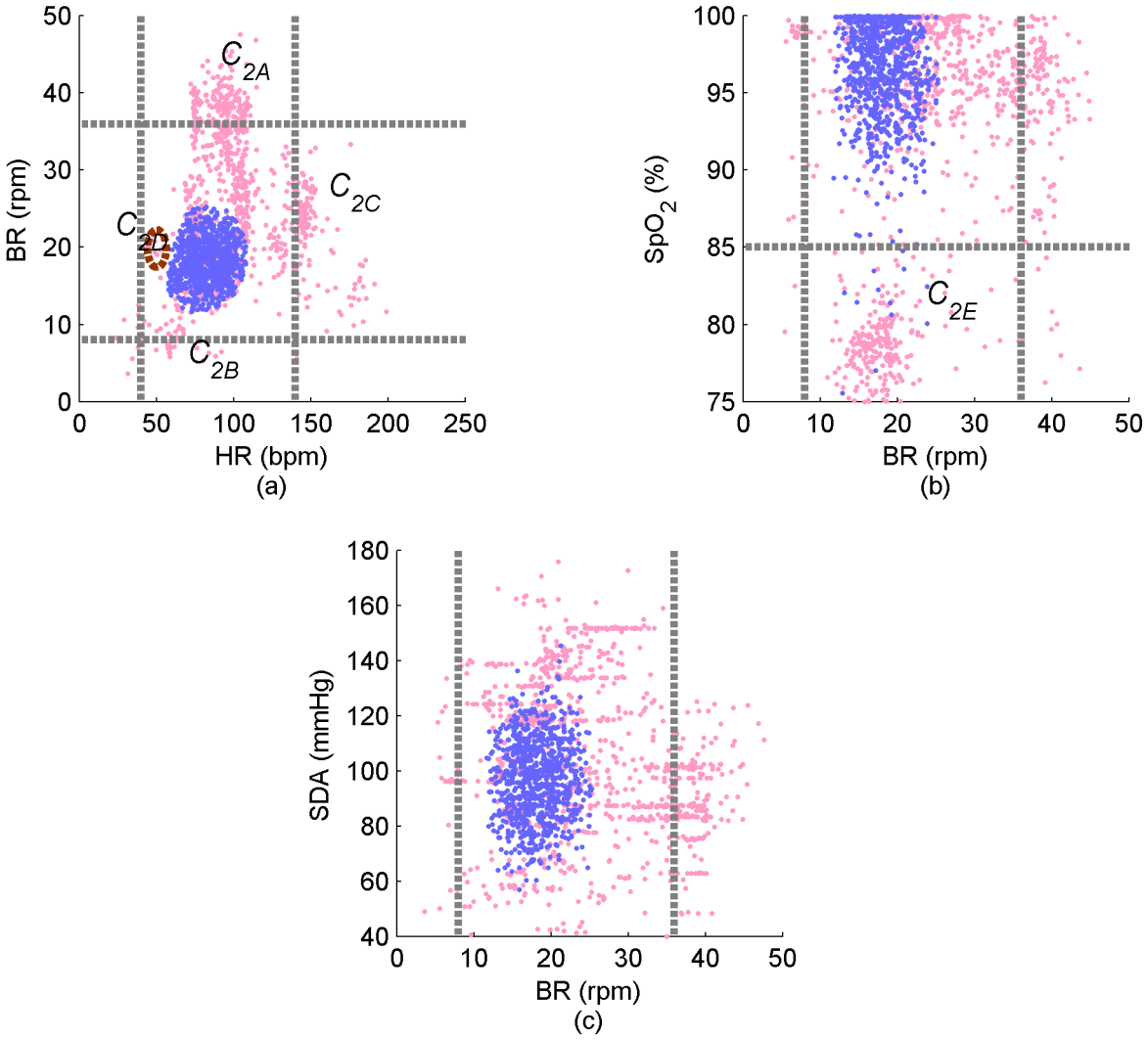


Figure 6.2: Two-dimensional scatter plots of dataset $\mathcal{D}_1^{test}$. Data from classes C_1^{test} and C_2^{test} are shown in light blue and pink, respectively. (a) BR v. HR, (b) SpO_2 v. BR, and (c) SDA v BR, are shown. Also shown by broken in lines in grey are the limits of normality as defined in the MET calling criteria (Table 3.2). "Abnormal" classes representing abnormality types are labelled according to the MET criteria: in (a) any pink points that lie within the boxes marked by the grey lines as C_{2A}, C_{2B} and C_{2C}, represent the classes labelled as high-BR, low-BR, and high-HR, respectively. In (b), any pink points that lie within the box marked as C_{2E} represent the class labelled as low-SpO_2. For C_{2D} in (a), the MET limits are ignored and an artificial rule is applied such that any pink points for which 15 rpm$\leq$BR$\leq$22 rpm, and for which 15 bpm$\leq$HR$\leq$55 bpm, would be included in the class C_{2D}, representing abnormality type low-HR, as shown by the circle in brown.

in C_{2A} + 55 in C_{2B} +116 in C_{2C} + 12 in C_{2D}). There are 391 data in C_{2E} which is projected in a third dimension, SpO_2, so that when considering the full 4-D dataset there are 395+ 391 data in C_2^{test} (and in C_2^{train}), some of which are overlapping in 4-D space.

Considering initially the events related to HR or BR, it is interesting to note that while each of the 45 such abnormal events was labelled by clinical experts as being associated with one vital sign, Figure 6.2 (a) shows that many (pink) points lie in regions outside the normal range for *both* vital signs. Some of the C_{2D} data, e.g., correspond to a low HR (as low as 55 bpm) as well as a low BR (less than 10 rpm). Similarly, C_{2C} includes data corresponding to a high BR (of up to 36 rpm) as well as a high HR (up to 160 bpm). In Figure 6.2 (b) a minority of C_2 data (in pink) have a low SpO_2 ($< 85\%$) well as a high BR (> 36 rpm). This suggests that some abnormal events are accompanied by a deterioration in more than one vital sign, and that some others may be related to abnormal values in a single vital sign (such as those belonging to the clusters representing low HR events, C_{2D}, and high BR events, C_{2A}). The method developed in this chapter aims to investigate the possibility of both types of events.

To construct the multi-class model, the normal class C_1 is also considered. In Chapter 5 the optimal threshold T_{opt} for the normal class C_1 was found to be at $T = 0.55$. Therefore only data in $\mathcal{D}_1^{test}$ and $\mathcal{D}_1^{train}$ for which $P(\kappa_{\mathbf{x}}) > T_{opt}$ were retained. As a result the number of C_1data in classes C_1^{test} and C_2^{train} is reduced from 1,073 to 854. As mentioned earlier in Section 6.3, the number of data in the five "abnormal" classes $C_{2A}...C_{2E}$, and in the normal class C_1, do not have to be balanced when using the one-class density-based approach for modelling these classes individually. The reduced training and test sets after applying the MET criteria to the "abnormal" data, and the optimal threshold T_{opt} to the "normal" data, are henceforth referred to as datasets $\tilde{\mathcal{D}}_1^{test}$ and $\tilde{\mathcal{D}}_1^{train}$. Finally, it is noted that this "pruning" has increased the separation between the six clusters in $\tilde{\mathcal{D}}_1^{test}$ and $\tilde{\mathcal{D}}_1^{train}$. However, as mentioned above, it is necessary to "artificially" label the abnormal data according to abnormality type in order to perform classification of abnormality type, hence the pruning is unavoidable.

Class	HR (bpm)	BR (rpm)	SpO_2(%)	SDA (mmHg)	No. data
C_1	84 ± 14	19 ± 4	97 ± 2	97±14	854
C_{2A}	94 ± 9	38 ± 4	97 ± 5	89 ± 17	212
C_{2B}	60 ± 9	7 ± 2	98 ± 3	88 ± 33	55
C_{2C}	146 ± 7	24 ± 4	96 ± 3	102 ± 17	116
C_{2D}	46 ± 4	20 ± 1	99 ± 2	65 ± 27	12
C_{2E}	92 ± 10	25 ± 10	74 ± 13	98 ± 26	391

Table 6.2: Vital-sign mean and standard deviation values for the clusters identified in $\tilde{\mathcal{D}}_1^{test}$. The right-most column shows the number of data in each cluster. Total no. data in C_2= 212+55+116+12+391= 786.

6.3 Constructing Models of Abnormality

In previous chapters, modelling of the "normal" and "abnormal" classes was carried out using both one-class methods and standard two-class classifiers. Classifiers were trained using dataset $\mathcal{D}_1^{train}$, and their performance was tested on the held-out dataset $\mathcal{D}_1^{test}$ (see Table 4.3 for a description of the two datasets). In this chapter, datasets $\tilde{\mathcal{D}}_1^{train}$ and $\tilde{\mathcal{D}}_1^{test}$ are used for training and testing the models, respectively. However, in this chapter the "abnormal" class has been divided into smaller classes $C_{2A}...C_{2E}$. In order to construct models of these classes using standard classifiers, the classifiers would have to be trained using balanced numbers of data from each class. The smallest class, class C_{2D}, contains 12 datapoints only (see Table 6.2), which is too small a number of data to train a classifier (such as the LR, MLP, and SVM used in the previous chapter) with. A one-class, density-based modelling approach is therefore a more suitable option, and is used to construct explicit models of the "abnormal" clusters considered in this chapter.

The multi-class approach is described in the following section, and is applied to the full 4-D dataset in Section 6.3.

6.3.1 A Multi-Class Output

The procedure for constructing a one-class density-based model of normality was explained in Chapter 3. The one-class approached previously used in Chapter 4 defined the probability of novelty $P(\kappa_{\mathbf{x}})$, where $\kappa_{\mathbf{x}} = p(\mathbf{x})$. Then using a threshold T on $P(\kappa_{\mathbf{x}})$, data which have $P(\kappa_{\mathbf{x}}) \geq T$ were considered to be "abnormal" with respect to the model of normality (see Equation 4.4). A model of abnormality can be formed in a similar way,

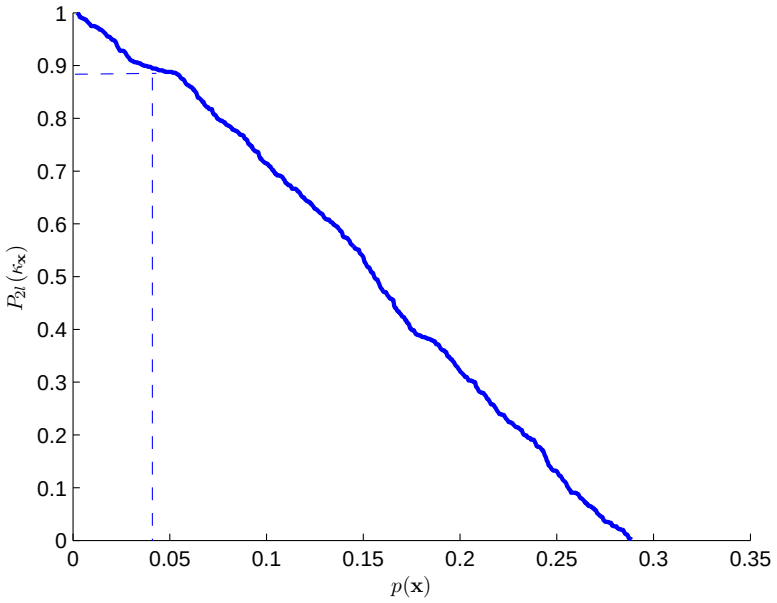


Figure 6.3: Probability $P_{2l}(\kappa_{\mathbf{x}})$ against pdf $p(\mathbf{x})$, for the model of abnormality constructed for class C_{2A}. Data closer to the mode of the model of abnormality for this class will take higher values of $p(\mathbf{x})$, and correspondingly lower values of $P_{2l}(\kappa_{\mathbf{x}})$, which the probability of being away from this model. $p(\mathbf{x}) = 0.04$ corresponding approximately to $P_{2l}(\kappa_{\mathbf{x}}) = 0.9$ is shown by the dashed lines.

estimating the probability density function (pdf), $p(\mathbf{x})$, of data from each "abnormal" class ($C_{2A}...C_{2E}$), and introducing the threshold $T_{2\ell}$ on $P_{2l}(\kappa_{\mathbf{x}})$, where $l = A,B,C,D,E$. In this case, the probability $P_{2l}(\kappa_{\mathbf{x}}) \geq T_{2l}$ is the probability of being away from the model of abnormality for class C_{2l}. The threshold optimisation procedure described in Chapter 5 is no longer required, as instead the threshold T for the model of normality, and the threshold $T_{2\ell}$ for each model of abnormality, is defined here to be that which include 90% or more data from the corresponding class, i.e. $T = 0.9$ for the "normal" class, and $T_{2\ell} = 0.9$ for each "abnormal" class. The choice of $T = T_{2\ell} = 0.9$ is discussed in detail in Section 6.5.

Figure 6.3 shows the relationship between the pdf $p(\mathbf{x})$ and the probability $P_{2l}(\kappa_{\mathbf{x}})$ for the model of abnormality constructed for class C_{2A}.

As in previous chapters, data for which $P(\kappa_{\mathbf{x}}) \geq T$ will be considered as being "abnormal" with respect to the model of normality, where $T = 0.9$. Further, data for which $P_{2l}(\kappa_{\mathbf{x}}) < T_{2l}$ will be considered as belonging to the model of abnormality for class C_{2l},

where $T_{2l} = 0.9$. Any data with $P_{2l}(\kappa_{\mathbf{x}}) > T_{2l}$ will not be considered as belonging to the corresponding "abnormal" class C_{2l}. Therefore it can be determined, for any test data: (1) whether it is "abnormal" with respect to the model of normality, and if so, (2) which model of abnormality it belongs to.

For any $\mathbf{x}$, two probabilities $P(\kappa_{\mathbf{x}})$, and $P_{2\ell}(\kappa_{\mathbf{x}})$ are relevant in defining outcomes of two types:

(i) type A outcome, where $P(\kappa_{\mathbf{x}}) \geq T$, and $P_{2\ell}(\kappa_{\mathbf{x}}) < T_{2l}$, i.e. when $\mathbf{x}$ is outside the model of normality and inside model of abnormality corresponding to class C_{2l}.

(ii) type B outcome, where $P(\kappa_{\mathbf{x}}) \geq T$, and $P_{2\ell}(\kappa_{\mathbf{x}}) \geq T_{2l}$. Type B refers to a case where $\mathbf{x}$ are are outside the model of normality but not within any one "abnormal" class, and are therefore considered as being "abnormal, but abnormality type not determined".

(iii) third type C outcome would occur if $P(\kappa_{\mathbf{x}}) < T$ and $P_{2\ell}(\kappa_{\mathbf{x}}) < T_{2l}$, i.e. when $\mathbf{x}$ is inside the model of normality and inside the model of abnormality corresponding to class C_{2l}. This case would occur when there is an overlap between class and C_1 and C_{2l}, for instance, between C_1 and C_{2D}.

$P(\mathbf{x} \notin C_1) = P(\kappa_{\mathbf{x}})$ and $P(\mathbf{x} \in C_{2l}) = 1 - P_{2\ell}(\kappa_{\mathbf{x}})$ are now refered to as the multi-class outputs. Setting $T = 0.9$, and $T_{2l} = 0.9$, if $P(\mathbf{x} \notin C_1) = P(\kappa_{\mathbf{x}}) \geq T$, then $\mathbf{x}$ is outside the model of normality. If $P_{2\ell}(\kappa_{\mathbf{x}}) \geq T_{2l}$, and consequently $P(\mathbf{x} \in C_{2l}) \geq 0.1$, then $\mathbf{x}$ is inside the model of abnormality for class C_{2l}.

A TP case was deemed to have occurred when a datapoint belonging to an "abnormal" class (according to the labelling described in Section 6.2) was classified correctly, i.e. in the case of a type A outcome. Similarly, a datapoint correctly classified as belonging to C_1 was deemed to be a TN case. A classification was deemed to be a FP case when a datapoint from C_1 was classified as belonging to an "abnormal" class C_{2l}, and an FN case when a datapoint from an abnormal class C_{2l} was classified as belonging to the normal class C_1.

6.4 Results

Models were trained using the 4-D dataset $\tilde{\mathcal{D}}_1^{train}$ and the classification performance of the multi-class approach was tested on the held-out test set $\mathcal{D}_1^{test}$. Table 6.3 shows the

results. The percentages in the TP column indicate, for an "abnormal" class C_{2l}, the proportion of "abnormal" data correctly estimated as being "abnormal with abnormality type l". For C_{2l}, the FN column shows the number of data which belong to that class but were misclassified as belonging to the "normal" class C_1. The TN column refers to the number (and percentage, in brackets) of correctly identified data belonging to the "normal" class C_1. The right-most column refers to "indeterminate" cases. The indeterminate cases for classes C_{2A}, C_{2B}, C_{2C} , and C_{2E} refer to Type B cases, i.e. those cases where $\mathbf{x}$ belongs to one of these classes, and was correctly identified to be "abnormal" with respect to "normal" class C_1, i.e., $P(\mathbf{x} \notin C_1) \geq T$, but where $\mathbf{x}$ is not near enough their respective model of abnormality, i.e.$P_{2l}(\mathbf{x} \in C_{2l}) \geq T_{2l}$. For class C_{2D}, the "indeterminate" column refers to Type C , i.e. where $P(\kappa_{\mathbf{x}}) < T$ and $P_{2\ell}(\kappa_{\mathbf{x}}) < T_{2l}$, as may be expected to occur due to the small size of C_{2D} and its overlap with C_1. The "indeterminate" cases for class C_1 refer to the Type B cases belonging to class C_1.

While the TP and and FN percentages indicate that the method is accurate at determining Type A cases corresponding to each "abnormal" class, the right-most column of Table 6.3 indicates that there are several indeterminate cases belonging to each class where the method does not determine the abnormality type. The Type B indeterminate cases occur when data lie outside the boundary of the "normal" class C_1 as well as outside the boundary of the relevant "abnormal" class C_{2l} ; the Type C indeterminate cases occur when data lie within the boundary of the "normal" class C_1 as well as well as within the boundary of the relevant "abnormal" class C_{2l} . These indeterminate cases are an expected limitation of the proposed approach, where the boundaries around each class have been defined according to T (and T_{2l}). By setting $T = T_{2l} = 0.9$, one is already causing 10% of data to be excluded from a given class. Increasing the threshold would lead to a reduction in the number of indeterminate cases, but also to an increase in the number of FP and FN cases. For example, increasing $T = T_{2l} = 0.98$, decreases the number of indeterminate cases in C_2, overall, from 135 to 92 cases, but increases the number of total number of FN cases from 8 to 18, and the number of FP cases from 60 to 93. Thus the two thresholds governs the trade-off between indeterminate and FP/FN cases.

However, even though the abnormality type is not assigned for the indeterminate cases,

	No. data	TP	FN	FP	TN	Indeterminate
C_{2A}	212	178 (84%)	0	–	–	34
C_{2B}	55	44 (80%)	6	–	–	5
C_{2C}	116	88 (76%)	0	–	–	28
C_{2D}	12	8 (67%)	0	–	–	4
C_{2E}	391	325 (83%)	2	–	–	64
C_1	854	–	–	60 (7 %)	728 (85 %)	66

Table 6.3: Classification results for the 4-D dataset $\tilde{\mathcal{D}}_1^{test}$ using the multi-class method. TP and FN numbers (and percentages in brackets) are presented for the "abnormal" classes C_{2A}, C_{2B}, C_{2C}, C_{2D}, and C_{2E}, and FP and TN numbers are presented for the "normal" class C_1. The rightmost column shows the number of Type B indeterminate cases for classes C_{2A}, C_{2B}, C_{2C}, C_{2E} and C_1, and the number of Type C cases for C_{2D}. The total number of indeterminate cases in the overall class C_2 is $34 + 5 + 28 + 4 + 64 = 135$. The total number of FN cases in the overall class C_2 is $6 + 2 = 8$.

they are still classified as being "abnormal" with respect to the model of normality. Thus for the 64 datapoints from C_2 for which the approach does not provide additional information about the type of abnormality, the method still correctly classifies them as being "abnormal" with respect to the model of normality. These results indicate that by considering explicit models of abnormality in addition to a model of normality, where possible (i.e. subject to the thresholds T and T_{2l}), novelty detection along with identification of abnormality type can be performed.

6.5 Discussion

Results presented in this chapter (Table 6.3) show that incorporation of explicit models of abnormality through the proposed multi-class approach enables classification of the type of abnormality, with reasonable accuracy. It may be noted that in the work presented in this chapter, the methodology of generating multi-class outputs was described; the methodology was demonstrated by setting the thresholds on the proposed model of normality and models of abnormality at an example value ($T = T_{2l} = 0.9$). In chapters 4 and 5, detailed approaches for selecting optimal values of thresholds for classification into "normal" and "abnormal" classes have already been described. The focus of this chapter was on developing a technique to estimate abnormality type, and therefore threshold optimisation was not repeated. Nevertheless, the effect of varying thresholds on classification performance was reported in Section 6.4.

Based on these results, we applied the multi-class method to the individual time-series of those patients in dataset $\tilde{\mathcal{D}}_1$ for whom the "abnormal" events were labelled by clinical experts. Only events originally labelled by the experts as associated with HR, BR, or SpO_2-related abnormalities were considered. The multi-class method was compared with the clinical labels to assess if the former correctly identified the abnormality type for the duration marked by clinical labels. As done throughout this chapter, a threshold of $T = 0.9$ was set on $P(\mathbf{x} \notin C_1)$, and threshold of $T_{2l} = 0.9$ on $P(\mathbf{x} \in C_{2l}). = 1 - P_{2\ell}(\kappa_{\mathbf{x}})$ as the multi-class outputs and if the multi-class output exceeded this threshold an alert was generated. The alerts generated by the multi-class method were seen to (i) either alert consistently throughout the labelled period, (ii) alert intermittently, (iii) or not alert at all. Figure 6.4 shows an example of the alerts generated by the multi-class method during the period labelled as a "low BR" event by clinical experts. The BR remains low for the entire episode marked by the clinical labels, whereas the other vital signs remain stable. The output for class C_{2B} is alerting intermittently throughout the episode, confirming that the adverse event is accompanied by bradypnoea. Another example is shown in Figure 6.5 which was labelled by clinical experts as a low SpO_2 event. Here the drop in the SpO_2 is accompanied by continuous alerting by the multi-class output for class C_{2E}, confirming that the labelled episode is accompanied by hypoxia. Finally, a third example is shown in Figure 6.6 to illustrate that in cases where there is an abnormal change in more than one vital sign, the multi-class output generates corresponding alerts. Recalling the discussion in Section 6.2, it was recognised that there are adverse events in dataset $\mathcal{D}_1$ in which the abnormality occurs in multiple vital signs. In the example shown, the event was labelled by clinical experts as a high-BR event, however from Figure 6.6 (b) it is also noted that the HR intermittently rises and falls throughout the labelled episode, indicating that apart from the high BR (39-41 rpm), Figure 6.6 (c), the patient's HR is also experiencing abnormal changes. This is reflected in the multi-class output which generates alerts to indicate both high HR (Figure 6.6 (f) and high BR (Figure 6.6 (h).

Clinical labels and multi-class output for all 84 out of 112 adverse events related to HR, BR or SpO_2 may be compared. However, comparison of the multi-class output and the reference clinical labels based on analysis of the times-series is not straightforward.

There are several limitations to this process. In many cases, the multi-class method generates alerts intermittently throughout the period labelled by the clinical experts. Here it was difficult to determine whether intermittent alerts should be considered as true positives. One solution, as proposed by [67], is to set a window length, and if an alert is generated at any time during that window, to consider it as a true positive. However this makes the system overly sensitive, since it assumes that an alert corresponding to a single sample in a given time window is sufficient to classify the episode as a true positive. Therefore a different approach was considered: if an alert was generated for 90% or more of the duration of the episode, then the multi-class output was deemed to be the same as the reference clinical label. It is noted here that the 90% criterion has been selected arbitrarily. Furthermore there were several cases of mis-assigned abnormality type in the clinical labels. To elaborate, an episode would be labelled by the clinical experts as, e.g., a "low SpO_2" event, but on visual inspection of the time-series, it could be seen that it was not a "low SpO_2" but possibly a "high BR" event or no event at all. This was especially true in the case of (60 % of the 39) events labelled as SpO_2 events.

Given these constraints, an initial comparison of multi-class outputs and clinical labels was performed, using the 90% alert criterion mentioned above. As discussed, many of the SpO_2-related events were mis-assigned, and therefore for this assessment only HR- and BR-related events were considered. Table 6.4 summarises the results for 42 events related to HR or BR. The remaining three events (including one low BR, and one low HR, and one high HR cases) were accompanied by very poor quality of either HR or BR data (i.e. where less than 10% data were available for the duration labelled by the clinical experts) and so were excluded. The results corroborate those in Table 6.3 and show that the multi-class method is accurate at identifying abnormality type. Also of note is the fact that the method correctly identifies two out of the three low HR, and the three out of the five low BR events correctly. This is an advantage of the proposed multi-class method over the one-class novelty detection method which, as discussed in Chapter 4, fails to correctly classify data belonging to the low HR and low BR clusters.

This analysis is only a first-step approach towards the assessment of the proposed method based on the time-series. The choice of the 90% alerting criterion was arbitrary,

Vital Sign	High	Low	Excluded
HR	12 out of 14	2 out of 3	2
BR	20 out of 20	3 out of 5	1

Table 6.4: Comparison of multi-class estimates with clinical labels. The number of events for which agreement was found between multi-class estimates and the reference clinical labels, according to the 90% alert criteria. The right-most column shows the number of events excluded from the comparison due to poor data quality.

and results (for the 42 events) were presented for a subset of the original dataset (which contains 112 labelled events in all). It is therefore not intended as an exhaustive or conclusive analysis. A full analysis and validation of the method would best be performed on an independent and fully labelled dataset. However this work is beyond the scope of this thesis.

6.6 Conclusions

The focus of Chapters 3 to 6 has been on modelling techniques for classification of "normal" and "abnormal" vital-sign data acquired from patients in an SDU, which covers the first of the two major themes of this thesis. The key findings of these chapters are summarised below.

6.6.1 Classifier Optimisation

Classifier optimisation began by constructing a probability density-based model of normality, inspired by earlier work on novelty detection, as described in Chapter 3. In Chapter 4, a probability-based threshold for classification was introduced. Based on this method, criteria for efficiently selecting data for training a classifier were also proposed. In Chapter 5, the performance of linear and non-linear classification models was investigated, and it was found that the support vector machine classifier handles non-linearly separable data better than the other classifiers, and so gave the best classification performance when applied to the dataset. Having optimised the choice of classification model, two-class classification was performed to refine our choice of probability threshold.

The performance of this "hybrid" classification method was assessed by comparing the numbers of true and false alerts generated with by standard two-class classifiers. Results

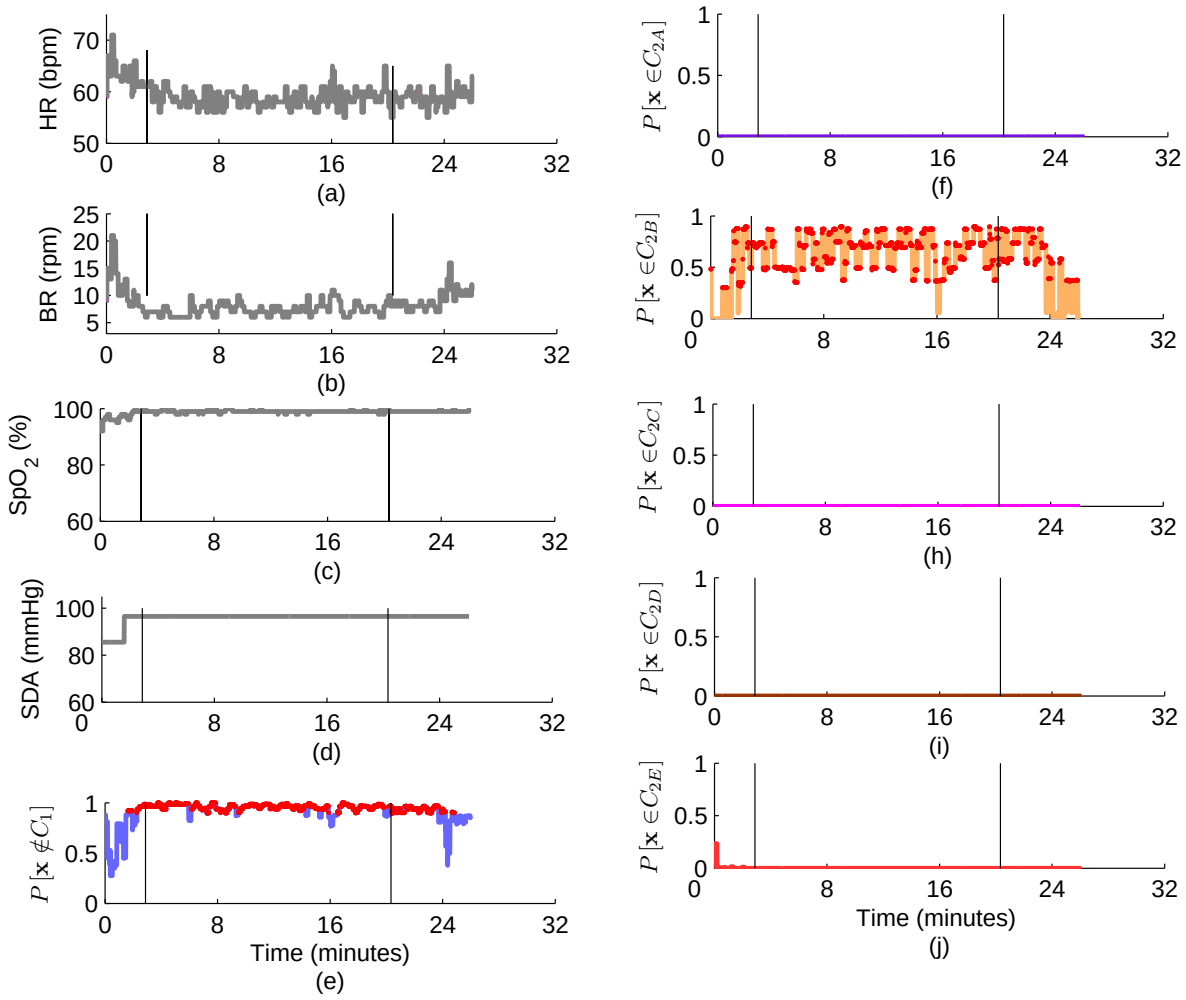


Figure 6.4: A clinically labelled low-BR event from dataset $\mathcal{D}_1$ belonging to Patient 159. (a) - (d) show the HR, BR, SpO_2 and SDA time-series during the event period labelled by clinical experts (marked by vertical black lines). (e) shows the time-series plot for $P[\mathbf{x} \notin C_1]$, and (f) - (j) show time-series for $P[\mathbf{x} \in C_{2l}]$ where $l = A,\ B,\ C,\ D$, and E, respectively. In (e), any sample for which $P[\mathbf{x} \notin C] > 0.9$ is marked in red. In (f) - (j) , any sample for which $P[\mathbf{x} \in C_{2l}] > 0.1$ is marked in red. During the labelled event, the alerts in red indicate that in (e) $P[\mathbf{x} \notin C_1]$ is high enough for $\mathbf{x}$ to be outside the model normality, and in (g) $P[\mathbf{x} \in C_{2B}]$ is high enough for $\mathbf{x}$ to be inside the model of abnormality for C_{2B}. In (f) and (h) - (j), $P[\mathbf{x} \in C_{2A}] = P[\mathbf{x} \in C_{2C}] = P[\mathbf{x} \in C_{2D}] = P[\mathbf{x} \in C_{2E}] = 0$, indicating that the event corresponds to class C_{2B}which represents abnormality of type low-BR.

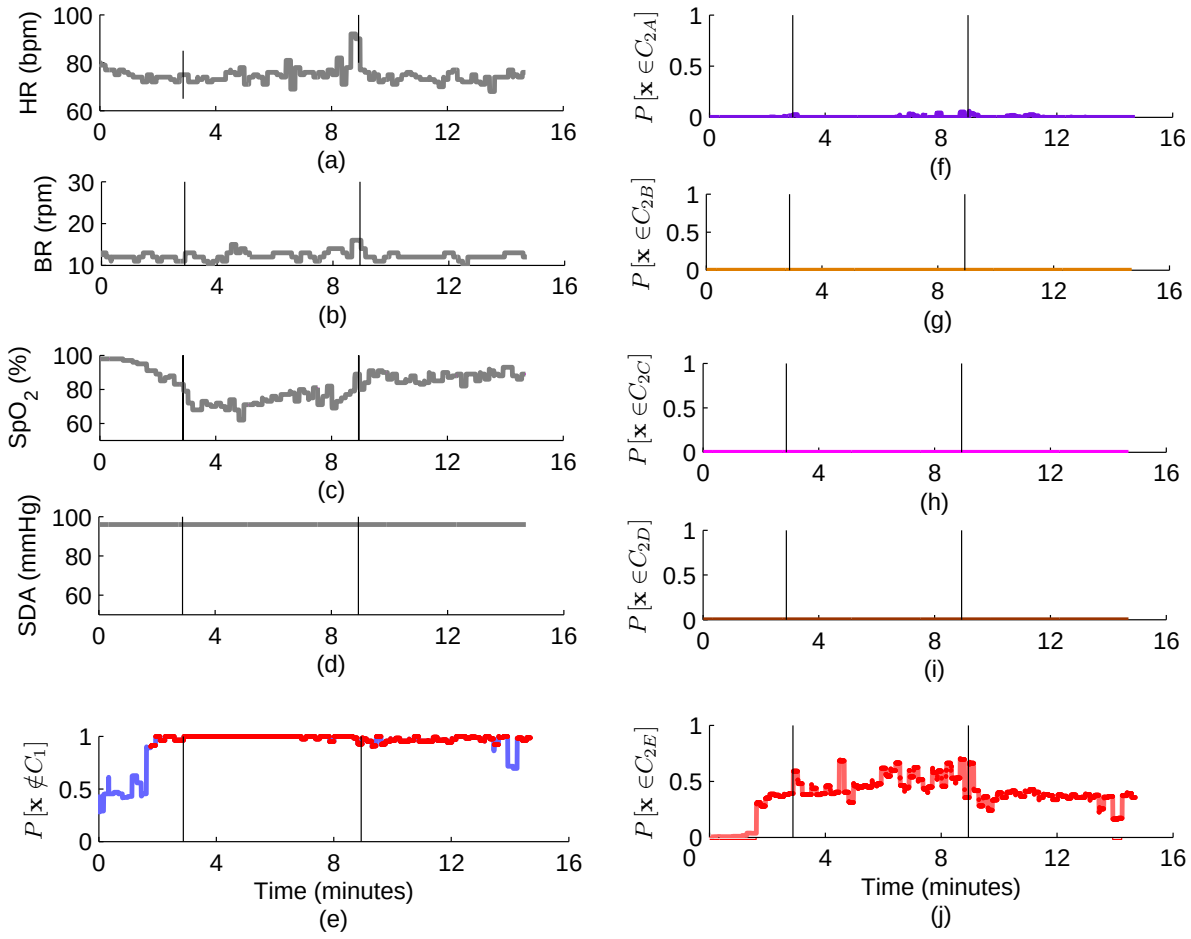


Figure 6.5: A clinically labelled low-SpO_2 event from dataset $\mathcal{D}_1$ belonging to Patient 206. (a) - (d) show the HR, BR, SpO_2 and SDA time-series during the event period labelled by clinical experts (marked by vertical black lines). (e) shows the time-series plot for $P[\mathbf{x} \notin C_1]$, and (f) - (j) show time-series for $P[\mathbf{x} \in C_{2l}]$ where $l = A$, B, C, D, and E, respectively. In (e) , any sample for which $P[\mathbf{x} \notin C] > 0.9$ is marked in red. In (f) - (j) , any sample for which $P[\mathbf{x} \in C_{2l}] > 0.1$ is marked in red. During the labelled event, the alerts in red indicate that in (e) $P[\mathbf{x} \notin C_1]$ is high enough for $\mathbf{x}$ to be outside the model normality, and in (j) $P[\mathbf{x} \in C_{2E}]$ is high enough for $\mathbf{x}$ to be inside the model of abnormality for C_{2E}. In (f) - (i) $P[\mathbf{x} \in C_{2A}] = P[\mathbf{x} \in C_{2C}] = P[\mathbf{x} \in C_{2D}] = P[\mathbf{x} \in C_{2B}] = 0$, indicating that the event corresponds to class C_{2E} which represents abnormality of type low-SpO_2.

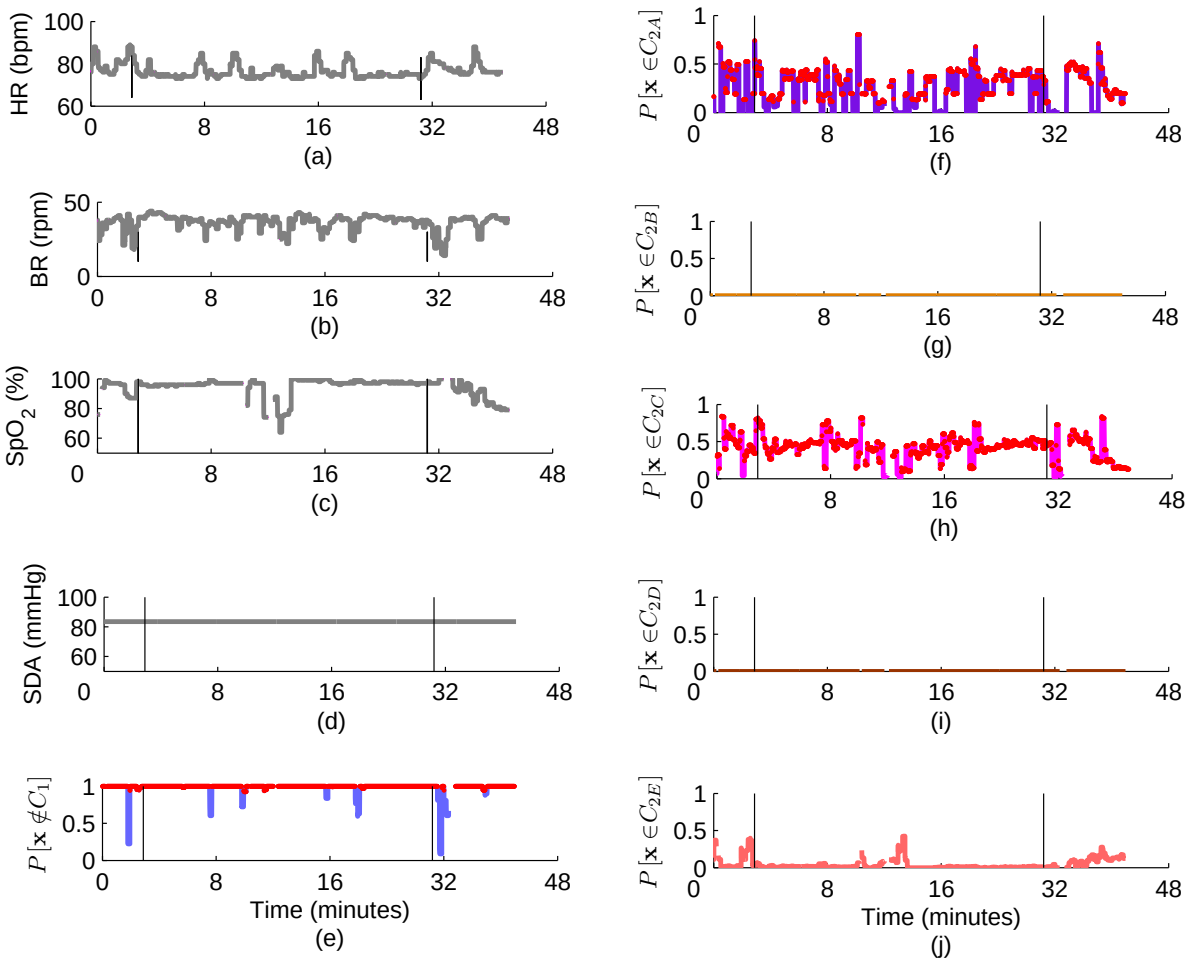


Figure 6.6: A clinically labelled high-BR event from dataset $\mathcal{D}_1$ belonging to Patient 225. (a) - (d) show the HR, BR, SpO_2 and SDA time-series during the event period labelled by clinical experts (marked by vertical black lines). (e) shows the time-series plot for $P[\mathbf{x} \notin C_1]$, and (f) - (j) show time-series for $P[\mathbf{x} \in C_{2l}]$ where $l = A, B, C, D,$ and E, respectively. In (e) , any sample for which $P[\mathbf{x} \notin C_1] > 0.9$ is marked in red. In (f) - (j) , any sample for which $P[\mathbf{x} \in C_{2l}] > 0.1$ is marked in red. During the labelled event, the alerts in red indicate that in (e) $P[\mathbf{x} \notin C_1]$ is high enough for $\mathbf{x}$ to be outside the model normality, in (f) $P[\mathbf{x} \in C_{2A}]$ is high enough for $\mathbf{x}$ to be inside the model of abnormality for C_{2A}, and in (h) $P[\mathbf{x} \in C_{2C}]$ is high enough for $\mathbf{x}$ to be inside the model of abnormality for C_{2C}. In (g) and (i), $P[\mathbf{x} \in C_{2B}] = P[\mathbf{x} \in C_{2D}] = 0$, while in (j) $P[\mathbf{x} \in C_{2D}] \leq 0.1$, indicating that the event corresponds to classes C_{2A} and C_{2B} which represent abnormalities of type high-HR and high-BR.

showed that the proposed method resulted in fewer FP and FN cases than the standard classifiers.

6.6.2 Incorporating Models of Abnormality

In this chapter, a method was presented to incorporate models of abnormality in a proposed multi-class classification method. It was shown using 2-D and 4-D datasets how regions of abnormal physiology may be modelled individually to generate models of abnormality. A multi-class classification method was proposed, which allows one to estimate the probability of test data belonging to the model of normality as well the probability of its belonging to each of the models of abnormality. This allows the proposed model not only to alert when vital signs deteriorate from normality, but also to indicate the type of abnormality associated with the deterioration. In addition to assessing the performance of the method as a classification system, a time-series based analysis was also carried out as a secondary check. A full time-series based assessment of the method on an independently labelled dataset will be considered in the future work chapter.

Chapter 7

Deterioration in Patients on Upper GI Ward

The work presented in previous chapters has focused on the detection of periods of single- or multi-channel vital-sign abnormality. The models of normality and abnormality were constructed from dataset $\mathcal{D}_1$ which comprises continuous monitoring data from unwell patients recovering from a stay in the ICU. Can the generic model so far developed be calibrated for a dataset representing a sub-population of patients who are recovering from cancer surgery and who are monitored intermittently on a post-operative ward? In this chapter an answer this question is attempted, by considering the distribution of vital signs for such a "condition-specific" dataset, and assessing whether the distribution of "normal" and "abnormal" data are similar for the generic and this dataset. The dataset has been obtained from patients recovering from upper gastrointestinal (GI) surgery on the Upper GI ward of the Oxford Cancer Hospital.

In the following sections the Cancer Hospital project is introduced, and the dataset obtained as part of the study is described. The remaining sections compare the generic dataset with the condition-specific dataset, and motivate the need for models calibrated for the latter.

7.1 The Cancer Hospital Project

The Cancer Hospital Project was set up to carry out the Computer ALerting and Monitoring System-2 (CALMS-2, Oxford, UK) trial[1]. The project is a collaboration between the Institute of Biomedical Engineering, and the Oxford Cancer Hospital in the Oxford University Hospitals NHS Trust, Oxford, UK. A key aim of the study is to investigate vital-sign trends during recovery following surgery, and to calibrate current monitoring systems for specific patient populations, with a view to (i) detecting deterioration, and (ii) reducing patients' length of stay (LOS) in the hospital.

The clinical trial has taken place in two phases - a "pre-intervention" phase in which track-and-trigger (as described in Section 1.3.1) data were recorded and an early warning score, EWS (see section 1.3.3) was used for clinical decision-making on the upper GI ward, and a second "post-intervention" phase in which a patient status index was displayed to help the nurses in their decision-making. Data for the analysis reported in this thesis was collected during the (pre-intervention) phase I, which took place between March 2010 and January 2011. The second phase began in the summer of 2011 and the collection of patient data for this phase was still on-going at the time of completion of this thesis.

7.1.1 The Clinical Trial

Patients undergoing upper GI surgery show a high incidence (up to 20 %) of post-operative complications which can result in emergency ICU admissions. This makes the cohort appropriate for the study of deterioration trends, since these complications commonly occur a few days after surgery.

Patients are admitted to the upper GI ward (Level 1) for the duration of their recovery phase unless a higher level of care is required, in which case they are admitted to the high dependency unit (HDU) (Level 2) [2]. Upper GI ward patients have their vital signs recorded periodically by nursing staff who enter the data manually onto a track-and-trigger chart. An early warning score (EWS) is calculated from the data, and an alert is generated if the individual score for a vital sign or the aggregate score for all vital signs

[1] Approved by the Oxford Research Ethics Committee, OxREC No. 08/H0607/79

[2] Levels of care are as described in Table 1.1 of section 1.2.

is above some pre-determined threshold.

7.1.2 Data Collection

Vital-sign observations for each patient were recorded by the duty nurse, and entered into a track-and-trigger paper chart. An observation set included the patient's heart rate, breathing rate, oxygen saturation, tympanic temperature, and systolic blood pressure. An example paper chart is shown in Figure 7.1. The EWS shown in Table 7.1 was the scoring system in use in the upper GI ward during phase I. The EWS score was also recorded on the paper chart. Observations were typically made every four hours, and the frequency of recording increased if the patient required greater medical attention.

Observation sets recorded on track-and-trigger paper charts were then transcribed into two separate electronic databases by two separate research nurses in order to minimise human error during transcription. The first transcription was carried out on a daily basis by a research nurse who would enter data from the previous 24 hours (or 72 hours, following a week-end) into a Personal Digital Assistance (PDA). The second transcription was carried out when a patient was discharged from the ward. It was performed by a second research nurse who entered the data into a second database (called the "excel database") without seeing the first "PDA database". The two databases were "merged" by a computer program into a "reconciled" database, and any transcriptional discrepancies in the two databases were removed by the program. If a discrepancy was too large (i.e., above a certain threshold), a third research nurse corrected the entry manually by consulting the patient notes. The reconciled database contains, for a given patient, the vital-sign values as recorded by the duty nurse, and the individual and aggregate EWS as recorded by the duty nurse. It also contains the time of entry to the ward and discharge from the ward (or time of death for any patient who died on the ward), as well as the times of any pre-scheduled or emergency admission to ICU and eventual return to the ward.

7.1.2.1 Data Gaps

Prior to analysis, data completeness was investigated. Time-series of track-and-trigger recordings were plotted for each patient, and any gaps identified. In several cases, a

Vital Sign	Score						
	3	2	1	0	1	2	3
HR		≤ 40	41-50	51-100	101-110	111-129	≥ 130
BR	≤ 8			9-18	19-24	25-29	≥ 30
SpO_2	≤ 91			≥ 92			
SBP	≤ 90	91-99		101-179			≥ 180
TMP		≤ 35.0		35.1-37.9			≥ 38.0

Table 7.1: The modified early warning score system used by nursing staff to monitor patients in the upper GI ward. Single-channel scores were calculated for the heart rate (HR, measured in beats per minute), breathing rate (BR, measured in breaths per minute), oxygen saturation (SpO_2, %), systolic blood pressure (SBP, mmHg), and temperature (TMP,^{o}C). A score of 3 in a single vital sign, or an aggregate score of 4 will trigger an alarm. EWS was the standard scoring system in practice at the upper GI ward during phase I of the CALMS-2 trial.

gap was found between the last track-and-trigger recording and the actual time of the end of stay in the ward. This was considered as an example of missing track-and-trigger data. Where possible, gaps were removed by updating the reconciled database manually according to updates provided by research nurses who consulted patient notes to verify information contained in the database. In some cases, however, it was found (during the work undertaken as part of this thesis) that no track-and-trigger observations had been recorded during the gap period. After consultation with the research nurses, these were identified as cases when the patient was removed from the ward for an emergency admission to the ICU. In such a situation, the gap between the last track-and-trigger recording and removal from ward was explained by the nurse assisting the patient and preparing for an emergency ICU admission, during which time no observations were recorded. The latter case is considered as a data gap and discussed further in the following section.

7.2 The $\mathcal{CH}_2$ Dataset

We will refer to the dataset obtained from phase I of the CALMS-2 trial as the $\mathcal{CH}_2$ dataset. The dataset contains a total of 16,672 track-and-trigger observation sets, where a complete observation set is comprised of the five vital signs (HR, BR, SpO_2, SBP, and Temperature). In addition to these five vital signs, the AVPU[3] score was also recorded.

[3]The Alert Voice Pain Unresponsive (AVPU) score is a measure of patient alertness. A score of 0 is recorded if the patient is alert or patient's eyes open spontaneously, 1 if the patient responds to voice, and 3 if the patient is unresponsive or responds to pain (or stimuli).

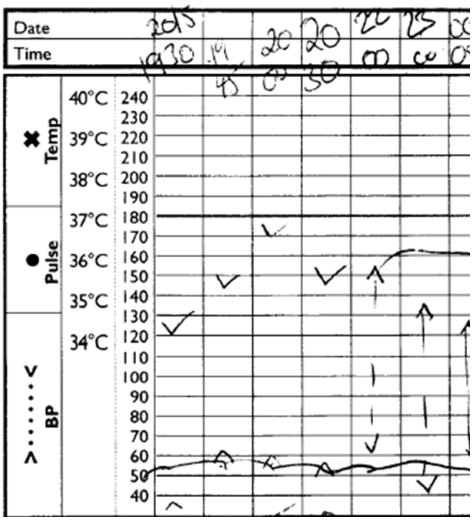


Figure 7.1: An anonymised extract from a track-and-trigger paper chart for a patient in the upper GI ward. Vital-sign observations are entered manually by the nurse.

Out of a total of 233 patients initially consented, 200 remained on the study. The rest were removed from the study due to revoked consent, transfer to a different ward, early discharge, or due to cancellation of surgery. The median age of all patients in the dataset was 64 years (with an inter-quartile range of 15 years), and the median length-of-stay (LOS) in the upper GI ward was 10 days (with an inter-quartile range of 7 days). All patients remaining in the study were recovering from upper gastrointestinal surgery, ranging from whipples (30%), liver resection (10%), hemi-hepatectomy (4%), total gasterectomy (8.5%), and others (28%) including pancreatectomy and splenectomy.

Figure 7.2 shows the categorisation of patients into groups according to their status during their stay on the upper GI ward. The 156 patients who had no emergency ICU admission, and were alive after 30 days of discharge from hospital with an LOS of at least four days but no more than 29 days were considered as being normal recovery cases and were included in a normal recovery (NR) group (indicated in the box in green). The reason for selecting this range of LOS was that the first 4 days after surgery were considered to be the recovery period in which a patient's vital signs are unstable [125]. A LOS of more than 30 days indicated that the patient had not recovered enough and was awaiting transferral from the upper GI ward.

Patients who had single or multiple emergency ICU admissions during their stay on the ward were included in an emergency ICU (EICU) group. This group includes a total of 19 patients, 16 of whom were discharged alive, and three who died in the ICU, as

indicated by the two pink boxes in Figure 7.2. Of the 16 patients in this group who were discharged alive, 14 had a single emergency ICU admission during their stay on the ward, whereas the remaining three had two. The median and interquartile (IQR) age and LOS for these two groups is shown in Table 7.2. It may be seen that the median LOS on the ward prior to emergency ICU admission for the EICU group (4.9 days) is almost half that of the median LOS in ward prior to discharge for the NR group (8.6 days) which suggests that an emergency ICU admission tends to occur half-way through what would be the normal recovery period. There were three patients who died in the ward before they could be admitted to the ICU. The median age of these three patients (81.6 years) was higher than that of the NR (63.5 years) and EICU (66.3 years) groups. These patients are not considered any further in the analysis.

Data gaps for the EICU patients, as defined in section 7.1.2.1, are shown in the last column of Table 7.2. Considering that a track-and-trigger entry is made every 4 hours on average, the gap duration of 3.6 hours for the EICU group suggests that approximately one observation set was not recorded prior to the admission, due to the fact that (as described in section 7.1.2.1) during this period the clinical staff are attending to the patient's condition and preparing them for ICU admission. The large interquartile range (IQR) of 12.3 hours for the EICU group is due to two patients whose track-and-trigger paper charts for their last 49 hours and 89 hours are missing, resulting in large data gaps for these patients. Excluding these two from the calculation results in a median gap duration of 3.2 hours with an IQR of 4.38 hours.

For a patient with two emergency ICU admissions or an emergency ICU admission followed by death, only the first emergency ICU admission event was considered for the analysis in the following section. A total of 19 such adverse events were identified in the $\mathcal{CH}_2$ dataset including 14 ICU admissions from the 14 patients with a single admission, the first of two events from each of the 2 patients with 2 ICU admissions, and 3 ICU admissions from the 3 patients who died in the ICU following their emergency admission.

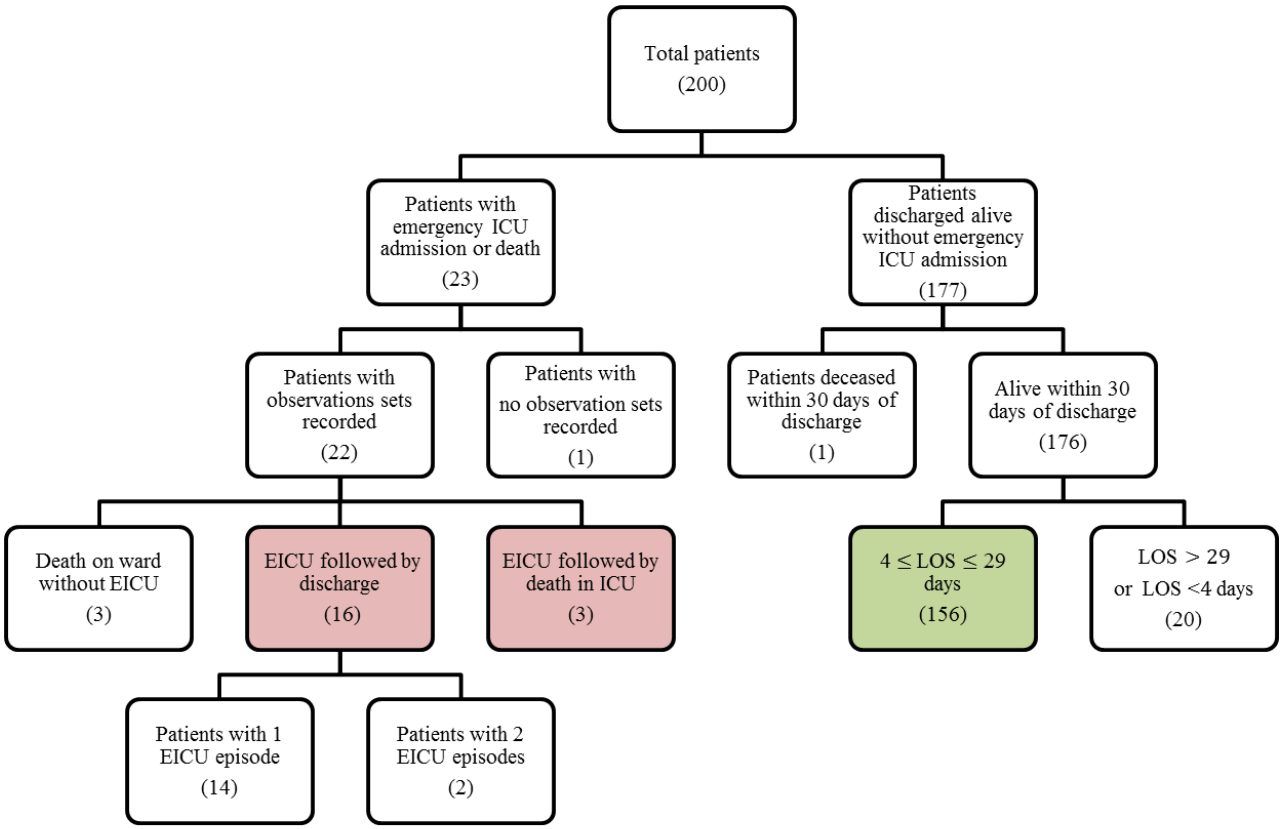


Figure 7.2: Categorisation of patients enrolled in phase I of the CALMS-2 trial. The number of patients belonging to each category is shown in brackets. LOS refers to the length of stay in the upper GI ward prior to discharge from the hospital. EICU refers to an emergency ICU admission. Boxes representing patients placed in the EICU group are indicated in pink, and boxes representing patients placed in the NR group are indicated in light green.

Patient Groups	# Patients	Age	Gender (f:m)	LOS	Data Gap
Normal Recovery (NR)	156	63.5 (15.6)	66:90	8.6 (6.2)	–
Emergency ICU (EICU)	19	66.3 (11.6)	8:11	4.9 (9.0)	3.6 (12.3)

Table 7.2: Patient groups in the $\mathcal{CH}_2$ dataset. For each patient group, the number of patients in the group is shown in the second column from the left. The median age in years, and the interquartile range in brackets, is shown in the third column. The median and interquartile range (in brackets) of the data gap duration in hours are shown in the last column for the EICU group. The LOS column refers to the length-of-stay prior to discharge for the NR group, and the length-of-stay prior to emergency ICU admission for the EICU group, respectively.

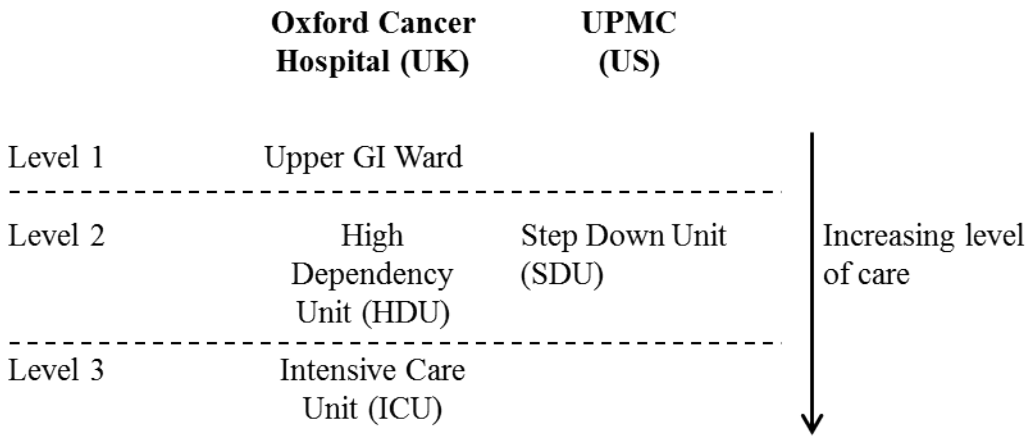


Figure 7.3: Equivalent levels of care in the UPMC (US) and the Oxford Cancer Hospital (UK).

7.3 Comparison of Datasets $\mathcal{D}_1$ and $\mathcal{CH}_2$

In order to assess if the methods developed using dataset $\mathcal{D}_1$ in previous chapters could be applied to dataset $\mathcal{CH}_2$, the vital-sign distributions belonging to the two datasets were compared. For comparison with the dataset $\mathcal{CH}_2$, we will use the test set $\mathcal{D}_1^{test}$ which is described in detail in section 3.4, and consists of a "normal" class C_1 which contains data representing examples of "normal" physiology, and an "abnormal" class C_2 which contains data representing examples of physiology during episodes of physiological deterioration (see Section 3.4).

As explained in section 3.2, adverse events in dataset $\mathcal{D}_1$ were labelled retrospectively by clinical experts, who reviewed the time-series plots of the vital signs in the period leading to the adverse event. The clinical experts were therefore able to annotate the beginning and end of an adverse event. Due to the intermittent nature of track-and-trigger monitoring, a similar annotation procedure is not possible for dataset $\mathcal{CH}_2$. The frequency of recording observations in track-and-trigger monitoring is variable (the typical monitoring frequency in the upper GI ward was 4-5 hours, which would increase subject to patient condition), whereas the continuous monitoring data in dataset $\mathcal{D}_1$ has a sampling rate of once every 20 seconds. Secondly, as described in section 7.1.2.1, the nurse may be occupied in preparing the patient for emergency ICU admission, which can create a gap in the data prior to the adverse event. From Table 7.2 it is known that, for the EICU group, there is a median data gap of 3.6 hours between the time of the last track-and-trigger

observation and the adverse event. Thus for a typical patient from the EICU group, the dataset $\mathcal{CH}_2$ contains data from before an adverse event rather than from both before *and* during the event as was the case with dataset $\mathcal{D}_1^{test}$.

Figure 7.4 presents one such example from the EICU group. Here, the patient's last track-and-trigger observation set was recorded approximately three hours before an emergency admission to ICU. Following an initial period of apparent recovery during the night, a clear rise in SBP, HR, and BR between 0900 and 1200 hrs indicates the onset of physiological deterioration. However the entire period of deterioration is not available for analysis as a result of the three-hour gap between 1230 and 1540 hrs (the time of ICU admission). One is therefore constrained to defining the period preceding an emergency ICU admission or death to be a *period of abnormality*, and define the "abnormal" class for dataset $\mathcal{CH}_2$ to be that which contains data from this period. Further, here the "normal" class is defined to be that which contains data belonging to patients from the NR group for a *period of normality* prior to discharge from the ward. A example of the vital-sign time-series of an NR patient's recovery is shown in Figure 7.5, where after an initial period of instability in the vital signs lasting approximately 4 days, they are settled for approximately the last 3 days prior to discharge, and hence an approximately 3-day period of normality. For other patients the period of normality may be longer or shorter than that. The selection of the period of abnormality and normality is explained next.

Figures 7.6 and 7.7 show a comparison of the distributions of HR and BR belonging to the "normal" and "abnormal" classes of datasets $\mathcal{D}_1^{test}$ and $\mathcal{CH}_2$. For dataset $\mathcal{CH}_2$ the "normal" class is considered to be that which contains data for all patients belonging to the NR group for the last n days (or hours) before discharge. Similarly, the "abnormal" class is considered to be that which contains data from patients from the EICU group for the last n days (or hours) before the emergency ICU admission. The figures show the change in the distribution of the vital signs belonging to $\mathcal{CH}_2$ with time, with the value of n changing from $n = 5$ days to $n = 0.25$ days (i.e. 6 hours) before discharge or event. For the "normal" class, in both HR and BR distributions, the overlap between datasets $\mathcal{D}_1^{test}$ and $\mathcal{CH}_2$ is consistent across the three different time windows shown in Figure 7.6 (a) - (c) and Figure 7.7 (a) - (c), for the HR and BR, respectively. While the distribution

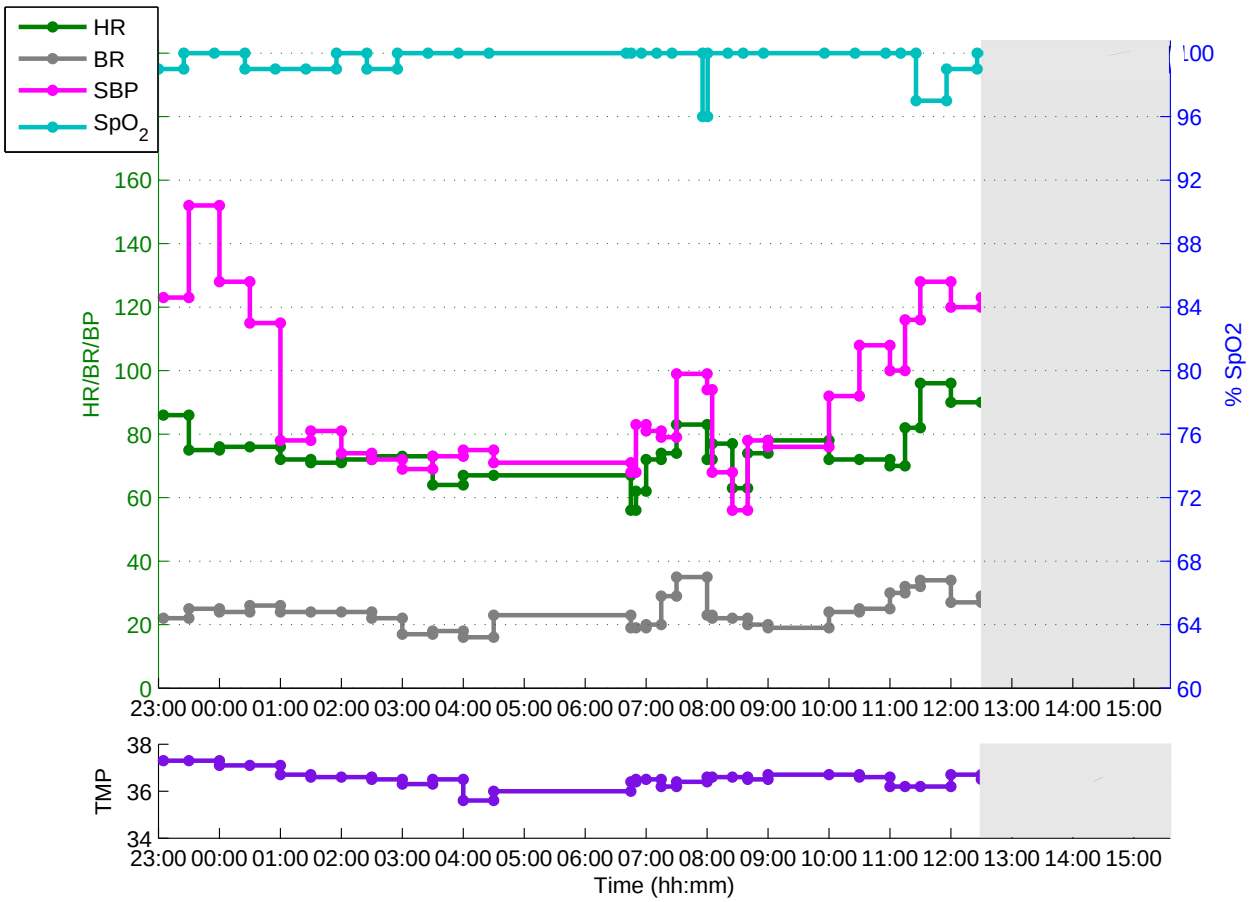


Figure 7.4: An example of deterioration in a patient's vital signs prior to emergency admission to the ICU. The region shaded in grey between 12:30 and 15:40 marks the period during which no track-and-trigger data were recorded. The last track-and-trigger observation set was recorded at 12:30. The time of admission to ICU is 15:40.

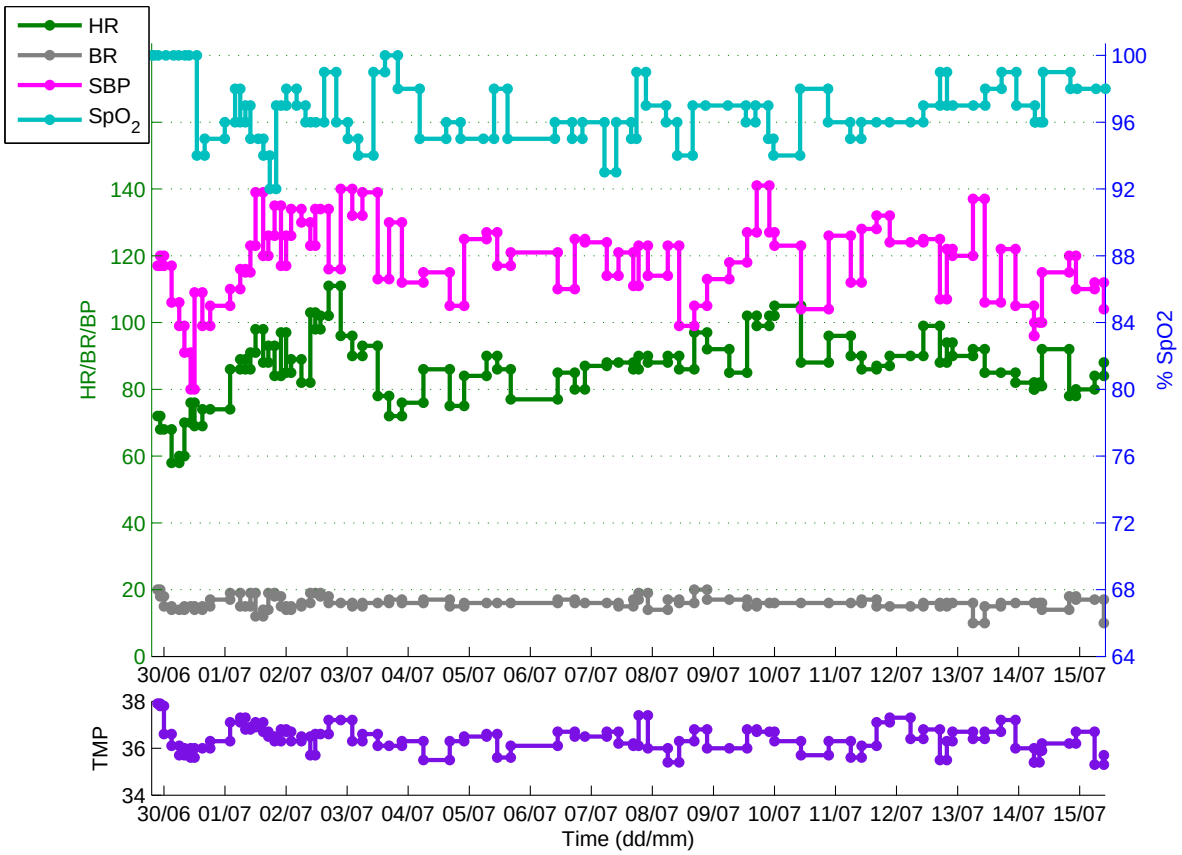


Figure 7.5: An example of the recovery time-series for the vital signs of a patient from the normal-recovery group. Following the initial admission to the upper GI ward on 30 June, the patient's vital signs undergo a period of recovery during the first four days, and are "settled" by the time of discharge on 15 July.

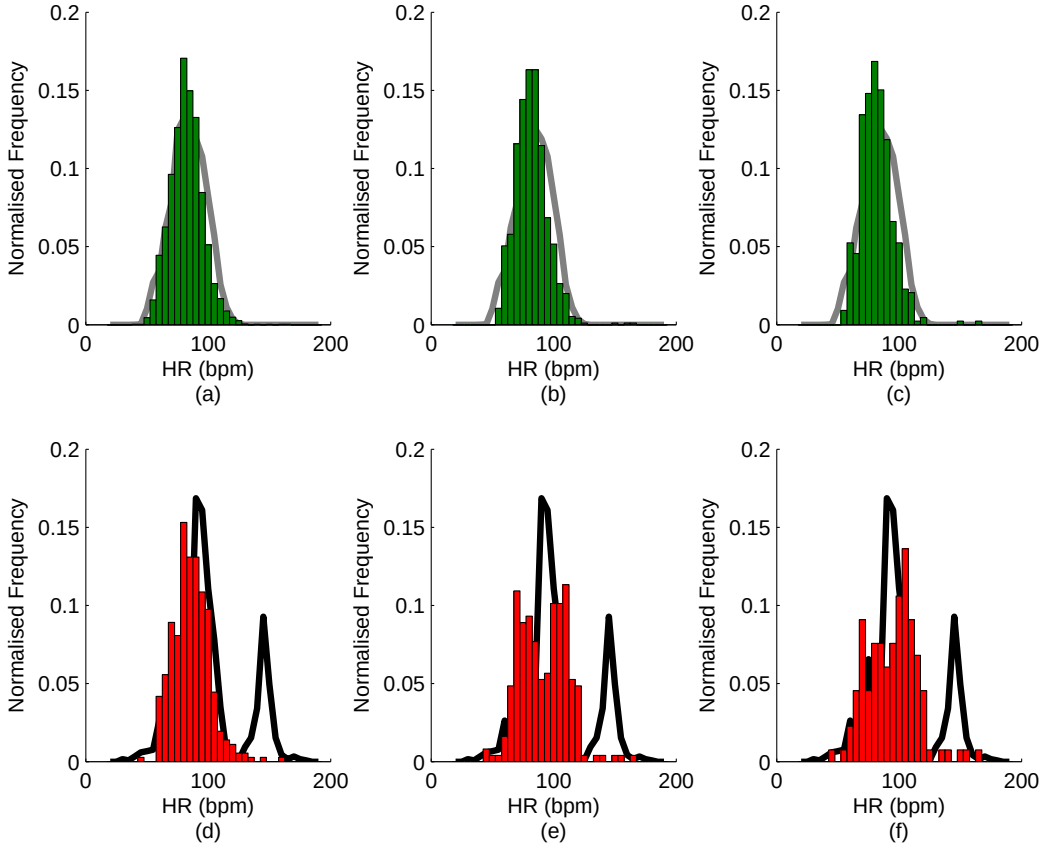


Figure 7.6: Comparison of the distribution of HR in datasets $\mathcal{CH}_2$ and $\mathcal{D}_1^{test}$. The distribution of HR for the "normal" class of $\mathcal{CH}_2$ for the last (a) five days, (b) one day, and (c) six hours before discharge is shown in green. Also shown in grey is the distribution of HR for the "normal" class of $\mathcal{D}_1^{test}$. The distribution of HR for the "abnormal" class of $\mathcal{CH}_2$ for the last (d) five days, (e) one day, and (f) six hours before event is shown in red. Also shown in black is the distribution of HR for the "abnormal" class of $\mathcal{D}_1^{test}$.

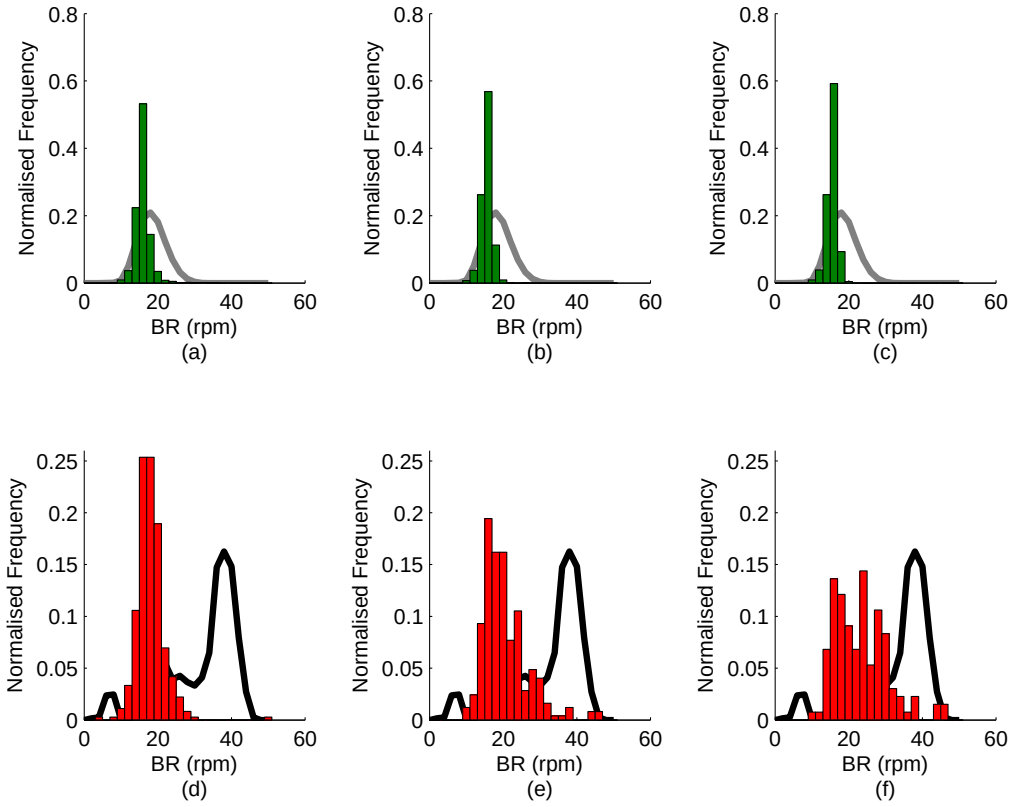


Figure 7.7: Comparison of the distribution of BR in datasets $\mathcal{CH}_2$ and $\mathcal{D}_1^{test}$. The distribution of BR for the "normal" class of $\mathcal{CH}_2$ for the last (a) five days, (b) one day, and (c) six hours before discharge is shown in green. Also shown in grey is the distribution of BR for the "normal" class of $\mathcal{D}_1^{test}$. The distribution of BR for the "abnormal" class of $\mathcal{CH}_2$ for the last (d) five days, (e) one day, and (f) six hours before event is shown in red. Also shown in black is the distribution of BR for the "abnormal" class of $\mathcal{D}_1^{test}$.

of HR for $\mathcal{D}_1^{test}$ and $\mathcal{CH}_2$ show a clear overlap, the BR distribution of $\mathcal{CH}_2$ is shifted to the left, and covers a much smaller range than that of the BR distribution for $\mathcal{D}_1^{test}$.

Next the distribution of the "abnormal" class for the two datasets is considered. The distribution of HR for $\mathcal{CH}_2$ is distributed around the lower mode of the distribution for $\mathcal{D}_1^{test}$ (centred at approximately 80 bpm) and is changing with time to spread out in the tails (Figure 7.6(d) - (f)). The difference between the two datasets is more prominent in the distribution of the BR for the "abnormal" class (Figure 7.7(d) - (f)). Here we see that at $n = 5$ days, the distribution for $\mathcal{CH}_2$ has a much lower mode (centred at approximately 18 rpm) than that of $\mathcal{D}_1^{test}$ (which is centred at approximately 40 bpm), and that the distribution for $\mathcal{CH}_2$ gradually spreads out to take higher values of BR (up to 45 rpm) as n changes from $n = 5$ days to $n = 0.25$ days (i.e., 6 hours). The smaller range of BR in the $\mathcal{CH}_2$ "normal" class as compared to that of the $\mathcal{D}_1^{test}$ "normal" class can be explained by what may be considered as an "observer effect", according to which the patient's BR can be increased in the presence of the nurse [126]. Nurses tend to account for that "observer"-related rise by underscoring, i.e., by not recording the most extreme value. Secondly, in $\mathcal{D}_1$ the BR is estimated using impedance plethysmography (as described in section 3.2), whereas in $\mathcal{CH}_2$ the BR is measured manually by the nurse. Typically, the nurse counts the number of breaths for 15 (or 30) seconds and multiplies by 4 (or 2) to estimate the per-minute breathing rate. Due to this, the nurse-estimated breathing rate typically takes even-numbered integer values, explaining the large peak in the histogram at the bin centred at 16 bpm.

The comparison of the bi-dimensional datasets can be seen in Figure 7.8. In Figure 7.8(a) it may be seen that for $n = 5$ days, the cluster of data (shown in red) representing the "abnormal" class of $\mathcal{CH}_2$ is overlapping the cluster (shown in green) representing its "normal" class, indicating that there is little separation between the "normal" and "abnormal" classes of $\mathcal{CH}_2$, where the "normal" class contains all data for the last five days before discharge and the "abnormal" class contains all data for the last five days before an emergency ICU admission. Further, it may be seen that the "normal" as well as the "abnormal" cluster for $\mathcal{CH}_2$ is overlapping the cluster representing the "normal" class for $\mathcal{D}_1^{test}$ (shown in grey).

The separation between the red and green clusters increases as n decreases from $n = 5$ days (Figure 7.8(a)) to $n = 1$ day (Figure 7.8(b)). At $n = 0.25$ days (Figure 7.8(c)), this separation becomes more distinct, as the "abnormal" class of $\mathcal{CH}_2$ (red dots) moves towards the tachycardia- and tachypnoea-related clusters of the "abnormal" class of $\mathcal{D}_1^{test}$ (black dots), but without overlapping them. Finally at $n = 0.042$ (= 1 hour), the separation is the greatest, however it can be seen that, as expected, there are very few samples from the "abnormal" class of $\mathcal{CH}_2$ (red dots) corresponding to one hour before the EICU admission.

This behaviour is as expected given that the "abnormal" data from $\mathcal{CH}_2$ are taken from the period *before* an event occurs and are therefore not as far from the region of "normality" as the $\mathcal{D}_1^{test}$ "abnormal" clusters are.

The reasons for focusing this analysis on the last five days are as follows. In [125], it was shown that the physiology of patients in the upper GI ward is unstable over the course of the first 4 days after surgery, and after 4 days the physiology becomes more stable for the majority of the "normal recovery" patients. Given that the average LOS for the NR group is 8.6 days (Table 7.2), this analysis has therefore been limited to the last five days. Further the analysis included HR and BR based on the knowledge that 4 of the 5 types of vital-sign abnormalities identified in the $\mathcal{D}_1^{test}$ are associated with the HR and BR, namely, tachycardia, bradycardia, tachypnoea, and bradypnoea. Based on the results, further analysis including SpO_2 information was not performed. The following points may be noted from the comparison of the $\mathcal{D}_1^{test}$ and $\mathcal{CH}_2$ datasets.

- At $n = 5$ days (Figure 7.8(a)), the $\mathcal{CH}_2$ "normal" cluster (in green) nearly overlaps the $\mathcal{D}_1^{test}$ "normal" cluster (in grey); however at $n = 0.042$ days, i.e. 1 hour (Figure 7.8(d)) before discharge, the $\mathcal{CH}_2$ "normal" cluster (in green) covers a much smaller subset of the region covered by the $\mathcal{D}_1^{test}$ "normal" cluster (in grey), indicating that the $\mathcal{CH}_2$ region of normality is much smaller than that of the model of normality constructed using $\mathcal{D}_1^{test}$. The reason for this becomes clear upon comparison of the patient characteristics of the two datasets: $\mathcal{D}_1^{test}$ contains data from patients in a step-down-unit which is classified as a Level 2 unit, which is one level below the ICU (see Figure 7.3). $\mathcal{CH}_2$, on the other hand, contains data from patients who are in the upper GI ward which is a Level 1 unit, and is two levels below the ICU.

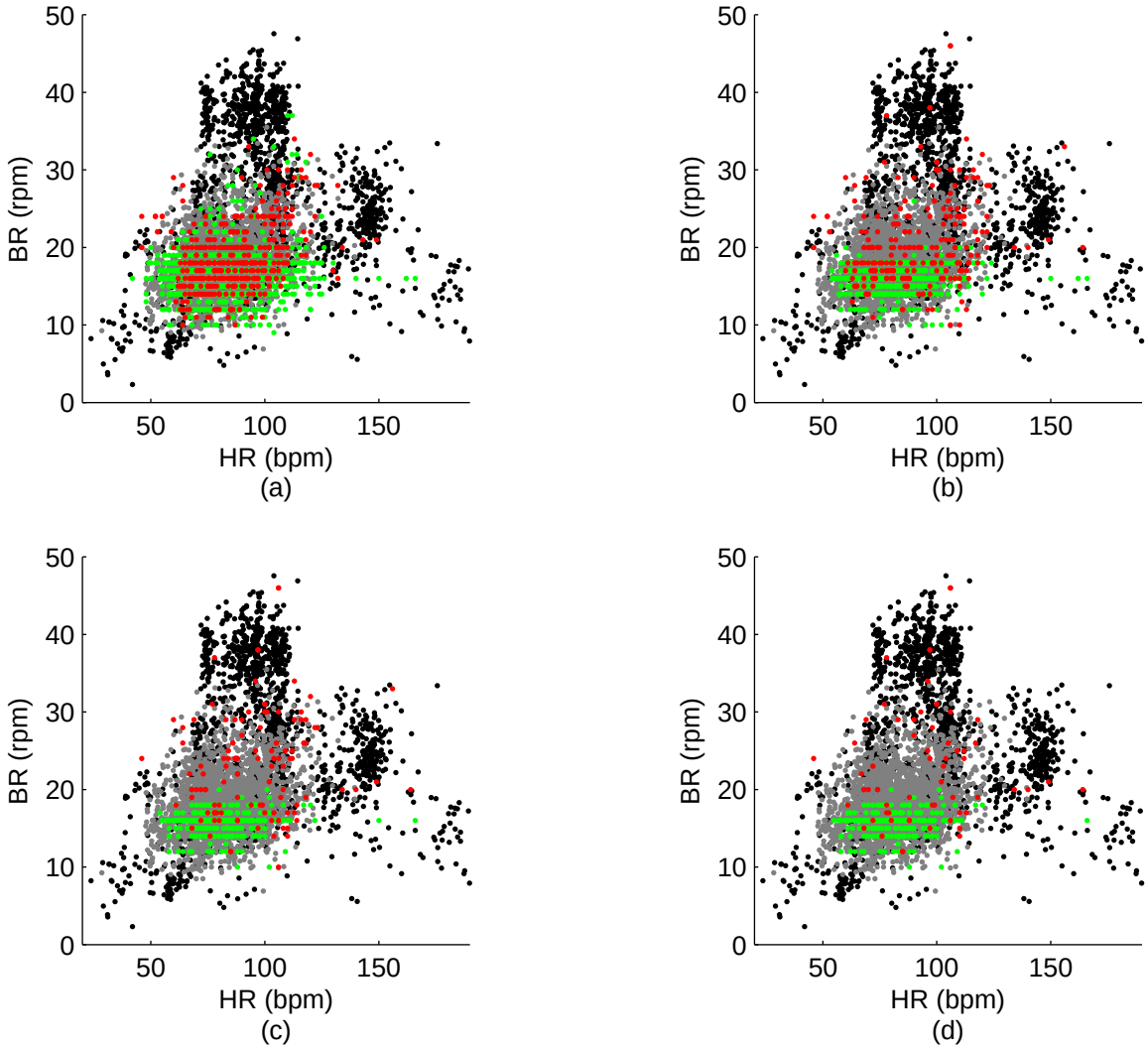


Figure 7.8: (a) $n = 5$ days, (b) $n = 1$ day, (c) $n = 0.25$ days, (= 6 hours), (d) $n = 0.042$ days (= 1 hour). The distribution of "normal" HR and BR data for patients in the NR group for the last n days prior to discharge is plotted in green. The distribution of "abnormal" HR and BR data for patients in the EICU group, for the last n days prior to emergency ICU admission, is plotted in red. "Normal" data from the "normal" class and "abnormal" class of dataset $\mathcal{D}_1^{test}$ are plotted in the background in black and grey, respectively, in (a) to (d).

The patients in this unit are expected to be less seriously ill than those in a level 2 step-down-unit. For this reason (as well as due to the nurse-observation effect described earlier in this section) one can expect to see fewer extreme data in $\mathcal{CH}_2$ as compared to $\mathcal{D}_1^{test}$.

- The distribution of "normal" HR and BR for $\mathcal{CH}_2$ from the last five days (shown in green in Figures 7.6(a) and 7.7(a)) is very similar to the distribution six hours before discharge (shown in green in Figures 7.6(c) and 7.7(c)), indicating that patients may become stabilised up to five days before discharge.

- The $\mathcal{CH}_2$ "abnormal" (red) and "normal" (green) clusters have significant overlap five days prior to event or discharge (Figure 7.8(a)), and as expected the overlap decreases as the event or discharge draws closer (Figure 7.8(a)). However even during the last 6 hours before the event, some $\mathcal{CH}_2$ "abnormal" data remain within the $\mathcal{CH}_2$ "normal" region. In the next section, the time-based trends in vital signs in dataset $\mathcal{CH}_2$ are investigated to understand better the differences in the physiology of the NR and EICU patients.

7.4 Time-varying Trends

The vital signs of NR and EICU patients may typically be expected to differ in mainly two ways: (a) vital sign(s) departs from the range of normality. In this case the deterioration can be indicated by the change in the amplitude of the vital sign(s); (b) The vital-sign trajectory does not depart from the range of normality. In this case the amplitude of the vital sign does not indicate deterioration, however it may be that (i) the deterioration is caused by increased variability in the vital sign, or (ii) that the deterioration is due to departure from a patient-specific normal range.

From Figure 7.9(a) it may be seen that for the NR group there is a clear peak in the distribution of the LOS at LOS=8 days, and from Table 7.2 it is known the median LOS on the ward before discharge is 8.6 days for the NR group. However, for the EICU group, except for two patients who have LOS of 22 and 23 days, the remaining 17 EICU patients have an LOS in the ward before emergency ICU admission of 12 days or fewer. The time

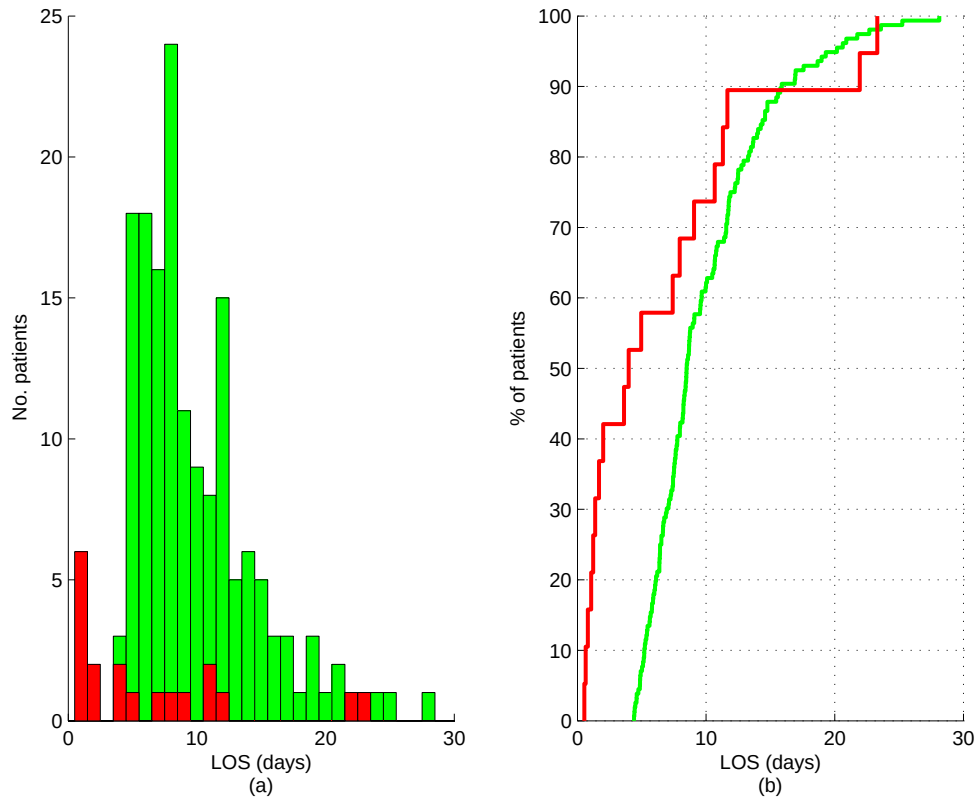


Figure 7.9: (a) Frequency distribution, and (b) cumulative distribution of the LOS of patients in the NR and EICU groups, shown in green and red, respectively.

Day	1	2	3	4	5	6	7	8	9	10	11	12
No. NR patients	156	156	156	156	145	127	109	90	68	59	–	–
No. EICU patients	19	13	11	11	9	8	8	7	6	5	5	3

Table 7.3: The number of NR and EICU patients who were monitored in the upper GI ward on a given day, starting with the first day on the ward.

course including the first 12 days of ward stay is therefore considered for the EICU group. For the NR group the first 10 days are considered.

7.4.1 Trends in Amplitude and Variability

For each patient in the NR group, the mean vital-sign value for each day was calculated, beginning with the day of ward entry. On day 1, for example, 156 patients were in the ward (see Table 7.3). However, on day 10, 59 out of the 156 patients were in the ward (as shown in the same table). Therefore the average of the mean daily vital-sign values for day 1 was computed by taking the average of the 156 patients' mean daily values; similarly, the average of the mean daily vital-sign values for day 10 was computed by taking the average of the 59 patients' mean daily values, and referred to as the average

daily vital-sign value for the group, as shown in Figure 7.10. This was done to account for the changing number of patients in the NR and EICU groups on different days. Table 7.3 shows the number of patients from the NR group in the ward on a given day. The numbers from Table 7.3 which are used to compute the daily average on a given day are repeated in Figure 7.10 (f) and (g).

The daily vital-sign variability was also computed for the two groups. Variability was defined as the total distance travelled by a normalised[4] vital-sign vector during a daily time window. 1-D variability was computed by simply calculating, e.g., for the HR: $\sum_n^N |\mathrm{HR}_n - \mathrm{HR}_{n-1}|$, where N is the number of vital-sign observations in a daily time window, and $n = 1...N$.

The variability for the 5-D vital-sign vector was also estimated, by calculating the total Euclidean distance travelled by the 5-D vital-sign vector in a daily window:

$$\text{5-d variability} = \sum_{n}^{N} \sqrt{\left(\mathrm{HR}_n - \mathrm{HR}_{n-1}\right)^2 + \left(\mathrm{BR}_n - \mathrm{BR}_{n-1}\right)^2 ... \left(\mathrm{TMP}_n - \mathrm{TMP}_{n-1}\right)^2} \tag{7.1}$$

The 1-D variability trends for the NR and EICU groups are shown in Figure 7.12 (a) - (e), and the 5-D trend for the two groups is shown in Figure 7.12 (f).

From Figures 7.10 and 7.12 the recovery trend may be seen in the NR group, e.g., the SBP starts low on the first day in the ward and by the fifth day it increases to reach a stable level which is maintained until discharge. For the other vital signs a clear trend in the amplitude is hard to discern. However, for each vital sign a decrease in the variability from the first day to the day of discharge may be seen for the NR group (Figure 7.12 (a) - (e)). Moreover, variability in the 5-D vector for the NR group is also reduced by half from the start of ward stay to the time of discharge (Figure 7.12 (f)). This indicates that both the variability in individual vital signs, as well as the overall variability in the five vital signs of the patient reduces as the patient recovers.

Comparing this with the EICU group, in Figure 7.12 (f) an offset can be seen in the

[4]Normalisation here was performed following the technique described in Equation 3.1, and using the respective mean and ±1 std of the NR and EICU groups. Normalisation was performed to ensure that the difference in the units of measurement of different vital signs did not influence the estimation of the variability.

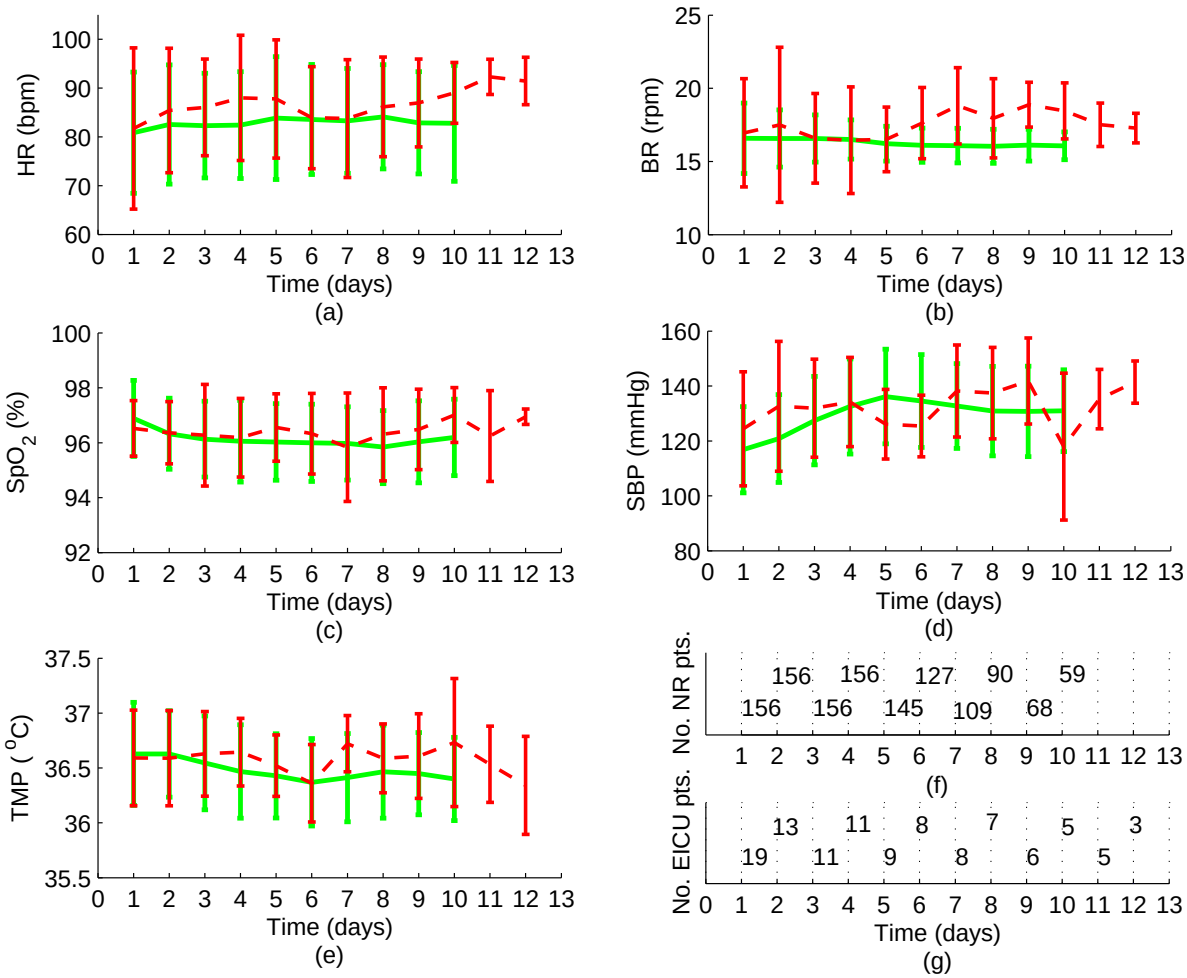


Figure 7.10: Vital-sign trajectories for the NR (green) and the EICU (red) groups. For each group, the average daily vital signs (defined in Section 7.4.1) are shown in (a) to (e). (f) The number of NR patients (also shown in Table 7.3) for whom the average daily vital-sign value for the NR group was calculated on the corresponding day in ward.(g) The number of EICU patients (also shown in Table 7.3) for whom the average daily vital-sign value for the EICU group was calculated on the corresponding day in ward.

EICU and NR 5-D variability, indicating that the 5-D variability is, as expected, higher in the EICU group. However, no clear differences between the NR and EICU groups emerge in the individual vital signs (Figure (7.12 (a) - (e)). This is not unexpected; there can be many different types of abnormality, and in fact, the definition of abnormality can be patient-specific (i.e., different for different patients). As there are only 19 patients in the EICU group, a generic picture is difficult to form. For this reason, the amplitude and variability trends for individual EICU patients were also calculated (Figures 7.11 and 7.13), and confirm that indeed the amplitude and variability trends are variable for individual patients, and that EICU patients are therefore best considered in this analysis individually.

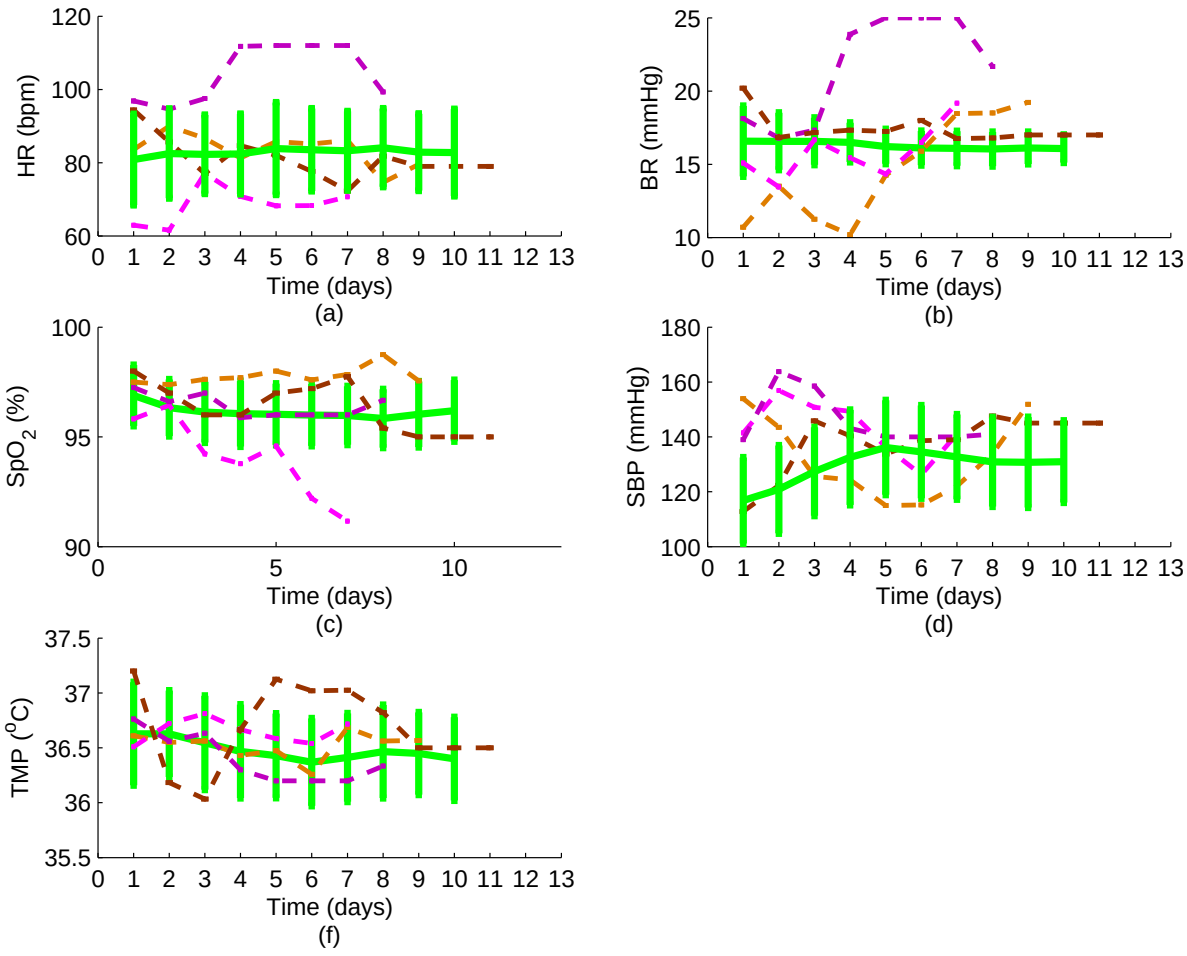


Figure 7.11: Vital-sign amplitude trajectories for the NR group, shown in solid green, and four example individual patients in the EICU group, shown in dashed coloured lines. For each EICU patient, the daily vital-sign values (defined in Section 7.4.1) are shown.

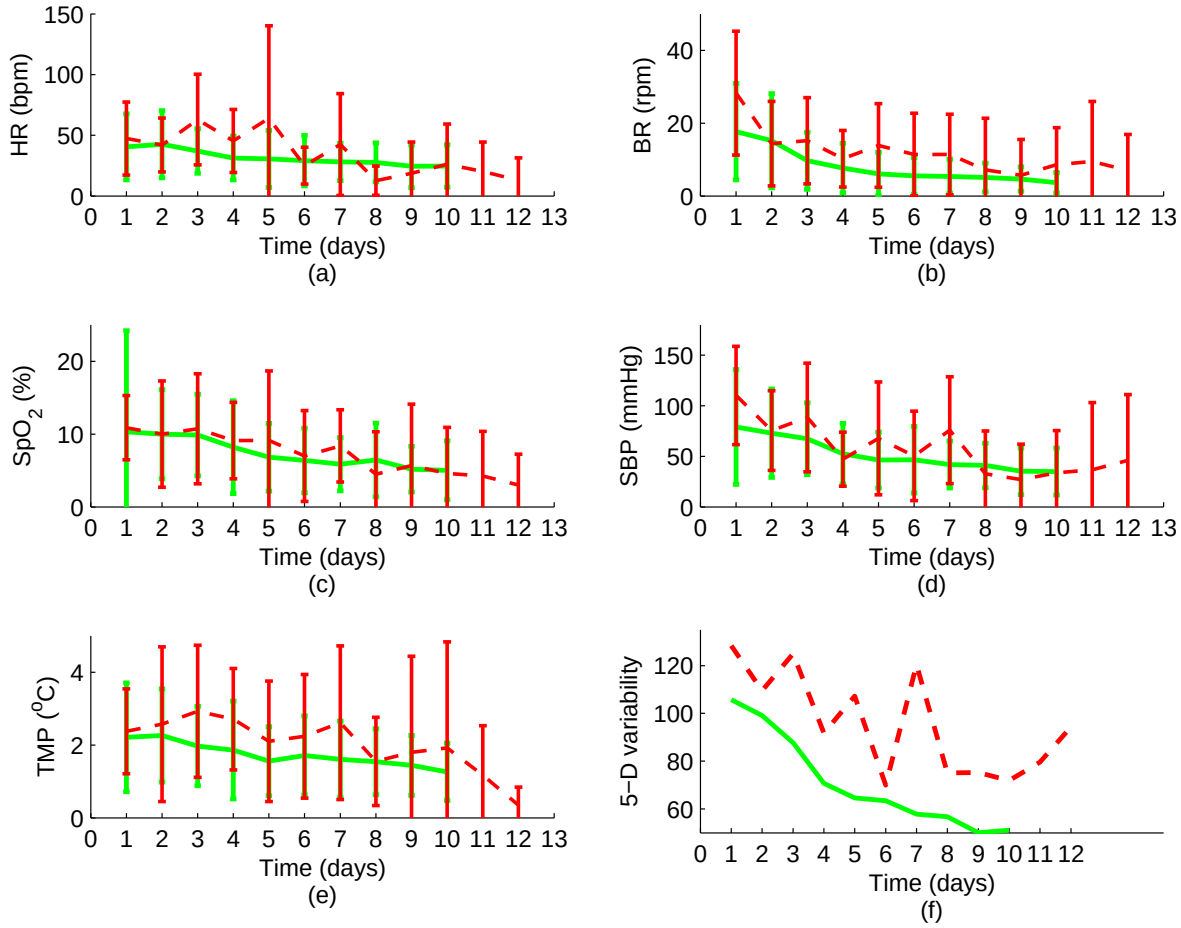


Figure 7.12: (a) to (e): 1-D vital-sign variability trajectories for the NR group, shown in solid green, and for the EICU group in dashed, red line. (f) the 5-D vital-sign variability trajectory shown for the NR in green and EICU in red.

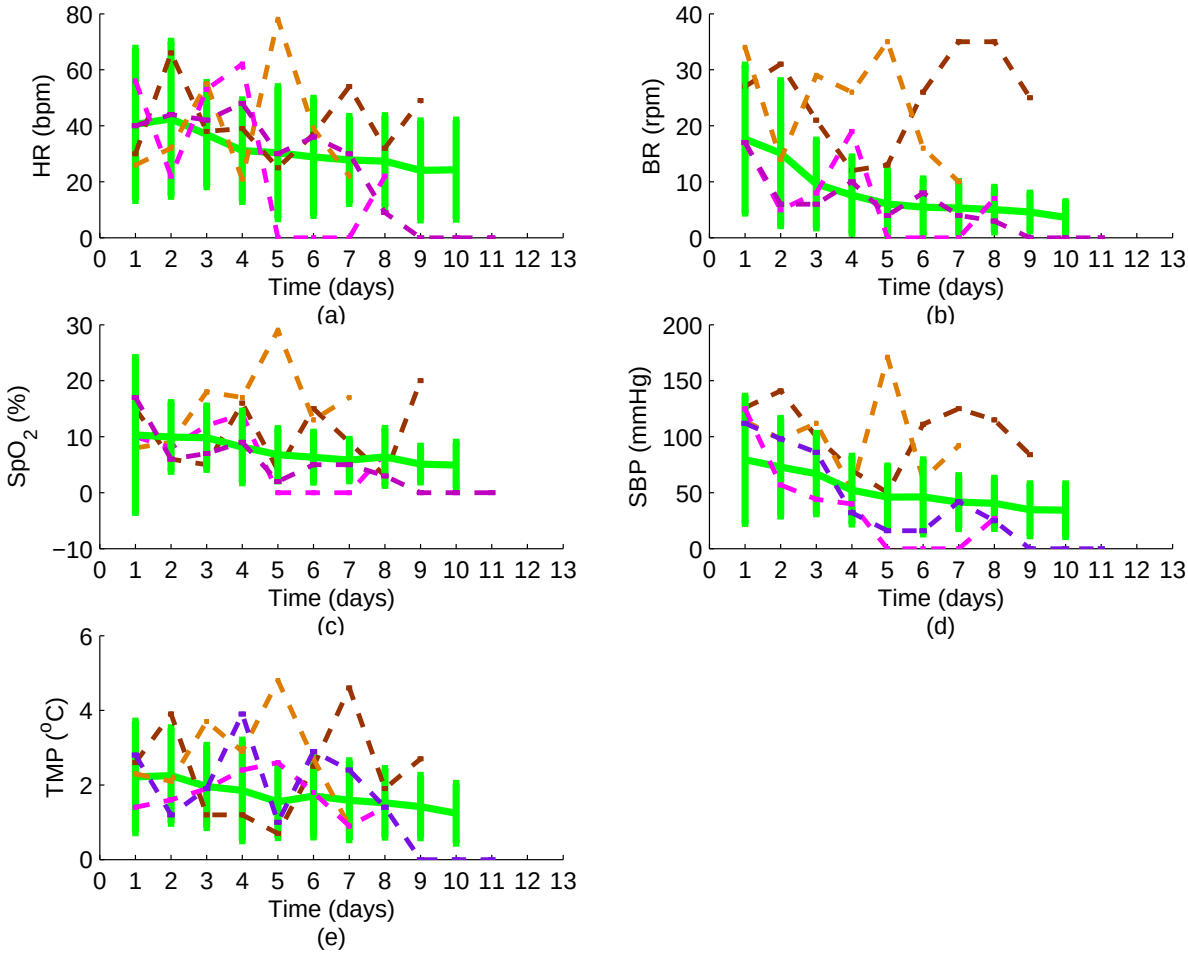


Figure 7.13: Vital-sign variability trajectories for the NR group, shown in solid green, and individual patients in EICU group, are shown in dashed coloured lines. For each EICU patient, the daily variability (described in Section 7.4.1) is shown.

7.4.2 Last Day Analysis

The analysis of the NR and EICU groups is brought to a close here by focusing on the last day of ward stay for patients from the two groups. It is expected that any differences between normally recovering patients and those who have an adverse event will be most evident in the hours just prior to the discharge or adverse event, respectively. With this view, the vital-sign trajectories of the two groups were plotted looking back from the time of the last observation. For the EICU group, the last observation refers to the last observation recorded in the upper GI ward prior to the emergency ICU admission, and for the NR group it refers to the last observation recorded before discharge from the hospital.

Here, mean vital-sign values were calculated at one-hourly intervals instead of daily intervals as in the previous section, and the plots are shown in Figure 7.14. Since observations were recorded for an NR patient on average, every 4 to 5 hours, a sliding window of 4 hours was used for the NR group. This means that, for example, the mean vital-sign value for a patient during the last hour interval is approximated by the mean of the vital-sign values in the previous four hours. For the EICU group, although the observation frequency of individual patients was variable, observations were recorded were a higher rate[5] than for the NR group, therefore for the EICU case a sliding window of 2 hours (shown in red) was considered, as well as one of 4 hours (shown in pink). The results show no clear separation between the NR and EICU trajectories for the SpO_2, SBP and temperature. However, for the HR, it may be seen that while there is no separation between the NR and EICU trajectories 48 hours before the last observation, the separation increases until there is a difference of 16 bpm at the last hour. Similarly, for the BR, there is a difference of only 3 rpm 48 hours before the last observation, which increases to a difference of 8 rpm in the last hour. Further, it may be seen that for all vital signs for the NR group change minimally in the last 24 hours.

Finally, the ± 1 standard deviation error in the mean vital-sign values is also plotted separately (in Figure 7.15), which show similar trends: (i) there is more variation in the vital signs of the EICU group in the last 24-hour period than there is in the NR group,

[5]The average number of observations recorded for an EICU patient in the last day is 12.2, whereas the average number of observations recorded for an NR patient in the last day is 5.3. Therefore the monitoring frequency for the EICU patients (once every $12.2/24 < 2$ hours) in the last day is approximately twice as high as it is for the NR patients (once every $5.3/24 \approx 4.5$ hours).

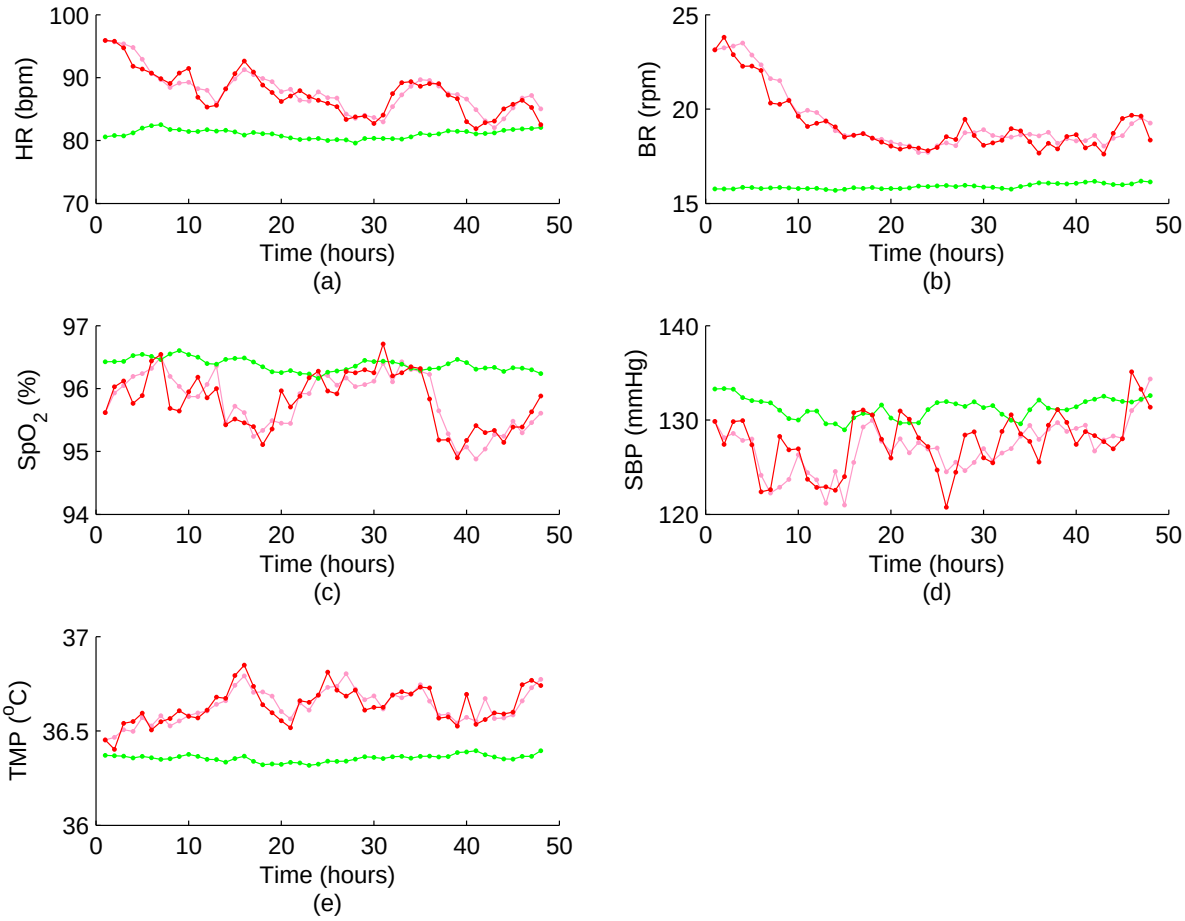


Figure 7.14: Trajectories of the mean vital-sign values for the last 48 hours of stay in the ward before discharge, for the NR group, and the last 48 hours in the ward before emergency admission to the ICU, for the EICU group. A sliding window of 4 hours was used for the NR group (green), and a sliding window of 4 hours (pink), and a sliding window of 2 hours (in red) was used for the EICU group.

and (ii) this variation continually increases in the last 24 hours in the EICU group but not in the NR group. These trends indicate that, although beyond the scope of this work, a study of the dynamics of vital signs could prove insightful.

7.5 Discussion

In this chapter, the vital-sign distributions of the $\mathcal{D}_1$ and $\mathcal{CH}_2$ datasets were compared. It was argued that models developed using $\mathcal{D}_1$ are not ideally suited for application to $\mathcal{CH}_2$, and hence new models should be trained for the latter dataset. $\mathcal{CH}_2$ does not contain clinical labels similar to those available for $\mathcal{D}_1$. However, outcomes such as "discharge from ward without event", "emergency ICU admission", "death on ward" are known for the patients in $\mathcal{CH}_2$. Therefore patients were categorised according to their outcomes into

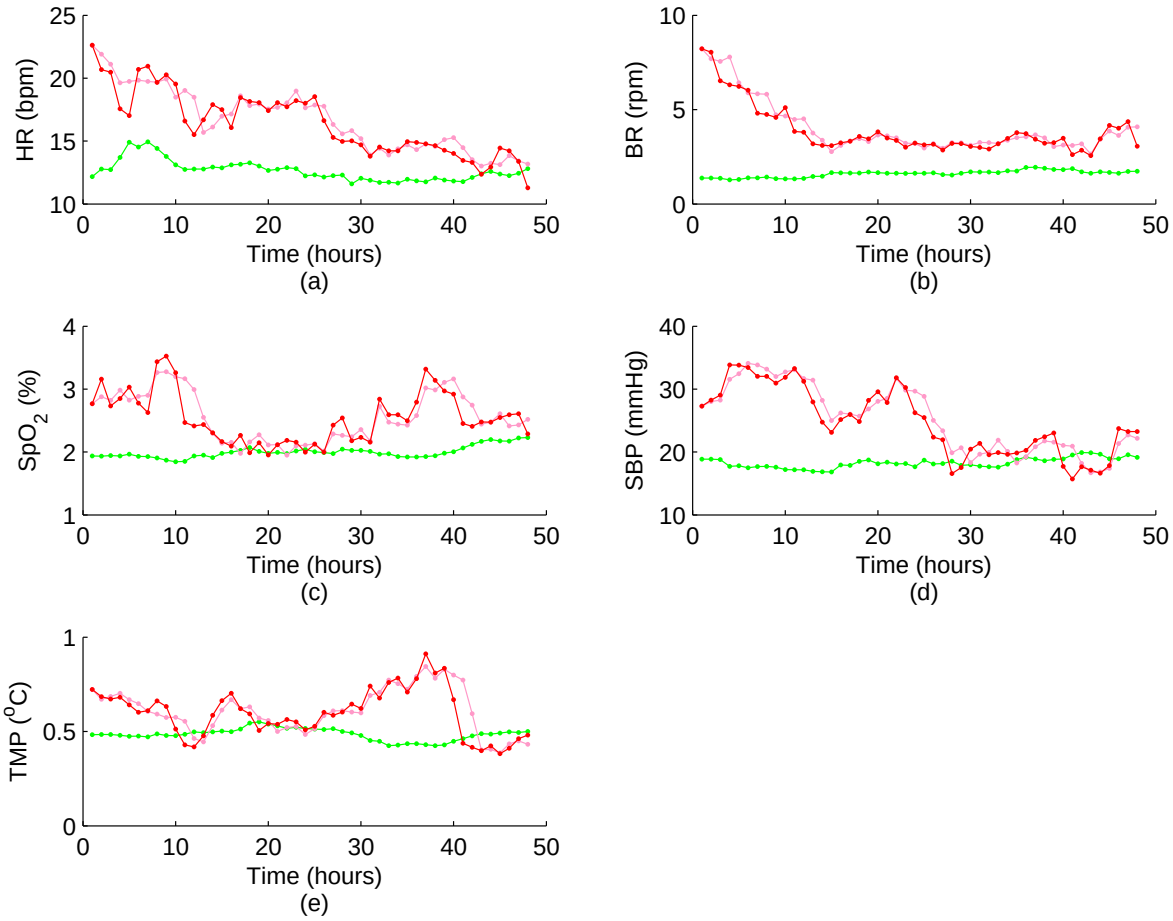


Figure 7.15: Trajectories of the ±1standard deviation in vital-signs for the last 48 hours of stay in the ward before discharge, for the NR group, and the last 48 hours in the ward before emergency admission to the ICU, for the EICU group. A sliding window of 4 hours was used for the NR group (green), and a sliding window of 4 hours (pink), and a sliding window of 2 hours (in red) was used for the EICU group.

two groups: NR, and EICU. The vital-sign trends for these two groups were compared and the results showed amplitude- and variability-based differences in the vital signs associated with normal recovery and deterioration.

In Section 7.4.1 it was shown that deterioration in vital signs can follow different trajectories for individual patients. It was found that the 5-D variability in the NR group reduces by half from day 1 on the ward to the last day before discharge. Further, since results were produced for a small set of EICU patients, a generic variability trend could not be established in individual vital signs, but it was found that EICU patients have higher 5-D variability than NR patients. However, despite small numbers it was shown in 7.4.2 that a clear increase in departure occurs in vital signs associated with deterioration on the upper GI ward in the last 48 hours. This latter analysis was limited to the last 48 hours only to reduce the influence of variable LOS on the results. Finally, the last 24-hour period was seen to be the most stable or "normal" period for the NR group (according to Figures 7.14 and 7.15).

In the next chapter, models are considered for detecting deterioration in patients in the upper GI ward. As vital signs for normal-recovery patients have been shown in this chapter to be stable in the last 24 hours, the last 24-hour period will be the focus of the next chapter.

Chapter 8

Models for Detecting Deterioration on the Upper GI Ward

8.1 Introduction

In Chapter 7, dataset $\mathcal{CH}_2$ was introduced and a comparison was made between datasets $\mathcal{CH}_2$ (which was collected during the CALMS-2 study) and $\mathcal{D}_1$ (which was collected from the UPMC study, as described in Chapter 3). The need for calibrating novelty detection models to dataset $\mathcal{CH}_2$ was identified, and it was shown that vital signs for patients in the normal-recovery (NR) group were most stable during the last 24 hours before discharge from the hospital. On the other hand, the vital signs of patients in the emergency ICU (EICU) group became increasingly deviant from those of the NR group in the last 24 hours, indicating that the deterioration in the EICU vital signs was greatest in the last 24 hours before the emergency ICU admission. This deterioration was evident in both the amplitude and the variability of the vital signs.

This chapter presents models for detection of deterioration in patients in dataset $\mathcal{CH}_2$. The methods discussed in this chapter focus on the major difference between datasets $\mathcal{CH}_2$ and $\mathcal{D}_1$; i.e., that of the rate of recording vital-sign observation sets. First, Section 8.2 describes how models are assessed in this chapter. Then a density-based novelty detection model for $\mathcal{CH}_2$ is constructed using Parzen windows in Section 8.4 as was done for $\mathcal{D}_1$ in previous chapters. The novelty detection (or one-class) approach is taken here as there are insufficient "abnormal" data to perform two-class classification (as was done with

dataset $\mathcal{D}_1$ in Chapter 5), given that there are only 19 emergency ICU examples. Further details of why a standard classification scheme is not used are provided in Section 8.2. In Section 8.5 it is explained how the low frequency of track-and-trigger observations in $\mathcal{CH}_2$ motivates the need for a *patient-based* novelty detection model. A Bayesian patient-based model based on this motivation is then presented in Section 8.6. Finally, in Section 8.9 the proposed method and the Parzen windows method are compared with the track-and-trigger early warning score (EWS) systems that are currently in use in hospitals in the UK.

8.2 Assessing Model Performance

In the design of a typical classification model, a train-validate-test approach is taken. Dataset $\mathcal{CH}_2$, as discussed in Chapter 7, does not contain clinical labels for "normal" and "abnormal" data. The dataset does not contain obvious episodes of abnormality that may be labelled retrospectively, due to the fact that, for a patient from the EICU group, the track-and trigger observations were recorded by the nurse up to some period prior to an emergency ICU admission. However, for dataset $\mathcal{CH}_2$ patient outcomes such as "emergency ICU admission", "death on the ward", and "discharge without event" are known. Therefore, *outcome-based* labels will be used here. In addition, if at any point during the last 24 hours of a patient's vital-sign recording an alert is generated, then the patient is considered to have had a positive event. (The choice of the last 24 hours is explained earlier in Section 8.1.) Thus, only one event may occur per patient. It might be argued that, in the clinical setting, it is only the first alert that counts[1]; i.e. one alert is sufficient to indicate that the patient is unwell and needs escalation of care.

The EICU and the NR groups are as described in Section 7.2. If an alert is generated for a patient from EICU group, it is deemed to be a true positive (TP), and if an alert is generated for a patient from the NR group, it is deemed to be false positive (FP).

Here an alert is considered to have been generated corresponding to a vital-sign observation set when, for a given method, some threshold T on a novelty score is crossed. The

[1]However, this may not always be the case. In a continous monitoring set-up, for instance, a predetermined "window" of alerts is considered to trigger an alarm, rather than simply the first alert.

	$\mathcal{CH}_2^{EICU}$	$\mathcal{CH}_2^{NR_{test}}$	$\mathcal{CH}_2^{NR_{train}}$
No. patients	19	19	137
Last-day (hours)	24.4	28.1 ± 0.9	28.1 ± 0.5
HR (bpm)	89.7	82.1 ± 3.3	81.5 ± 0.5
BR (rpm)	20.6	15.8 ± 0.2	15.8 ± 0.0
$SpO_2(\%)$	96.2	96.4 ± 0.3	96.4 ± 0.0
SBP (mmHg)	123.9	131.0 ± 2.9	130.7 ± 0.4
TMP (°C)	36.6	36.4 ± 0.1	36.4 ± 0.2

Table 8.1: Training set, $\mathcal{CH}_2^{NR_{train}}$, and the test set containing $\mathcal{CH}_2^{EICU}$ and $\mathcal{CH}_2^{NR_{test}}$. The "last-day" refers to the mean duration of the last-day period. As $\mathcal{CH}_2^{NR_{test}}$ and $\mathcal{CH}_2^{NR_{train}}$ are randomly selected, the mean and ±1std. of the last-day duration, HR, BR, SpO_2, SBP, and temperature, taken by averaging the results of 50 experiments, are shown for these two groups.

novelty score and its corresponding threshold T may be defined differently for different methods, as explained in subsequent sections of this chapter. The performance of a model is assessed over a range of values of T.

Now it is assessed if an alert is generated during the last 24 hours before the emergency ICU admission for a patient from the EICU group, or before discharge from the hospital for a patient from the NR group. r_1 is defined to be the percentage of patients for whom at least one alert is generated during the last 24 hours. A better model is expected to result in a low value of r_1 for the NR group and a high value of r_1 for the EICU group. Thus the r_1 rate for the EICU group measures the *sensitivity* of the system, and the r_1 rate for the NR group measures the *false alert rate*.

8.3 Datasets

From the NR and EICU groups the datasets $\mathcal{CH}_2^{NR_{train}}$, $\mathcal{CH}_2^{EICU}$, and $\mathcal{CH}_2^{NR_{test}}$ are derived, as shown in Table 8.1. These datasets are used for training and testing the methods used in this chapter, and are described as follows.

Dataset $\mathcal{CH}_2^{NR}$ Dataset $\mathcal{CH}_2^{NR}$ is defined here such that it contains data belonging to each NR patient for the last 24 hours prior to discharge from the upper GI ward, for whom the length-of-stay (LOS) before discharge in the upper GI ward was at least four days and at most 29 days long. The reason for selecting this range of LOS was that the first 4 days were considered to be the recovery period in which the vital signs of a given

patient are unstable [125]; a LOS of more than 30 days indicated that the patient had not recovered enough and was awaiting transferral. For a patient in $\mathcal{CH}_2^{NR}$, the vital-sign observation sets recorded in the last 24 hours are included. The observation set recorded closest to 24 hours before the last observation set recorded prior to discharge was included. For example, if a patient has an observation set recorded 28 hours before, and the next observation set recorded 20 hours before the last observation prior to discharge, but no observation set *precisely* 24 hours prior to the last observation set, the observation set from 28 hours before the last observation set was included. Due to this, the number of hours per patient in $\mathcal{CH}_2^{NR}$ may be greater than 24 for some patients. This time period is generally referred to in this chapter as the *last-day* period. The mean duration of the last-day period for $\mathcal{CH}_2^{NR}$ is 28.3 hours, and there are 156 patients in this group.

Dataset $\mathcal{CH}_2^{EICU}$ This contains the 19 patients belonging to the EICU group defined in Chapter 7. For a patient in this group, vital-sign observation sets recorded in the last-day period prior to an EICU admission were included. For patients with multiple emergency ICU admissions, or an EICU admission followed by death at some point after the EICU episode, only the first emergency ICU admission was included. For patients who had an emergency ICU admission within 24 hours or less of entering the upper GI ward, this entire period (which was less than 24 hours) was included. The mean duration of the last-day period for $\mathcal{CH}_2^{EICU}$ is 24.4 hours. For the three patients in this group who had an emergency ICU admission in less than 24 hours of their arrival in the ward, their lengths of stay on the ward were 14.9 hours, 13.4 hours, and 16.3 hours.

Ideally, if there were a much larger number of patients in the EICU group, then these latter three patients would not be included in $\mathcal{CH}_2^{EICU}$ since instability in vital signs is expected during the first 24 hours of arriving on the ward in both NR and EICU patients. However, doing so would reduce the number of patients in $\mathcal{CH}_2^{EICU}$, and consequently the number of patients to test the model on, to only 16. Therefore it was decided to retain these three patients, and where possible in the results, indicate them separately.

Dataset $\mathcal{CH}_2^{NR_{test}}$ As there are 19 patients in $\mathcal{CH}_2^{EICU}$, in order to test models on the same number patients from the NR group, 19 randomly selected patients from dataset

$\mathcal{CH}_2^{NR}$ were included in dataset $\mathcal{CH}_2^{NR_{test}}$.

Dataset $\mathcal{CH}_2^{NR_{train}}$ This dataset includes all 137 patients from $\mathcal{CH}_2^{NR}$ that are not included in $\mathcal{CH}_2^{NR_{test}}$.

Since $\mathcal{CH}_2^{NR_{train}}$ and $\mathcal{CH}_2^{NR_{test}}$ contain a random selection of patients, any experiment in which these datasets are used will be repeated 50 times, each time with a different random selection of patients, and the results will be averaged over the 50 iterations.

The distribution of vital signs in these datasets is shown in Figure 8.1. As there are many more vital-sign observations in $\mathcal{CH}_2^{NR_{train}}$ than there are in $\mathcal{CH}_2^{NR_{test}}$ and $\mathcal{CH}_2^{EICU}$, for comparison of the datasets, the histograms are normalised. From the distributions it may be seen that the mode of the BR for $\mathcal{CH}_2^{EICU}$ takes a higher value ($\sim$ 20 rpm) compared to that of the $\mathcal{CH}_2^{NR_{train}}$ and $\mathcal{CH}_2^{NR_{test}}$. Further the HR distribution of $\mathcal{CH}_2^{EICU}$ is bi-modal with the higher mode at 110 bpm, which does not exist in $\mathcal{CH}_2^{NR_{train}}$ and $\mathcal{CH}_2^{NR_{test}}$. A lower SBP mode may also be seen for the $\mathcal{CH}_2^{EICU}$, as well as a slight rise in the temperature with a mode at 36.6°C compared with modes of $\mathcal{CH}_2^{NR_{train}}$ and $\mathcal{CH}_2^{NR_{test}}$ being at 36.2°C, indicating that $\mathcal{CH}_2^{EICU}$ patients may be slightly pyrexic (i.e. have a high temperature).

8.4 Parzen Windows Model

As done for dataset $\mathcal{D}_1$ earlier in this thesis, a Parzen windows model was constructed for $\mathcal{CH}_2$, first for the 2-D dataset comprising only HR and BR, and then for the full 5-D dataset. Construction of a density-based model of normality using Parzen windows (PW) is described in detail in section 3.5.3. Here the Parzen windows model is constructed using $\mathcal{CH}_2^{NR_{train}}$, where the number of vital-sign dimensions, d=2 for the 2-D dataset corresponding to the HR and the BR, and d=5 for the 5-D dataset which also contains SpO_2, SBP, and temperature. Recalling Equation 3.5, the Parzen windows model can be described using

$$p(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sigma^d}\frac{1}{(2\pi)^{d/2}}\exp\left\{-\frac{|\mathbf{x}-\boldsymbol{\mu}_i|^2}{2\sigma^2}\right\}. \tag{8.1}$$

As explained in Chapter 4, the probability of being away from the model of normality

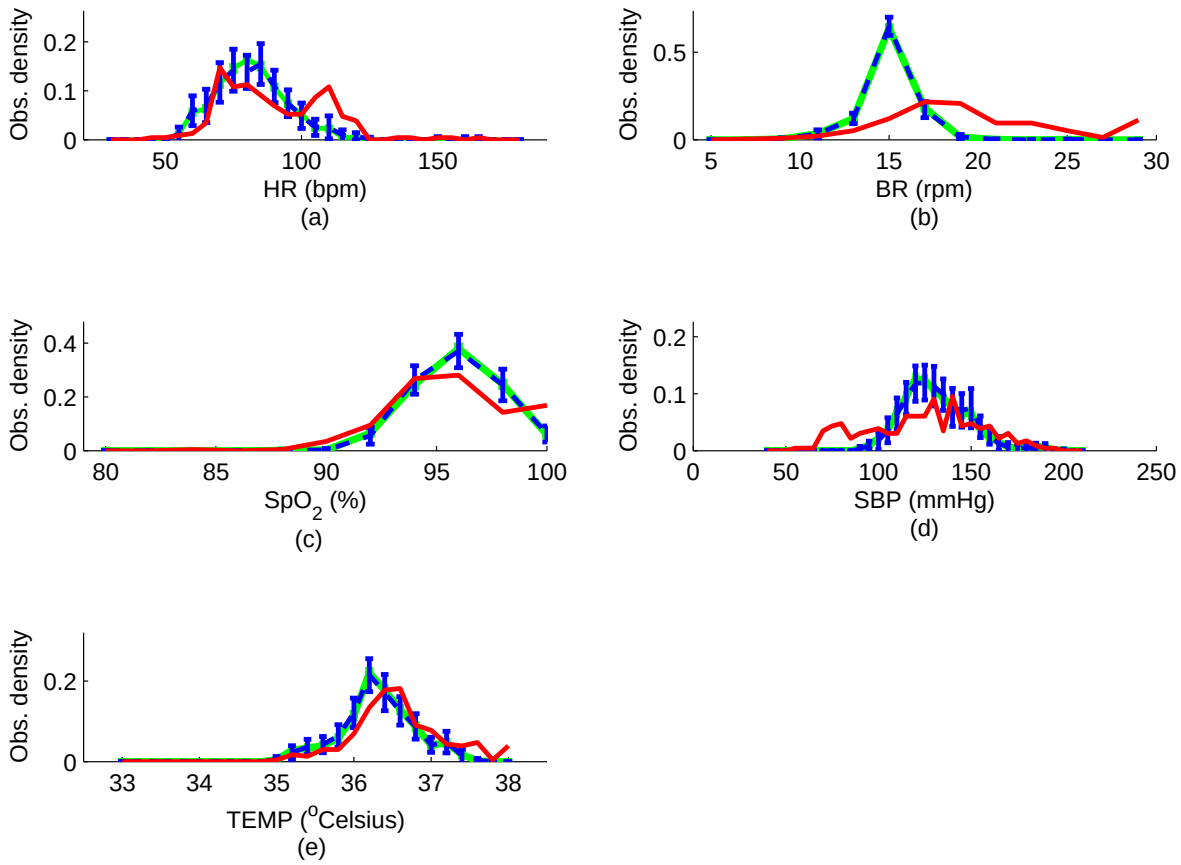


Figure 8.1: Normalised histograms for the (a) HR, (b) BR, (c) SpO_2, (d) SBP, and (e) temperature, for the datasets used in this chapter. The distribution for $\mathcal{CH}_2^{EICU}$ is shown in red. For the distributions of $\mathcal{CH}_2^{NR_{train}}$ and $\mathcal{CH}_2^{NR_{test}}$, the mean value after averaging over 50 experiments, is shown in blue and green, respectively. The mean and ± 1 standard error over the 50 experiments is shown by the blue and green vertical bars, for $\mathcal{CH}_2^{NR_{train}}$ and $\mathcal{CH}_2^{NR_{test}}$, respectively.

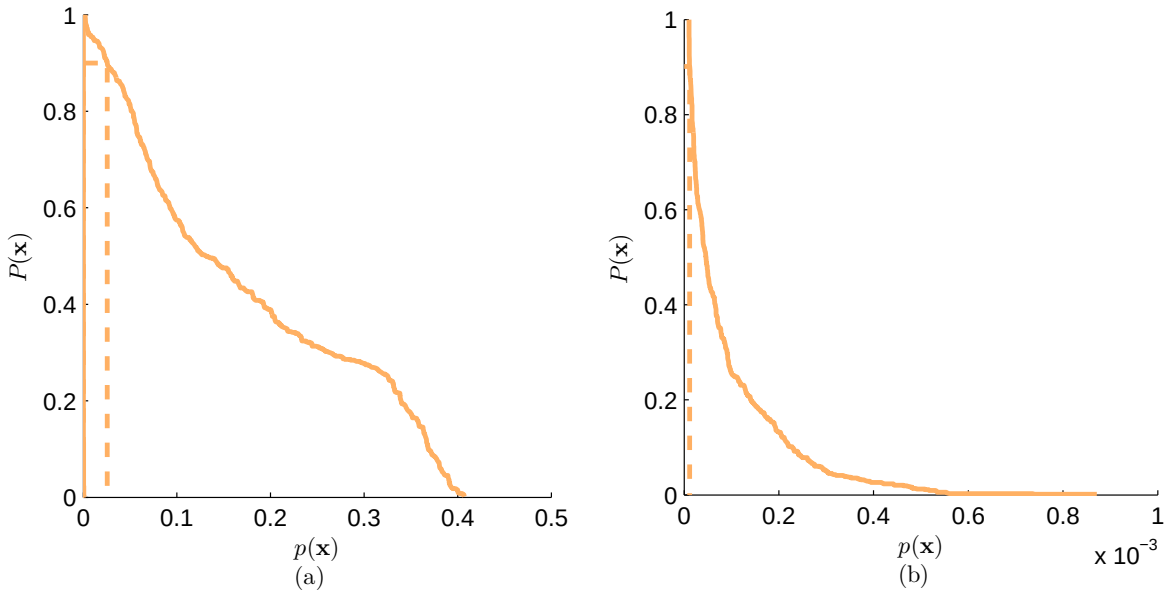


Figure 8.2: $P(\mathbf{x})$ plotted against the $p(\mathbf{x})$ for the Parzen window models of normality for the (a) 2-D, and (b) 5-D training set $\mathcal{CH}_2^{NR_{train}}$. In (a), $p(\mathbf{x}) = 0.026$ corresponding to $P(\mathbf{x})= 0.9$, and in (b) $p(\mathbf{x}) = 1.1 \times 10^{-5}$ corresponding to $P(\mathbf{x})= 0.9$ are shown by a dashed line.

(or the probability of novelty) is then defined as:

$$P(\mathbf{x}) = 1 - \tilde{P}(\mathbf{x}) \tag{8.2}$$

An alert is generated when, for a single observation set $\mathbf{x}$, $P(\mathbf{x})\geq T$, where T is a threshold on $P(.)$. Figure 8.2 shows the relationship between $p(\mathbf{x})$, and $P(\mathbf{x})$ for the models constructed using the 2-D and the 5-D datasets. The example threshold of $T = P(\mathbf{x}) = 0.9$ is also shown in Figure 8.2.

8.4.1 Parzen Windows Results

Figures 8.3(a) and (b) show the results obtained when assessing the 2-D and 5-D models, respectively, using the r_1 metric defined in Section 8.2. Both 2-D and 5-D models were tested over the complete range of values of $0.1 \leq T\leq 0.99$, with a step size of 0.01. Model performance is discussed at the middle of the threshold range (i.e. at $T = 0.5$) and at $T = 0.99$.

Model performance is similar for the 2-D and 5-D datasets; at $T = 0.5$, $r_1 = 100\%$ for $\mathcal{CH}_2^{EICU}$ both 2-D and 5-D models, and similarly the corresponding r_1 for the $\mathcal{CH}_2^{NR_{test}}$ is 87% for both 2-D and 5-D models, which indicates that the models have a false-alert

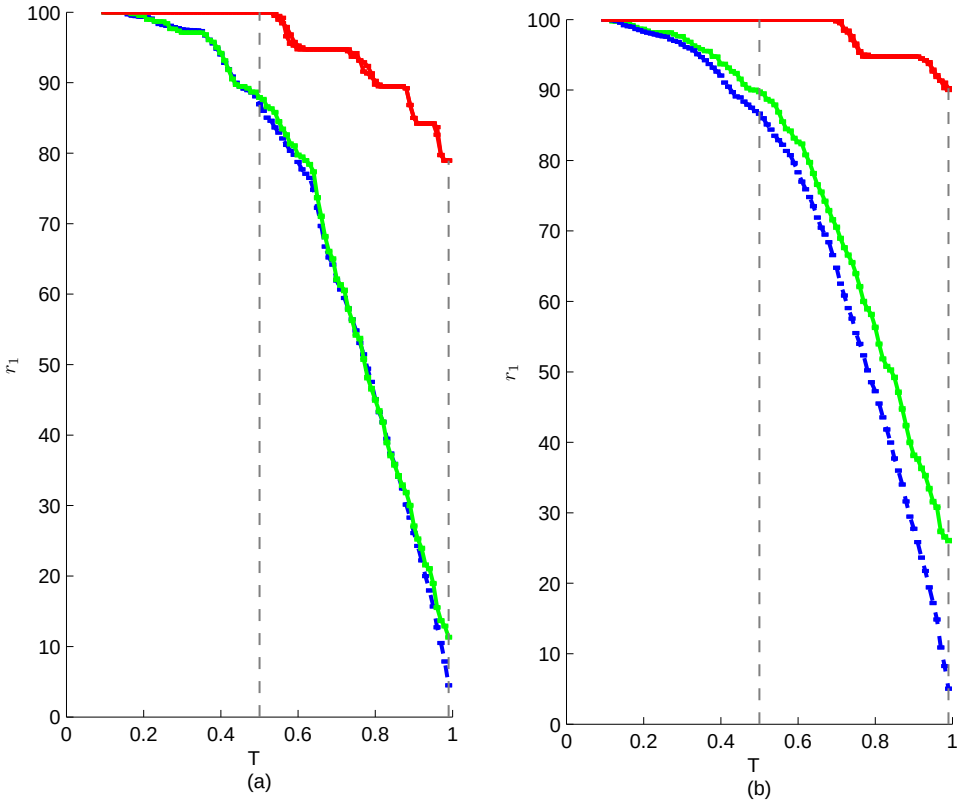


Figure 8.3: r_1 (%) for the (a) 2-D, and (b) 5-D models. r_1 over $0.1 \leq T \leq 0.99$, for the test sets $\mathcal{CH}_2^{NR_{test}}$ and $\mathcal{CH}_2^{EICU}$ is shown in green and red, respectively. r_1 for the $\mathcal{CH}_2^{NR_{train}}$ is also shown in the background in blue. Grey dashed lines indicate results at $T = 0.5$ and $T = 0.99$.

rate of 87% at $T = 0.5$.

At $T = 0.99$, $r_1 = 79.7\%$ for $\mathcal{CH}_2^{EICU}$ using the 2-D model, and $r_1 = 90.2\%$ using the 5-D model, indicating that model sensitivity increases as T is increased, and further that the sensitivity of the 5-D model is higher than that of the 2-D model.

For $\mathcal{CH}_2^{NR_{test}}$, at $T = 0.99$, r_1 drops to 11.3% for the 2-D model. However for the 5-D model, it only drops to 26.0% , indicating that while the 5-D model has a higher sensitivity than the 2-D model, it also has a lower specificity.

These results indicate that on average, considering the 5-D model, the Parzen windows model generates alerts for 17 out of the 19 (i.e. 90%) patients in the $\mathcal{CH}_2^{EICU}$ group, but that it also generates false alerts for 5 out of 19 (i.e. 26%) of the $\mathcal{CH}_2^{NR_{test}}$ patients. As a background check to assess the performance on a larger number of patients, it may be seen from the blue curve that if tested on the 137 patients in the training set $\mathcal{CH}_2^{NR_{training}}$, the 5-D model has a false alert rate as low as 5%, i.e. approximately 7 out of 137 patients.

8.5 Motivation for Patient-based Model of Normality

In dataset $\mathcal{D}_1$, the vital-sign data in the training set were collected at a sampling rate of 0.05 Hz, i.e. one sample every 20 seconds. 18,000 hours of data were collected from a total of 326 patients, with a total number of samples in the training set of 2×10^6 (see Table 8.2). This gives an average of 54 hours per patient, and an average of 9,720 samples per patient. Thus the resulting Parzen windows based probability density function obtained using the training set could be expected to be an accurate representation of the distribution of data for all patients. In comparison, the average number of vital-sign observations on the last day in the $\mathcal{CH}^{NR}$ dataset is 5.3 ± 0.9 observations per patient, i.e. the data are relatively "under-sampled". (The mean number of samples per patient for the $\mathcal{CH}_2^{EICU}$ dataset is 12.2 ± 6.1, and the sampling rate for a patient who is deteriorating is higher in the last 24 hours prior to an event). In the previous section, it was shown that the Parzen windows model alerts in almost a third of the patients who did not have an event. An alternative modelling approach is now considered that takes into account the under-sampled nature of the data recorded from each individual patient. As with Parzen windows, this alternative approach also performs novelty detection with respect to a model of normality, but here

the model of normality contains *patient-based hypotheses* (explained in the next section). It is therefore refered to as a patient-based model.

In order to construct the patient-based model, the form of the distribution of an individual patient's vital signs during a 24-hour period was investigated. As discussed, the track-and-trigger data are too under-sampled to obtain an accurate representation of the vital-sign distribution. However, for some patients in the study who had an intermediate stay in the high dependency unit (HDU) before being transferred to the upper GI ward, some vital signs were recorded continuously, with a sampling frequency of approximately 4 seconds[2]. Therefore the 24-hour distribution of continuously-recorded HR, BR, SpO_2, and SBP was considered, for each patient from the NR group for whom continuous data were recorded for at least 24 hours during their HDU stay. Temperature data could not be recorded continuously. Here it is recognised that most of these patients were still in their recovery phase while continuous recording was being carried out, however it is considered that this continuously acquired dataset is the best available record to investigate the distribution of the vital signs of a patient. HR was recorded for a minimum of 24 hours for 33 patients in the NR group, BR for 32 patients, and SpO_2 and SBP for 34 patients. Figure 8.5 shows the daily distribution over the last 24 hours of continuous monitoring for each patient (blue curves), as well the average of all patients (pink curve), which appear to be Gaussian (or Normal) with the exception of the SpO_2 distribution. A chi-squared test [3] was performed to test the gaussianity of the HR, BR, and SBP distributions of each individual patient. The null hypothesis, which was that the vital-sign data are a random sample from a Normal distribution, was accepted for the HR in 81% of the patients, for the BR in 82%, and the for the SBP in 80% of the patients. Based on these results, it was considered that a Gaussian (or Normal) form can be assumed for the distribution of vital-sign data for an individual patient. Hence in the following section, a Normal distribution is considered in order to construct a patient-based model. Here it is noted that that the Normal assumption is not valid for the SpO_2, however in the first instance this

[2]Following surgery, most patients go directly to the upper GI ward after their surgery; a second group goes first to the ICU for a scheduled admission, while a third group goes to the HDU, where they are monitored more closely via continuous bed-side monitoring (using Phillips IntelliVue MP50 monitors), before being transferred to the upper GI ward.

[3]In the chi-squared ($\mathcal{X}^2$) test, the number of observed values in a histogram bin are compared with some expected value. If the p-value (i.e. the probability of obtaining a value as extreme as the one observed) is less than the 5% significance level, the null hypothesis is rejected.

Dataset	Hours	No. Patients	No. Labelled events	Sampling rate
$\mathcal{D}_1$	18,000	326	112	0.05 Hz
$\mathcal{CH}_2^{NR}$	3,876	156	–	5.3 samples/patient
$\mathcal{CH}_2^{ETCU}$	464	19	–	variable

Table 8.2: A comparison of datasets belonging to the UPMC and CALMS-2 studies. The second column shows the total number of hours for which vital signs were recorded. No. patients refers to the number of patients in the dataset, and no. labelled events refers to the total number of labelled events. Events were not labelled in the CALMS-2 study, indicated by –.

variable will be treated in the same way as the others (including temperature, for which no continuous data were available to test). In the final chapter, models considering other forms of distribution for the SpO_2 variable, and indeed for those patients for whom the distribution of vital signs is not Normal, are discussed.

8.6 Patient-Based Modelling

In the patient-based modelling approach, a model of normality that contains "patient-based" hypotheses is considered here. Hypothesis H_i is defined here such that H_i holds true if some data $\mathbf{x}$ are generated from the distribution of data for patient Pt_i. To elaborate, here an assumption is made that a patient's physiology may be represented by the distribution of the patient's vital signs, and that the vital signs are distributed with Normal form, such that for the i^{th} patient Pt_i, the distribution of some d-dimensional vital-sign vector $\mathbf{x}$ may be expressed as

$$\mathcal{N}((\mathbf{x}\,|\Theta_i, H_i) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|}\exp\left\{-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_i\right)'\left(\mathbf{x}-\boldsymbol{\mu}_i\right)\right\} \tag{8.3}$$

where $\Theta_i = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$, $\boldsymbol{\mu}_i$ is the mean vector and $\boldsymbol{\Sigma}_i$ is the covariance matrix for the Normal distribution representing patient H_i.

Let a hypothetical situation be considered where there are $I = 100$ such patients with "normal" physiology, and the goal is to determine how similar a new, previously-unseen test patient's physiology is to that of these I "normal" patients. Figure 8.4 presents a schematic illustration of this concept. The hypothetical boundary placed around the model of normality represents some threshold such that if any test data lie within the boundary of normality with respect to the threshold (as shown by the dotted black line

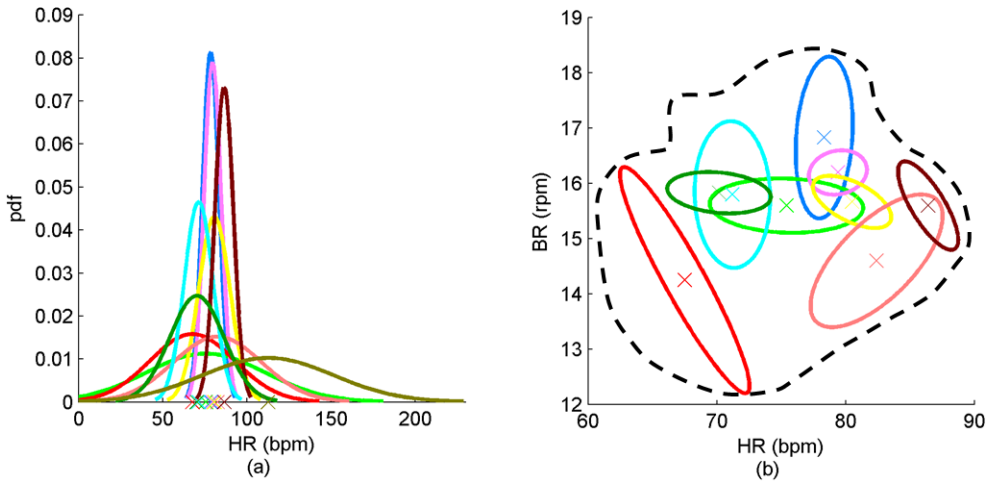


Figure 8.4: A schematic diagram illustrating the concept of patient-based modelling. (a) a hypothetical uni-dimensional model. A single curve shows the uni-dimensional Normal distribution $\mathcal{N}(\mu_i, \sigma_i^2)$ for the HR data belonging to a hypothetical patient H_i. Ten curves are shown for ten hypothetical patients, with μ_i of a patient shown by a cross in the corresponding colour. (b) A single ellipse shows the bi-dimensional ($d = 2$) Normal distribution $p(\mathbf{x}) = N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ belonging to a hypothetical patient H_i. $\boldsymbol{\mu}_i$ is shown by a cross of the same colour as the ellipse. In (b), the hypothetical "model of normality" is bounded by the threshold shown in dashed black line.

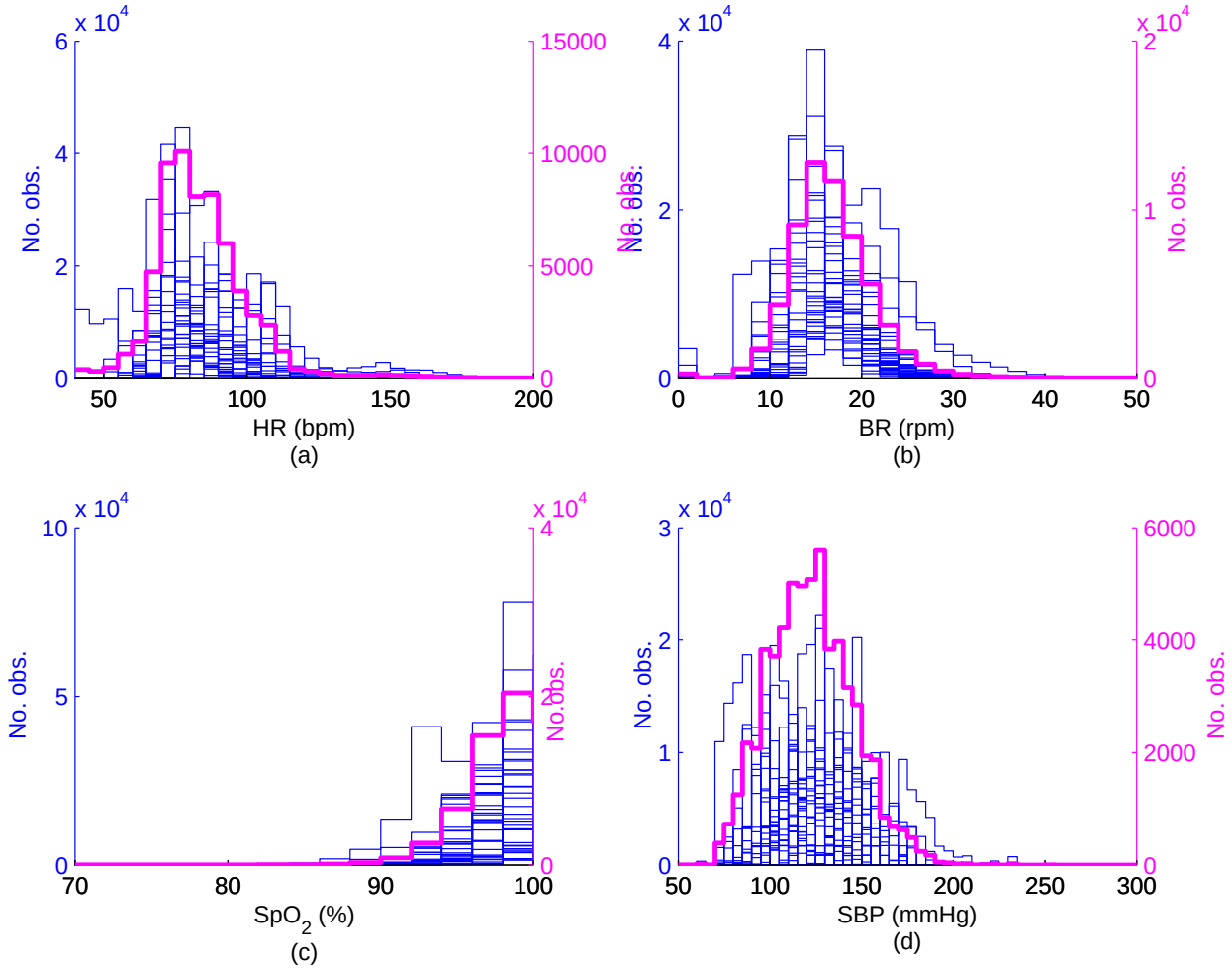


Figure 8.5: Histograms for (a) HR, (b) BR, (c) SpO_2, and (d) SBP. A blue curve represents the histogram distribution of data belonging to an NR patient, where the data represent a 24-hour period of continuous vital-sign recording. The pink curve shows the average distribution over all patients who had continuous recordings for the corresponding vital sign.

in Figure 8.4 (b)), they are considered as being "normal" and "abnormal" otherwise.

One way to build such a model could be to consider each patient as a "component" of a density-based model such as Parzen windows, and estimate the pdf of the model by placing a Normal kernel around the mean of the distribution of the patient's data. According to this approach, a patient's physiology would be represented by the mean $\boldsymbol{\mu}_i$ and variance $\boldsymbol{\Sigma}_i$ of a Normal distribution fitted to this patient's vital-sign data, $\mathcal{N}(\mathbf{x}\,|\Theta_i, H_i)$, using Equation (8.3). Thus here one is choosing to *summarise* the patient's physiology in the form of the point-estimates of the two variables $\Theta_i = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$. In doing so, it is being assumed that the point-estimates of Θ_i are sufficiently accurate, in that they represent the underlying Normal distribution from which $\mathbf{x}$ is presumably generated. Point-estimates of Θ_i are often deemed to be appropriate in cases for which there are sufficient examples of $\mathbf{x}$ (in keeping with the central limit theorem, for $N \to \infty$); i.e., when the data vector $X = \mathbf{x}_1 \ldots \mathbf{x}_N$ is large. However, in $\mathcal{CH}_2^{NR_{train}}$, there are only 5.3 samples per patient on average. While a Normal distribution may in theory be fitted to such a dataset, the point-estimate of Θ_i is largely uncertain, in that many values of Θ_i may be possible given the small number of observed samples.

In section 8.6.1 a Bayesian method is described which considers Θ_i not to be a point-estimate, but rather a random variable over which a distribution, such as a Normal distribution or a Wishart distribution, may be defined. In the following section, it is explained how a Bayesian modelling approach takes into account uncertainty in the estimation of Θ_i in a principled manner, and how this is suited to modelling the data in dataset $\mathcal{CH}_2^{NR_{train}}$, in which there is a very limited number of data per patient available for modelling.

8.6.1 Bayesian Inference

According to Bayes' theorem, considering hypothesis H_i to be true,

$$p(\Theta_i \mid \mathbf{x}_1, H_i) = p(\mathbf{x}_1|\ \Theta_i, H_i)p(\Theta_i\ |H_i) \tag{8.4}$$

where the prior distribution, $p(\Theta_i \mid H_i)$, is the "guess" or the prior belief about the parameters Θ_i, before observing any data X. The conditional distribution, $p(\mathbf{x}_1 \mid \Theta_i, H_i)$, is the probability distribution of the data $X = \mathbf{x}_1 \ldots \mathbf{x}_N$ given the hypothesis H_i, upon the

arrival of the first sample of data $\mathbf{x}_1$. The posterior probability distribution, $p(\Theta_i|\mathbf{x}_1, H_i)$, is the "updated" estimate of the belief after observing sample $\mathbf{x}_1$.

When the next sample $\mathbf{x}_2$ arrives, the new prior distribution is the previous posterior distribution $p(\Theta_i \mid \mathbf{x}_1, H_i)$, the conditional distribution is $p(\mathbf{x}_2 \mid \Theta_i, H_i)$, and the new posterior distribution is $p(\Theta_i \mid \mathbf{x}_1, \mathbf{x}_2, H_i)$. The relationship between the prior and the posterior distributions now looks like

$$p(\Theta_i \mid \mathbf{x}_1, \mathbf{x_2}, H_i) = p(\mathbf{x}_2 \mid \Theta_i, H_i) p(\Theta_i \mid \mathbf{x}_1, H_i) \tag{8.5}$$

This process, known as recursive Bayesian inference, can be repeated iteratively as more data arrive, and can be generalised as

$$p(\Theta_i \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, H_i) \propto p(\mathbf{x}_N \mid \Theta_i, H_i) p(\Theta_i \mid \mathbf{x}_1, \ldots, \mathbf{x}_{N-1}, H_i) \tag{8.6}$$

According to Equation 8.6, the posterior and prior distributions must be of the same form in order for the posterior at the $(N-1)^{th}$ iteration to be used as the prior in the N^{th} iteration. A prior distribution that has the same form as that of the posterior distribution is known as the *conjugate prior*, and together the prior and posterior distributions are known as the *conjugate pair* [112]. Conjugate pairs are used in Section 8.6.2.

Recursive Bayesian inference allows one to learn the posterior distribution of parameters Θ_i of a hypothesis H_i in a manner that lets the incoming data update one's initial "uncertain" belief about the values Θ_i. This is illustrated with an example in the next section.

8.6.2 Coping with Uncertainty in Parameter Estimation

For the purpose of explanation, let us assume that an artificial uni-dimensional dataset $X = \mathrm{x}_1 \ldots \mathrm{x}_N$ belongs to an underlying Normal distribution, where $N = 5$. A non-Bayesian will find point-estimates of $\Theta = \{\mu, \sigma^2\}$ from the $N = 5$ samples. For uni-dimensional data, this can be performed using the maximum likelihood estimation (MLE) approach, which results in the standard formulae

$$\mu = \frac{1}{N}\sum_{n=1}^{N} \mathrm{x}_n \tag{8.7}$$

and

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N} \left(\mathrm{x}_n - \mu\right)^2 . \tag{8.8}$$

From Equation (8.7) it can be seen that μ is simply the arithmetic mean of the five samples. While the arithmetic mean of a large number of samples may be expected to converge to some expected value, according to the the central limit theorem, the same may not necessarily be true for a small sample size, such as when $N = 5$. In the latter case, the point-estimates of μ and σ^2 may or may not be accurately representative of the underlying distribution to which the samples belong, as shown in [33]. In the Bayesian approach, the "uncertainty" in the estimation of the parameters Θ of a distribution is taken into account explicitly.

Consider

$$p\left(\mu\right) \sim \mathcal{N}\left(\mu \mid \mu_0, (\kappa_0\lambda)^{-1}\right) \tag{8.9}$$

and

$$p\left(\lambda\right) \sim \mathcal{G}\left(\lambda \mid \alpha_0, \beta_0\right) \tag{8.10}$$

where μ_0 and $(\kappa_0\lambda)^{-1}$ are the mean and the variance of the distribution of μ which takes the form of a Normal distribution, and κ_0 is the hyper-parameter which governs the dependence of μ on λ (where $\lambda = \frac{1}{\sigma^2}$, used here instead of σ^2 for ease of computation). Hyper-parameters α_0 and β_0 are the shape and scale, respectively, of the distribution of λ, which takes the form of a Gamma distribution. Thus combining Equations 8.9 and 8.10 the Normal-Gamma ($\mathcal{NG}$) form of the prior distribution $p(\Theta_i \mid H_i)$ is expressed as

$$p\left(\mu, \lambda\right) \sim \mathcal{N}\left(\mu \mid \mu_0, (\kappa_0\lambda)^{-1}\right) \mathcal{G}\left(\lambda \mid \alpha_0, \beta_0\right) \tag{8.11}$$

and the Normal-Gamma ($\mathcal{NG}$) form of the posterior distribution $p(\Theta_i \mid \mathbf{x}_N, H_i)$ as

$$p(\mu, \lambda|\mathbf{x}_N)\ \mathcal{N}(\mu|\mu_N, \frac{1}{\kappa_N \lambda})\mathcal{G}(\lambda|\alpha_N, \beta_N) \tag{8.12}$$

where the dependence on H_i in the left- and right-hand sides of Equations 8.11 and 8.12 are temporarily removed given that we are discussing only one hypothesis at a time.

The conditional distribution $p(\mathbf{x}_N \mid \Theta_i)$ takes Normal form, since under hypothesis H_i we assume $\mathbf{x}_N$ to be generated from a multi-dimensional Normal distribution,

$$p(\mathbf{x}_N \mid \mu, \lambda) = \frac{1}{(2\pi)^{d/2}} \lambda^{N/2} \exp\left(-\frac{\lambda}{2} \sum_{n=1}^{N} (\mathbf{x}_N - \boldsymbol{\mu}_n)^2\right) \tag{8.13}$$

The relationship between the $\mathcal{NG}$ conjugate prior distribution $p(\mu, \lambda)$ and the posterior distribution $p(\mu, \lambda|\mathbf{x}_N)$, and the hyper-parameters α and β is described in Appendix A.

Figure 8.6 illustrates how recursive Bayesian inference is used to learn $p(\Theta_i \mid \mathbf{x}_N, H_i)$ for an artificial dataset $\mathbf{x} \in X$, which is generated from a Normal distribution, with $p(X) = \mathcal{N}(0, 1)$. In Figure 8.6(a), $N = 0$, and the prior distribution $p(\Theta_i \mid H_i)$ is shown. Figure 8.6(b) shows the posterior distribution $p(\Theta_i \mid \mathbf{x}_N, H_i)$ after observing $N = 5$ samples from X. The coloured circles indicate four example values of $\Theta_i = \{\mu_i, \lambda_i\}$. Figure 8.7 shows the corresponding distributions $p(\mathbf{x}_N \mid \Theta_i, H_i)$ at each of the example $\Theta_i = \{\mu_i, \lambda_i\}$ which are shown in 8.6(b). It may be seen that after each iteration of Equation (8.6), the $\mathcal{NG}$ distribution of the parameters becomes more compact in support. In fact, for $N \gg 0$, μ and λ are expected to converge to the MLE-estimated μ_{MLE} (defined in Equation 8.7) and $\lambda_{MLE} = \frac{1}{\sigma^2_{MLE}}$ (defined in Equation 8.8), as shown in Figure 8.6.

Recall that the average number of samples in the last 24 hours of a patient from the normal-recovery group is 5.3 samples per patient. Hence if it was wished to fit a Normal distribution to data from a single patient, the Bayesian approach discussed in this section would require a $\mathcal{NG}$ distribution describing the parameters of the Normal distribution responsible for generating the data. Supposing for the sake of explanation that the artificial dataset X belongs to a patient from the training set, then the $\mathcal{NG}$ distribution over the parameters of the Normal generating distribution for this patient's data would look like Figure 8.6(b). A non-Bayesian estimator would represent this as a

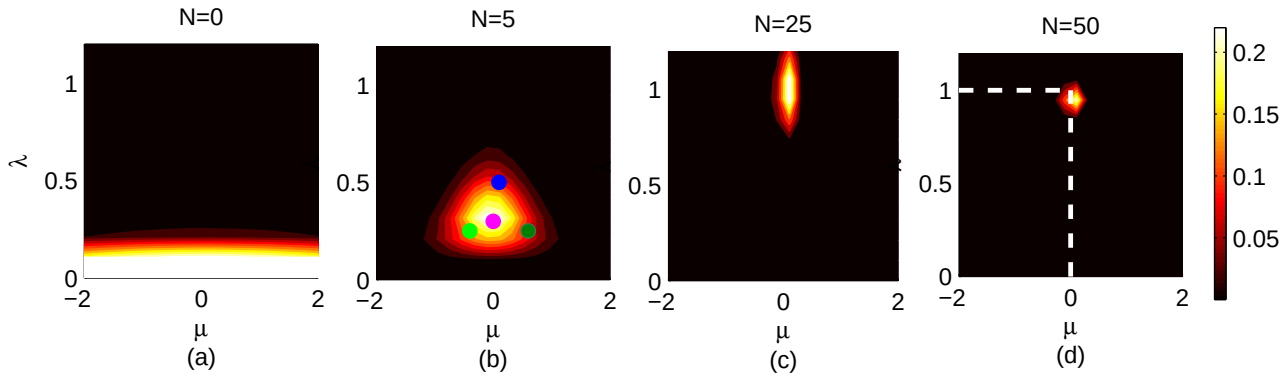


Figure 8.6: An example of learning the $\mathcal{NG}$ distribution over the parameters $\Theta_i = \{\mu_i, \lambda_i\}$ using recursive Bayesian inference. μ_i is shown on the x-axis, and λ_i is shown on the y-axis. (a) The prior probability distribution $p(\Theta_i \mid H_i)$, i.e. the distribution of the parameters Θ_i prior to the arrival of any data, where $N = 0$. (b) to (d) show the posterior distribution $p(\Theta_i \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, H_i)$ upon the arrival of data, for $N = 5$, $N = 25$, and $N = 50$, respectively. The artificial data $\mathbf{x} \in X$, and $p(X) = \mathcal{N}(0, 1)$. In (b), four sample points on $p(\Theta_i \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, H_i)$ are shown by the coloured dots, where each dot represents some $\Theta_i = \{\mu_i, \lambda_i\}$. In (d), MLE point-estimates $\mu_{MLE} = 0$ and $\lambda_{MLE} = 1$ are marked by dashed lines in white. The colour bar on the right indicates the value of the $p(\Theta_i \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, H_i)$ for the corresponding colour in the bar.

point-estimate which would be a point on this figure such as that indicated on Figure 8.6(d). However, as shown by the circles in Figure 8.6(b), $\Theta_i = \{\mu_i, \lambda_i\}$ can take many different values, which can result in many $p(\mathbf{x}_N \mid \Theta_i, H_i)$ as shown in Figure 8.7. This highlights the main difference between the Bayesian and non-Bayesian approaches in the context of taking into account uncertainty in parameter estimation.

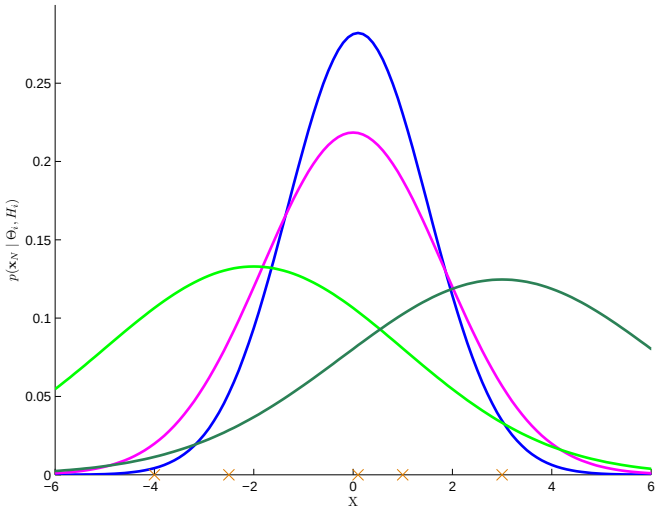


Figure 8.7: Distributions representing $p(\mathbf{x}_N \mid \Theta_i, H_i)$ corresponding to the four samples of $\Theta_i = \{\mu_i, \lambda_i\}$ shown by the coloured dots in Figure 8.6 at $N = 5$. The five samples of $\mathbf{x} \in X$ are shown by crosses in orange on the x-axis.

8.6.3 Evidence of a Hypothesis

The evidence $P(X \mid H_i)$ for hypothesis H_i is defined as

$$P(X \mid H_i) = \int p(X \mid \Theta_i, H_i) p(\Theta_i \mid H_i) d\Theta_i \tag{8.14}$$

and is in effect the probability that hypothesis H_i describes the observed data, after seeing dataset X [99]. It is also known as the marginal likelihood or the integrated likelihood of hypothesis H_i [11]. As the evidence of a hypothesis is obtained after integrating over Θ_i, it encompasses the prior uncertainty about Θ_i.

Using Bayes theorem, given that $P(H_i)$ is the prior probability of hypothesis H_i being that which is responsible for generating the observed data, the posterior probability $P(H_i \mid X)$ of hypothesis H_i is defined as

$$P(H_i \mid X) = \frac{P(X \mid H_i) P(H_i)}{\sum_{i=1}^{I} P(X \mid H_i) P(H_i)} \tag{8.15}$$

Here, the term in the denominator may equivalently be written as $P(X)$, which is the overall probability of dataset X given all possible hypotheses, and is a constant. Thus Equation (8.15) reduces to

$$P(H_i \mid X) \propto P(X \mid H_i) P(H_i) \tag{8.16}$$

Furthermore, if it is assumed that all hypotheses under consideration, $H_{i=1\ldots I}$, have equal prior probability $P(H_i)$, then $P(H_i) = \frac{1}{I}$, so that effectively, Equation 8.16 reduces to

$$P(H_i \mid X) \propto P(X \mid H_i) \tag{8.17}$$

which means that the posterior probability of a hypothesis after observing data is directly proportional to the evidence of that hypothesis after observing the data. From this it follows that competing hypotheses may be compared by evaluating their evidence [99].

Thus, evidence can be used to rank candidate hypotheses; for this reason it is often used in model selection applications [63]. In the next section it is shown how the evidence

can be used to combine hypotheses based on the concept of *Bayesian model averaging*. This will allow for the construction of a Bayesian model of normality for novelty detection.

8.6.4 Constructing a Bayesian Model of Normality

Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ represent the data from patient Pt_i in the training set (where the terms patient and hypothesis will be used interchangeably in this section, according to the context). Using the recursive Bayesian approach described in the previous section, one may learn the posterior distribution over the parameters of the generating distribution of data observed from patient Pt_i, recalling Equation (8.6):

$$p(\Theta_i \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, H_i) \propto p(\mathbf{x}_N \mid \Theta_i, H_i) p(\Theta_i \mid \mathbf{x}_1, \ldots, \mathbf{x}_{N-1}, H_i) \tag{8.18}$$

Here $p(\Theta_i \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, H_i)$ is refered to as the distribution over the parameters for hypothesis H_i. A patient-based model is then defined such that it contains hypotheses $H_{i=1\ldots137}$. $I = 137$ as there are a total of 137 patients in the training set (see Table 8.1). As shown in Equation 8.17, the probability of hypothesis H_i after observing the training data is $P(H_i \mid X)$, which is the weight assigned to that hypothesis in the model.

The posterior probability of H_i upon seeing a new test sample, $\mathbf{x}_{new}$, may then be expressed as $P(H_i \mid \mathbf{x}_{new})$, which is proportional to the evidence of the hypothesis H_i upon seeing a new test sample, $P(\mathbf{x}_{new} \mid H_i)$. That means, recalling Equation 8.17, $P(H_i \mid \mathbf{x}_{new}) \propto P(\mathbf{x}_{new} \mid H_i)$, so that, considering all 137 hypotheses,

$$P(\mathbf{x}_{new} \mid H_{i\ldots I}) = \frac{1}{I} \sum_{i=1}^{I} P(H_i \mid \mathbf{x}_{new}). \tag{8.19}$$

This summation of $P(H_i \mid \mathbf{x}_{new})$ for hypotheses $H_{i\ldots I}$ is based on the concept of Bayesian model averaging (BMA) [74], in which the premise is that a weighted average of candidate hypothesis incorporates the uncertainty about individual hypotheses [122]. In the context of patient monitoring, it addresses the limitation posed by having available only a small number of training data per patient. $P(\mathbf{x}_{new} \mid H_{i\ldots I})$ is therefore an estimate of how probable $\mathbf{x}_{new}$ is with respect to the hypotheses included in the patient-based model. For simplicity, $P(\mathbf{x}_{new} \mid H_{i\ldots I})$ is henceforth referred to as $P(\mathbf{x}_{new})$.

Note that, since the posterior probability $P(H_i \mid X)$ assigned to each hypothesis is inversely proportional to the number of hypotheses included in the model,I, the resulting $P(\mathbf{x}_{new})$ is expected to take very low values. A novelty score $z(\mathbf{x}_{new})$ may therefore be computed.

$$z(\mathbf{x}_{new}) = \log_e \frac{1}{P(\mathbf{x}_{new})} = -\log_e P(\mathbf{x}_{new}) \tag{8.20}$$

such that, if $\mathbf{x}_{new}$ belongs to a patient whose physiology is similar to those patients with representative hypotheses within the patient-based model, the corresponding novelty score $z(\mathbf{x}_{new})$ is expected to be low, and if $\mathbf{x}_{new}$ belongs to a patient whose physiology is not similar to hypotheses in the patient-based model, the corresponding novelty score $z(\mathbf{x}_{new})$ will be high. Hence this method may be used to test the novelty of a test sample with respect to the patient-based model. Similarity to the patient-based model is determined by placing a threshold T on $z(\mathbf{x})$ such that if $z(\mathbf{x}_{new}) \geq T$, $\mathbf{x}_{new}$ is considered "abnormal" with respect to this patient-based model of normality, and "normal" otherwise.

Here it is important to clarify that the proposed method is based on the BMA approach, and is not an example of model combination. Indeed, in the limit $N \to \infty$, $\mu_N \to \mu_{MLE}$, and $\lambda_N \to \lambda_{MLE}$, and a single hypothesis H_i in the patient-based model would resemble a single component H_i in what could then be considered as a Gaussian mixture model. The latter is an example of model combination, but this is not the case for the problem discussed in this work, where the number of data observed per test patient is small, $N \ll \infty$.

Finally, here it is wished to clarify that while the proposed method is based on BMA, in the traditional sense BMA compares hypotheses under which the distribution of X take different forms [122], or different orders of model complexity [168]; for instance, a hypothesis under which the distribution of X takes a Normal form may be compared with a hypothesis under which the distribution of X takes the form a Cauchy distribution. In this work, only Normal distribution is considered over X, as it is assumed (based on Figure 8.5) that the vital-signs of a patient in the NR group can be modelled using a Normal distribution. As mentioned earlier in this thesis, this assumption is not best suited in the case of the non-Gaussian SpO_2 distribution, but since it upholds for the other vital

signs considered in this work, it is adopted as a first approach. Nevertheless, in the final chapter of this thesis, an extension of the proposed BMA-based model to include other distributions is discussed as part of future work. The proposed BMA-based model is henceforth referred to as a Bayesian Model of Normality (BMN).

8.7 A Model for Multi-dimensional Data

So far the BMN has been introduced using uni-dimensional data. In this section an extension of the same techniques to multi-dimensional data is presented. The Bayesian modelling process for multi-dimensional data is the same as in the uni-dimensional case. The form of the distributions taken by candidate models for multi-dimensional data is the multi-dimensional analogue of the Normal-Gamma ($\mathcal{NG}$) distribution, called the Normal-Wishart ($\mathcal{NW}$) distribution. The following analysis is presented for a bi-dimensional dataset, where the number of dimensions in the dataset, $d = 2$, and will be extended for $d = 5$ later in this section.

Let X be a bi-dimensional dataset generated from a Normal distribution, containing $n = 1...N$ bi-dimensional vectors, $X = [\mathbf{x}_1...\mathbf{x}_N]$, where the parameters $\Theta = \{\boldsymbol{\mu}, \Lambda\}$ of the $\mathcal{NW}$ distribution are the mean $\boldsymbol{\mu}$ =$[\mu_1\ \mu_2]$ and the precision matrix Λ=$\begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}$.

The closed form expressions for the $\mathcal{NW}$ conditional distribution $p(\mathbf{x}|\boldsymbol{\mu},\Lambda, H_i)$, and the conjugate prior $p(\boldsymbol{\mu},\Lambda \mid H_i)$, are as follows:

$$p(\mathbf{x}|\mu, \Lambda, H_i) = (2\pi)^{-(Nd)/2}|\Lambda|^{N/2} \exp\left(-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})\Lambda(\mathbf{x}_n - \boldsymbol{\mu})'\right) \tag{8.21}$$

$$p(\boldsymbol{\mu}, \Lambda|H_i) = \mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\kappa_0\Lambda)^{-1}\right)\mathcal{W}\left(\Lambda \mid T_0\right) \tag{8.22}$$

where $\mathcal{W}(.)$ is the Wishart distribution over the precision Λ, and the posterior $p(\boldsymbol{\mu}$, $\Lambda|\mathbf{x}, H_i)$ is

$$p(\mu, \Lambda \mid \mathbf{x}, H_i) = \mathcal{N}\left(\boldsymbol{\mu} \mid \boldsymbol{\mu}_N, (\kappa_N\Lambda)^{-1}, H_i\right)\mathcal{W}\left(\Lambda \mid T_N, H_i\right). \tag{8.23}$$

The derivation of T_N from T_0, which is the prior covariance, is shown in Appendix B.

8.7.1 Evaluating the Evidence Integral

Although the evidence integral shown in Equation 8.14 can be evaluated for very simple integrands, it is generally difficult to compute analytically, and approximations such as Laplace's approximation[4] [99] are often used. Indeed a major reason for the slow uptake of BMA since it was first proposed in the literature in the 1990's has been the challenge of evaluating the evidence integral [74]. In recent years, as computational power has increased, this limitation has been overcome through the use of Markov Chain Monte Carlo (MCMC) methods such as Gibb's sampling to generate samples from a probability distribution [14].

Here a distribution with five parameters (μ_1,μ_2,Λ_{11},Λ_{12},Λ_{22}) is being dealt with, so Laplace's approximation, which is suitable for a smaller parameter set, is no longer appropriate. It is therefore decided to resort to a numerical approximation of the evidence integral.

Inserting Equations 8.21 and 8.22 in to Equation 8.14,

$$P(\mathbf{x} \mid H_i) = \iint \underbrace{p(\mathbf{x}|\mu, \Lambda, H_i)}_{\text{likelihood}} \underbrace{p(\boldsymbol{\mu}, \Lambda|H_i)}_{\text{prior}} \mathrm{d}\boldsymbol{\mu}_i \mathrm{d}\Lambda_i \tag{8.24}$$

From Equation (8.24), it can be seen that the evidence of a given hypothesis can be approximated by computing the product of the likelihood (Equation 8.21) and the prior (Equation 8.22), for finite $\Theta_i = \{\boldsymbol{\mu}_i, \Lambda_i\}$. The task becomes one of generating sufficient samples of Θ_i from the distribution $p(\Theta_i \mid H_i)$, and computing $p(\mathbf{x} \mid \Theta_i, H_i)$ using Equation 8.24. For this, the WinBUGS [97] tool that implements Gibb's sampling to generate finite samples of Θ_i for multi-dimensional datasets was used.

The steps for estimating the evidence for a given hypothesis are summarised in Algorithm 8.1, and are explained as follows for the uni-dimensional case where $\Theta = \{\mu, \lambda\}$ i.e., μ and λ are the only two parameters. Later the same steps will be repeated for multi-dimensional data where Θ_i is a set of five parameters, $\Theta_i = \{\mu_1, \mu_2, \Lambda_{11}, \Lambda_{12}, \Lambda_{22}\}$.

[4]Laplace's approximation assumes that the evidence integral is peaked quite tightly around the most probable value(s) of parameters w, so that the evidence may be approximated by multiplying the height of the peak of the integral with the width of the integral $\sigma_{\Theta|X}$. The reader is referred to [99] for a full explanation of the method.

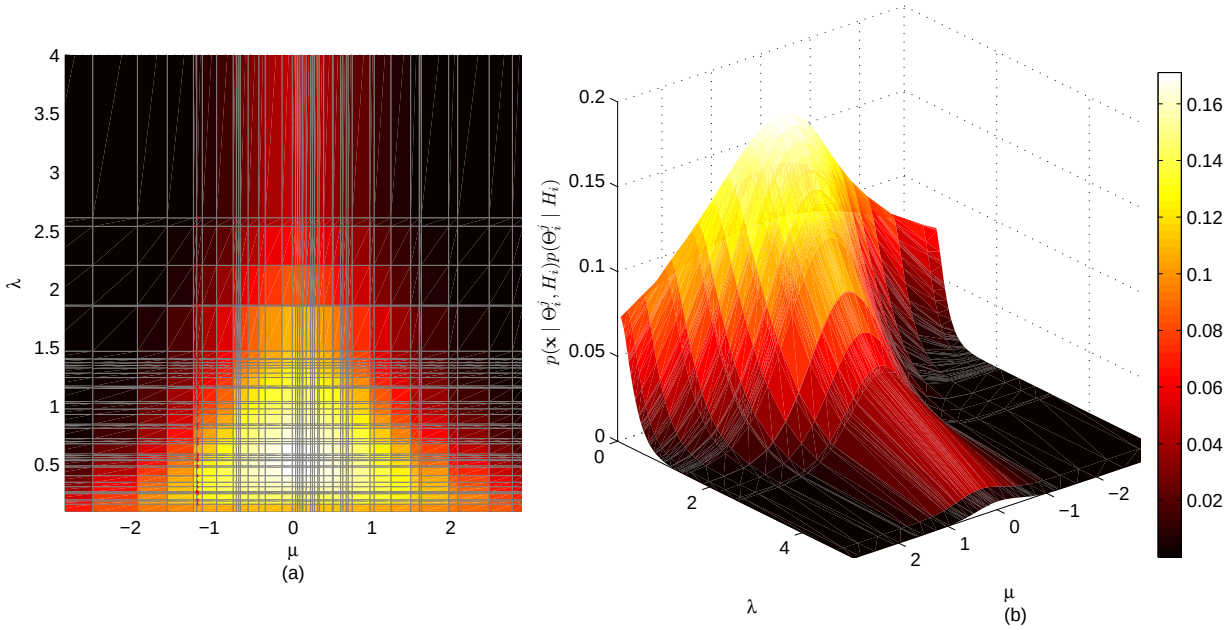


Figure 8.8: Approximating the evidence integral. (a) The $J \times J$ grid is shown in grey. (b) The corresponding value of the product $p(\mathbf{x} \mid \Theta_i^j, H_i)\ p(\Theta_i^j \mid H_i)$ for the j^{th} sample is shown on the z-axis.

Algorithm 8.1 An approximation of the evidence integral.

1. Generate J samples of Θ_i from prior distribution $p(\Theta_i \mid H_i)$
2. Compute the product $p(\mathbf{x} \mid \Theta_i^j, H_i)\ p(\Theta_i^j \mid H_i)$, for $j = 1, ..., J$
3. Compute the area $A_j = \delta_{\mu_i^j} \times \delta_{\lambda_i^j}$
4. Compute the incremental volume $\left[p(\mathbf{x} \mid \Theta_i^j, H_i)p(\Theta_i^j \mid H_i)\right] \times A_j$
5. Compute the sum of all incremental volumes using $\sum_{j=1}^{J} \left[\left(p(\mathbf{x} \mid \Theta_i^j, H_i)p(\Theta_i^j \mid H_i)\right) \times A_j\right]$

A $J \times J$ grid can be visualised on an x-y plane, representing a finite set of $J \times J$ samples of Θ_i generated from $p(\Theta_i \mid H_i)$. At each $\Theta_i^j = \left\{\mu_i^j, \lambda_i^j\right\}$, $p(\mathbf{x} \mid \Theta_i^j, H_i)\ p(\Theta_i^j \mid H_i)$ is computed. Figure 8.8 illustrates this concept. Let δx be an incremental distance on the x-axis representing $\delta_{\mu_i^j}$, and δy be an incremental distance on the y-axis representing $\delta_{\lambda_i^j}$. δz on the z-axis represents the corresponding $p(\mathbf{x} \mid \Theta_i^j, H_i)\ p(\Theta_i^j \mid H_i)$. Thus, the integrand in Equation 8.24 can be approximated by the volume $\delta x \times \delta y \times \delta z$. The sum of all these incremental volumes gives us an approximation of the probability mass which approximates $P(\mathbf{x} \mid H_i)$.

8.7.2 Verification of the Approximation

In order to verify Algorithm 8.1, the evidence was calculated for a hypothesis H_{source} after observing data $\mathbf{x}_{source}$ which is known to be generated under this same hypothesis H_{source},

and so H_{source} is known to be the "true" hypothesis given data $\mathbf{x}_{source}$. Therefore, it is expected that the evidence $P(\mathbf{x}_{source} \mid H_{source}) = 1$. Artificial data $\mathbf{x}_{source}$ was sampled from a synthetic hypothesis $H_{source} \sim \mathcal{N}(0, 1)$. By carrying out the steps summarised in Algorithm 8.1, $P(\mathbf{x}_{source} \mid H_{source}) = \sum_{j=1}^{J} \left[\left(p(\mathbf{x}_{source} \mid \Theta_i^j, H_i) p(\Theta_i^j \mid H_i) \right) \times A_j \right] = 0.99$ was obtained. Figure 8.8 shows samples generated from $H_{source} \sim \mathcal{N}(0, 1)$, and the resulting $\mathcal{NG}$ distribution.

Extension of the approximation to multi-dimensional datasets is described in Appendix B.2. Using the approximation, $P(\mathbf{x}_{test} \mid H_{source})$ was calculated for (a) $\mathbf{x}_{test} = [0]$, (b) $\mathbf{x}_{test} = [2]$, and (c) $\mathbf{x}_{test} = [10]$. The synthetic hypothesis $H_{source} \sim NG(\mu, \lambda \mid \mu_0 = 0, \lambda_0 = 1)$. As $\mu_{0_{source}} = 0$ and $\lambda_0 = 1$, between the three test cases, this hypothesis is most highly expected to be the "true" hypothesis for $\mathbf{x}_{test} = [0]$, and so the evidence $P(\mathbf{x}_{test} \mid H_{source})$ is expected to be the highest when $\mathbf{x}_{test} = [0]$.

Figure 8.9 shows $P(\mathbf{x}_{test} \mid H_{source})$ for cases (a), (b), and (c). As expected, the hypothesis has the highest evidence (i.e. $P(\mathbf{x}_{test} \mid H_{source}) = 1$) when $\mathbf{x}_{test}$=0, and the evidence for H_{source} decreases to $P(\mathbf{x}_{test} \mid H_{source}) \approx 0.35$ when test data $\mathbf{x}_{test} = [2]$, and further to $P(\mathbf{x}_{source} \mid H_{source}) \approx 0$ when $\mathbf{x}_{test} = [10]$, indicating that $\mathbf{x}_{test} = [0]$ is the most likely, and $\mathbf{x}_{test} = [10]$ is the least less likely, to have been generated by the uni-dimensional Normal distribution corresponding to hypothesis H_{source}.

8.7.3 BMN Results

Figure 8.10 shows a typical example time-series and the corresponding 5-D novelty score $z(\mathbf{x}_{new})$ for an example patient from $\mathcal{CH}_2^{NR_{test}}$ for the last-day period before discharge from the hospital. It may be seen that for this patient, as expected, the vital signs are largely unchanging, and the novelty score is constantly low, indicating that the patient's vital signs are "normal". For comparison, Figure 8.11 shows the track-and-trigger time-series of a patient from the $\mathcal{CH}_2^{EICU}$ group, for the last-day period before emergency ICU admission. At 18:00 the patient's HR is 102 bpm, the SBP is 140 mmHg, and the BR is 18 rpm. At 06:00 the HR has increased to 118 bpm, the SBP to 150 mmHg, while the BR is at 21 rpm. The increase in the HR and SBP is reflected in the novelty score which increases from 5.8 to 7.5. By 15:00 hours, the HR has risen up to 122 bpm, while the

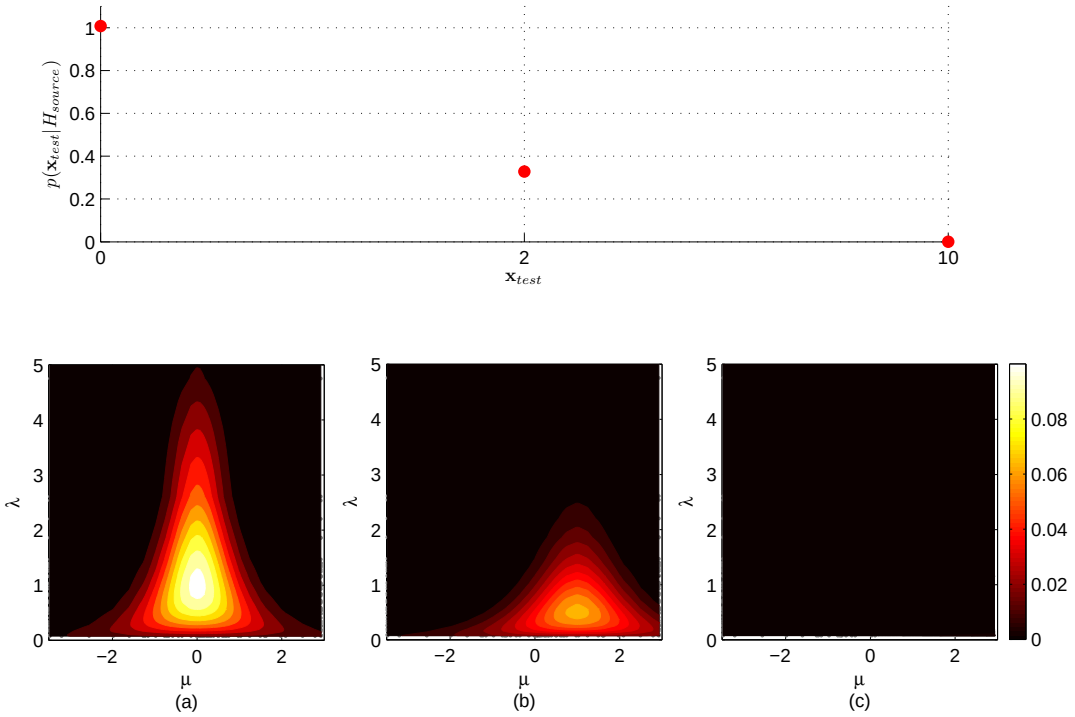


Figure 8.9: Evidence for an example hypothesis after seeing example artificial data. The upper plot shows $P(\mathbf{x}_{test} \mid H_{source})$ at the corresponding value of $\mathbf{x}_{test}$. The lower plots show the corresponding posterior distribution of the hypothesis after observation of each test data, for (a) $\mathbf{x}_{test} = [0]$, (b) $\mathbf{x}_{test} = [2]$, and (c) $\mathbf{x}_{test} = [10]$.

SBP is 142 mmHg. However, the BR has also increased from 21 rpm to 28 rpm, which is reflected in the novelty score $z(\mathbf{x}_{new})$ which further increases to 9.5. From this time onwards, the novelty score remains above the example threshold set at 6.2, indicating that the patient is deteriorating[5]. The last observation was recorded at 20:00 hours, and the patient had an emergency admission to the ICU at 21:27 hours.

The posterior probability $P(H_i \mid \mathbf{x}_{new})$ of 5 randomly selected hypotheses is also shown in Figures 8.10 (b) and 8.11(b) to illustrate how $P(H_i \mid \mathbf{x}_{new})$ changes upon the arrival of each $\mathbf{x}_{new}$. Appendix C contains further examples to illustrate how the posterior probability $P(H_i \mid \mathbf{x}_{new})$ and the novelty score $z(\mathbf{x}_{new})$ change over time for patients from the $\mathcal{CH}_2^{EICU}$ and $\mathcal{CH}_2^{NR_{test}}$ groups.

The BMN was trained using $\mathcal{CH}_2^{NR_{train}}$, and was tested using test sets $\mathcal{CH}_2^{EICU}$ and $\mathcal{CH}_2^{NR_{test}}$ over a range of thresholds T on $z(\mathbf{x}_{new})$ where T is as defined in Section 8.6.4. The selection of the range of T using $\mathcal{CH}_2^{NR_{train}}$ is explained as follows.

For the 2-D model, the lowest novelty score obtained for a $\mathcal{CH}_2^{NR_{train}}$ patient was

[5]The reasons for placing the threshold on the novelty score at 6.2 are explained in detail later in this section

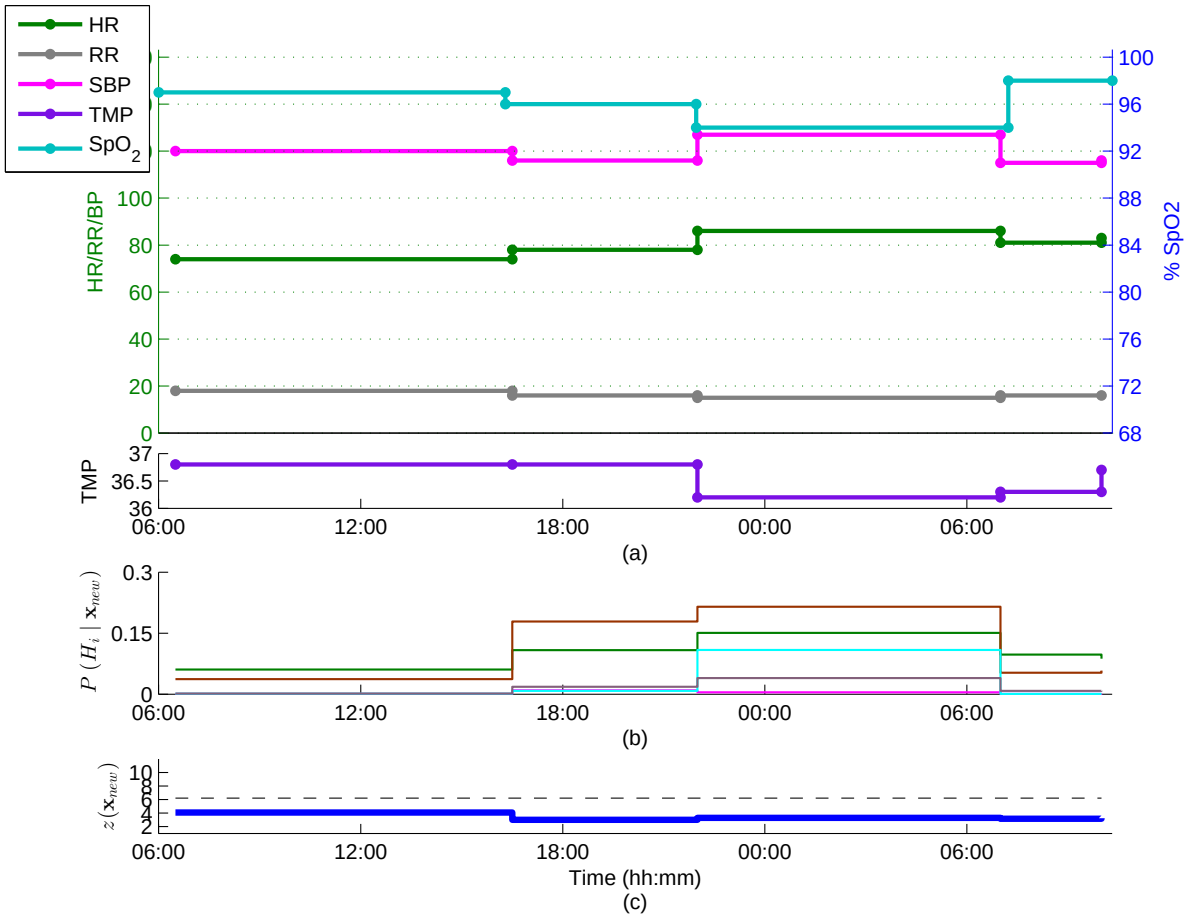


Figure 8.10: Time-series of the last-day period up to the time of recording of the last vital-sign observation before discharge from the hospital, for a patient (Study ID 3) from the $\mathcal{CH}_2^{NR_{test}}$ group. (a) Vital sign observations recorded using track-and-trigger. (b) The posterior probability $P(H_i \mid \mathbf{x}_{new})$ of 5 randomly chosen hypotheses from the 137 hypotheses H_i. (c) The novelty score $z(\mathbf{x}_{new})$. An example threshold of $z(\mathbf{x}_{new}) = 6.2$ is shown by the dashed line in black. An alert is indicated in red if $z(\mathbf{x}_{new}) > 6.2$.

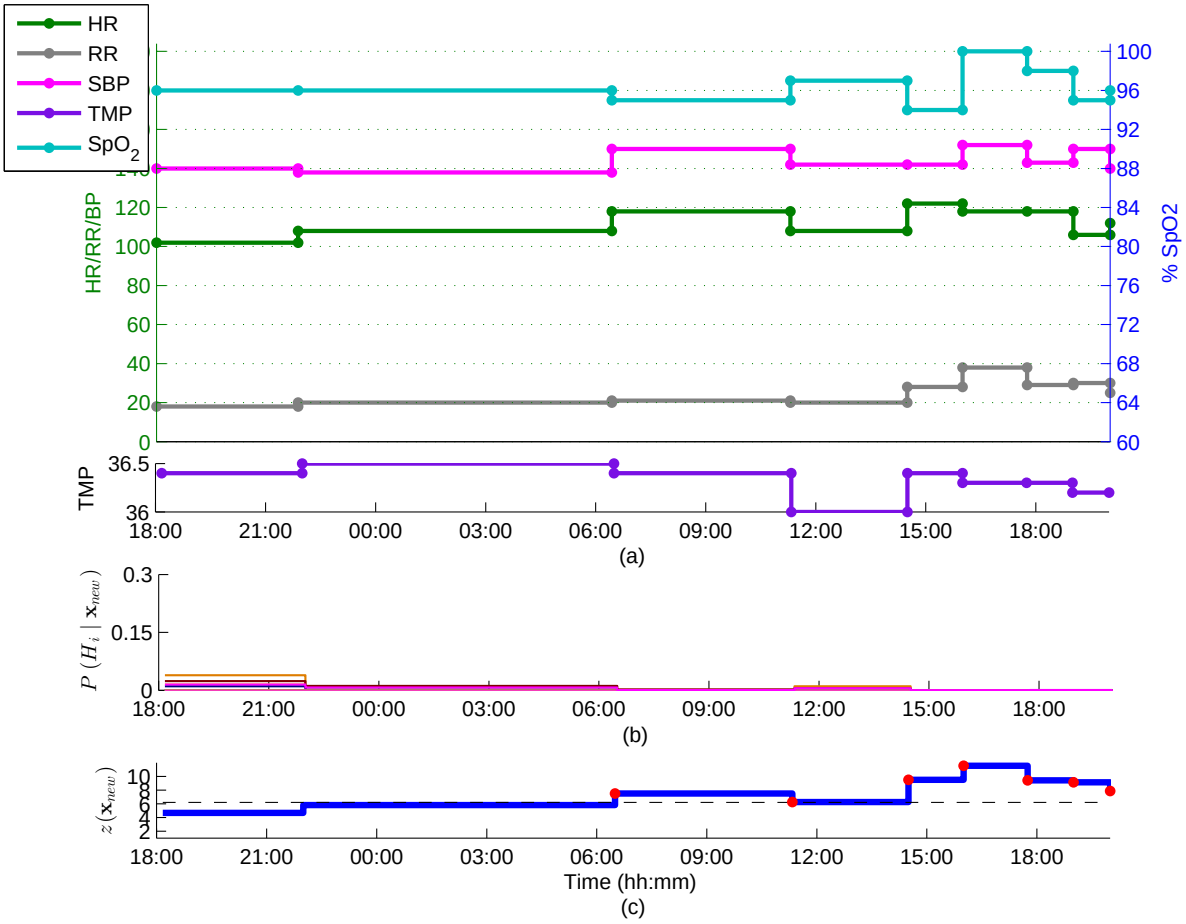


Figure 8.11: Time-series of the last-day period up to the time of recording of the last vital-sign observation before emergency admission to the ICU, for a patient (Study ID 125) from the $\mathcal{CH}_2^{ICU}$ group. (a) Vital-sign observations recorded using track-and-trigger. (b) The posterior probability $P\left(H_i \mid \mathbf{x}_{new}\right)$ of 5 randomly chosen hypotheses from the 137 hypotheses H_i. (c) Novelty score $z(\mathbf{x}_{new})$. An example threshold on $z(\mathbf{x}_{new}) = 6.2$ is shown by the dashed line in black. An alert is indicated in red if $z(\mathbf{x}_{new}) > 6.2$.

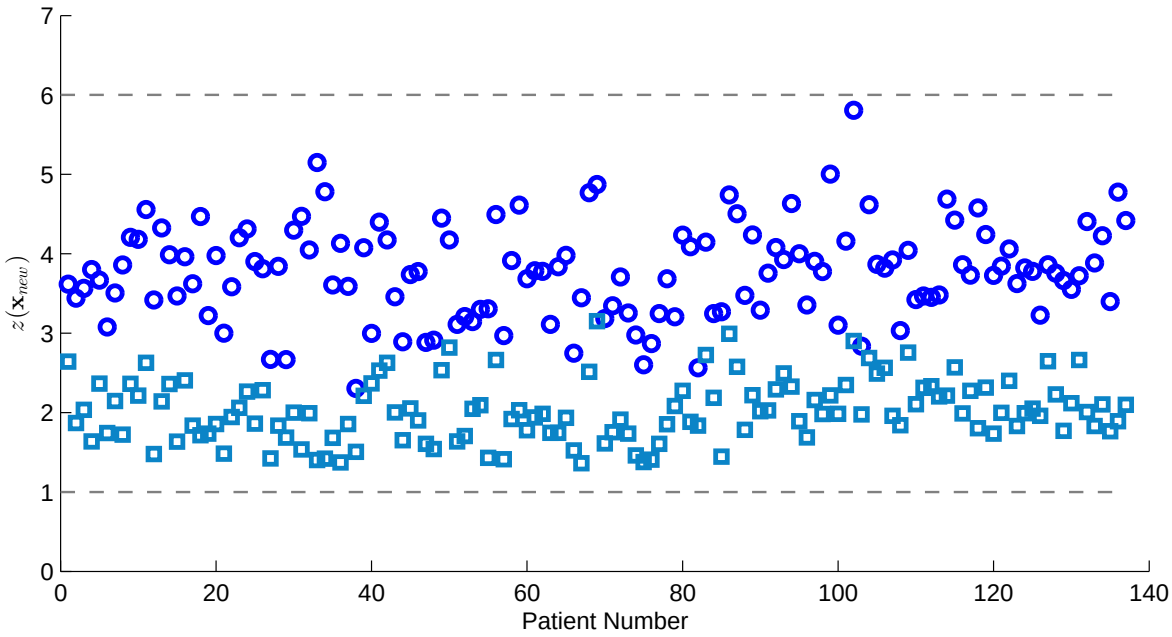


Figure 8.12: The highest novelty score $z(\mathbf{x}_{new})$ for each of the 137 patients in the training set $\mathcal{CH}_2^{NR_{train}}$. The mean of the highest $z(\mathbf{x}_{new})$ for a patient over 50 experiments is shown by a dark blue circle, and the mean of the lowest $z(\mathbf{x}_{new})$ for that patient over 50 experiments is shown by a light blue square.

$z(X_{new}) = 3.0$ and the highest score for a patient was $z(\mathbf{x}_{new}) = 7.8$. Therefore, the 2-D model was tested over the range $3 \leq T \leq 9$. Following the same reasoning, for the 5-D model, the lowest score for a patient was $z(\mathbf{x}_{new}) = 1.3$, and the highest novelty score obtained for a $\mathcal{CH}_2^{NR_{train}}$ patient was $z(X_{new}) = 5.9$, as shown in Figure 8.12. Therefore, the 5-D model was tested over the range $0 \leq T \leq 7$. The range of T is different for the 2-D and 5-D models because $z(\mathbf{x}_{new})$ depends on $P(\mathbf{x}_{new})$ which covers a different range for the 2-D model than for the 5-D model.

The results of testing the BMN using test sets $\mathcal{CH}_2^{EICU}$ and $\mathcal{CH}_2^{NR_{test}}$ are shown in Figure 8.13(a) for the 2-D model, and in Figure 8.13(b) for the 5-D model. At approximately the middle of the range of T, for example at $T = 5.5$, the 2-D model generates alerts for 92.1% of the patients in the $\mathcal{CH}_2^{EICU}$ group (shown in red), and 47.3% of the patients in the $\mathcal{CH}_2^{NR_{test}}$ group (shown in green). At $T = 7.8$, this false alert rate (i.e., r_1 for the $\mathcal{CH}_2^{NR_{test}}$ group, shown in green) drops to 0%, whereas the model still generates alerts for 68.4% of the patients in the $\mathcal{CH}_2^{EICU}$ group. In Figure 8.13(b), the performance of the 5-D model is even better; at $T = 6.2$, it generates alerts for 82.0% of the patients in the $\mathcal{CH}_2^{EICU}$ group (shown in red), and for only 2.6% of the patients in the $\mathcal{CH}_2^{NR_{test}}$ (shown

in green). At approximately the middle of the range for the 5-D model (i.e. at $T = 3$), although the 5-D model generates alerts for 100% of the patients in the $\mathcal{CH}_2^{EICU}$ group (shown in red), it also generates false alerts for 79.4% of the patients in the $\mathcal{CH}_2^{NR_{test}}$ group (shown in green).

The false-alert rate (i.e., r_1 for the $\mathcal{CH}_2^{NR_{test}}$ group) stays constant at 0% for $T \geq 7.8$ for the 2-D model, and $\leq 2\%$ at $T \geq 6.2$ for the 5-D model, therefore $T = 7.8$ and $T = 6.2$ are considered to be the optimal thresholds for the 2-D and 5-D models, respectively.

The performance of the BMN is compared with that of other methods in the next section. Here it is important to mention that the performance of the BMN at the optimal 2-D and 5-D thresholds will be used for this comparison. However, as can be seen in Figure 8.13, the performance is subject to the choice of threshold T. If, for a given model, a much higher or much lower threshold is selected instead of the optimal threshold set here, then the performance (in terms of the sensitivity and false alert rate) of the BMN can be expected to change correspondingly, which must be kept in mind when comparing this method with those introduced in the next section.

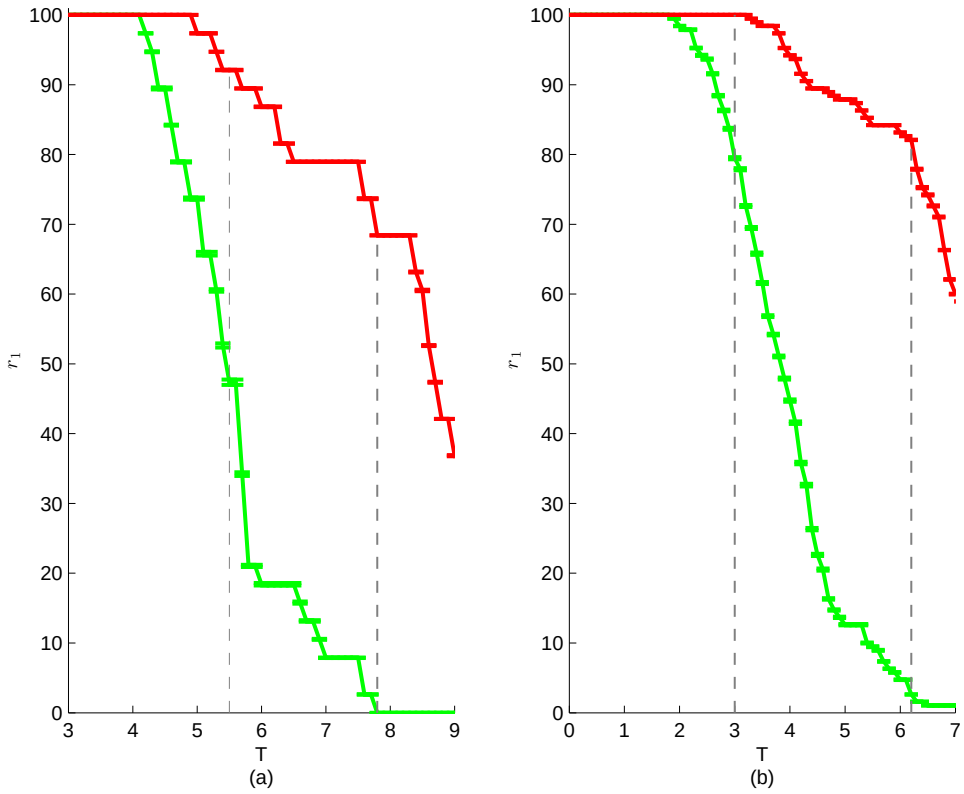


Figure 8.13: (a) r_1 (%) for the 2-D model, over $3 \leq T \leq 9$. (b) r_1 (%) for the 5-D model, over $0 \leq T \leq 7$. Results for sets $\mathcal{CH}_2^{NR_{test}}$ and $\mathcal{CH}_2^{EICU}$ are shown in green and red, respectively. r_1 for the $\mathcal{CH}_2^{NR_{train}}$ is also shown in the background in blue. Grey dashed lines indicate results at $T = 5.5$ and $T = 7.8$ in (a), and at $T = 3$ and $T = 6.2$ in (b). r_1 for the $\mathcal{CH}_2^{NR_{test}}$ group represents the false alert rate.

8.8 EWS Models

In order to compare the performance of the Parzen windows model and the BMN with the system(s) used in current practice in hospitals in the UK, three early warning score (EWS) systems were considered.

The EWS system which was in use in the upper GI ward during phase 1 of the CALMS-2 trial (Table 7.1) is reproduced here in Table 8.3 as the CALMS-2 EWS.

Also compared was the centile-based early warning score (CEWS [158]) model (shown in Table 8.4) which is the current system in use in the Oxford University Hospitals in Oxford; and lastly, the more recent national early warning score (NEWS [117]) model (shown in Table 8.5), which is being introduced in some NHS hospitals in the UK as of 2013.

If, for a given EWS model, the corresponding alert criteria for that model are met, an

Vital Sign	Score						
	3	2	1	0	1	2	3
HR		≤ 40	41-50	51-100	101-110	111-129	≥ 130
BR	≤ 8			9-18	19-24	25-29	≥ 30
SpO_2	≤ 91			≥ 92			
SBP	≤ 90	91-99		101-179			≥ 180
TMP		≤ 35.0		35.1-37.9			≥ 38.0

Table 8.3: CALMS-2 EWS. A score of 3 in an individual vital sign, and a score of 4 in the aggregate score of the five vital signs generates an alert.

alert is generated; i.e. if the individual or aggregate score of the vital-sign observation-set for a patient exceeds the alert threshold used by that model, then an alert is generated. In addition to the vital signs, these EWS also score level of consciousness. However, the three EWS use different measure of consciousness [6]; further while these latter scores are recorded on the track-and-trigger paper chart, they are not always included by the nurse while calculating the aggregate score, and are therefore not considered in this work.

Vital Sign	Score						
	3	2	1	0	1	2	3
HR	≤ 42	$43-49$	$50-53$	$54-104$	$105-112$	$113-127$	≥ 128
BR	≤ 7	$8-10$	$11-13$	$14-25$	$26-28$	$29-33$	≥ 34
SpO_2	≤ 84	$85-90$	$91-93$	≥ 94			
SBP	≤ 85	$86-96$	$97-101$	$102-154$	$155-164$	$165-184$	≥ 185
TMP	≤ 35.4		$35.5-35.9$	$36.0-37.3$	$37.4-38.3$		≥ 38.4

Table 8.4: The centile-based early warning score (CEWS) system. A score of 3 in an individual vital sign, and a score of 3 in the aggregate score of the five vital signs generates an alert.

Vital Sign	Score						
	3	2	1	0	1	2	3
HR	≤ 40		$41-50$	$51-90$	$91-110$	$110-130$	≥ 131
BR	≤ 8		$9-11$	$12-20$		$21-24$	≥ 25
SpO_2	≤ 91	$92-93$	$94-95$	≥ 96			
SBP	≤ 90	$91-100$	$101-110$	$111-219$			≥ 220
TMP	≤ 35.0		$35.1-36.0$	$36.1-38.0$	$38.1-39.0$	≥ 39.1	

Table 8.5: The national early warning score (NEWS) system. An aggregate score of 5, or an individual score of 3 in a single vital sign, generates an alert.

[6]These include Alert Voice Pain Unresponsive (AVPU), Glasgow coma scale (GCS), Canadian neurological score (CNS).

8.9 Overall Results

The EWS models were applied to the vital-sign data of each patient in the test sets $\mathcal{CH}_2^{EICU}$ and $\mathcal{CH}_2^{NR_{test}}$. As dataset $\mathcal{CH}_2^{NR_{train}}$ was not used to train these EWS models, the models were also applied to this dataset. Both datasets contain the last-day data for patients, as defined in Section 8.3.

Table 8.6 shows the results for the three EWS models, with the sensitivity and false alert rate as defined in Section 8.2 are reported. Also shown are the results for the 5-D Parzen windows and BMN models, at their respective optimal thresholds T; i.e. for the Parzen windows, the sensitivity and specificity at $T = 0.99$ are shown, and for the BMN, the sensitivity and specificity at $T = 6.2$ are shown. (As highlighted at the end of Section 8.7.3, although the BMN is discussed at the optimal threshold value of $T = 6.2$, for a fair comparison, it may be recalled that model performance is subject to the choice of this optimal threshold; if a different threshold is considered, for instance, the model's performance may be worse than that at the optimal threshold.)

The Parzen windows and BMN models, as described in earlier sections, were trained using $\mathcal{CH}_2^{NR_{train}}$ and were tested using the test sets $\mathcal{CH}_2^{EICU}$ and $\mathcal{CH}_2^{NR_{test}}$. The following may be noted from these results.

Sensitivity and False Alert Rate

- CALMS-2 EWS and NEWS have low false-alert rates (of 10.2% and 6.9%, respectively), implying that they perform well at not generating false alerts for the 19 patients in the test set who recovered without event and were discharged from the hospital. However, NEWS and CEWS each have a sensitivity of $\sim 78\%$, indicating that they both fail to generate alerts for approximately 4 out of the 19 patients who had emergency ICU admissions. The CALMS-2 EWS has an even lower sensitivity of $\sim 68\%$, indicating that it fails to generate alerts for 6 out of the 19 patients. As shown in the right-most column of Table 8.6, the false alert rates of these three EWS models remain consistent when tested on the larger dataset $\mathcal{CH}_2^{NR_{train}}$ which contains 137 patients. With a sensitivity similar to that of CEWS, and the lowest false alert rate among the three EWS, NEWS may be said to perform the best

Methods	*sensitivity* (%)	*false alert rate* (%)	*false alert rate** (%)
CALMS-2 EWS	68.4	10.2	10.3
CEWS	78.9	22.5	23.8
NEWS	78.8	6.9	6.3
Parzen windows	90.3	26.1	–
BMN	82.0	2.6	–

Table 8.6: A comparison of the performance of models considered in this chapter, tested using the 5-D dataset. The second and third columns show the sensitivity and false alert alert obtained using datasets $\mathcal{CH}_2^{EICU}$ and $\mathcal{CH}_2^{NR_{test}}$. *For the three early warning score models (i.e., CALMS-2 EWS, CEWS, and NEWS), the false alert rate obtained by applying the models to dataset $\mathcal{CH}_2^{NR_{train}}$ is also shown in the right-most column.

among these EWS models.

- Compared to the EWS systems, the Parzen windows model has a much higher sensitivity (90.3%), but it also has a 26.1% false alert rate, implying that the model is perhaps over-sensitive; in that, while it generates alerts for, on average, 17 out of the 19 patients who had an adverse event, it also generates alerts for ~ 5 out of the 19 of patients who did not have an adverse event, on average.

- In contrast, the BMN, which was designed to overcome the limitation of small number of per-patient training data, has a much lower false alert rate (2.6%). The low false alert rate may be attributed to the design of the model which takes into account the fact that 5.35 samples per patient may not be an accurate representation of the patient's "normal" physiology. Its sensitivity (82.0%), however, is only less than 10% lower than that of the Parzen windows model, which is the most sensitive of all models considered in Table 8.6.

First Alert

For a $\mathcal{CH}_2^{EICU}$ patient for whom an alert was generated in their last-day period, the timing of the generation of the first alert was investigated. As the last-day duration of individual patients was variable, this work also investigated how many hours before the last observation of that patient, the first alert was generated using a given model. The first alert timing estimates how early a model can detect the emergency ICU admission. Figure 8.14 shows the results for the Parzen Windows, and BMN models, as well as for

CEWS (which is the system in current use in hospitals within the Oxford University Hospitals NHS Trust). As there are only 19 patients in the $\mathcal{CH}_2^{EICU}$ group, the timing of the first alert for each individual patients is shown, instead of average results for the entire group. In particular, the three patients (whose study IDs were 140, 155, and 191) who had an emergency ICU admission within 24 hours of their admission to the upper GI ward are indicated in the figure.

While these results are based on only 19 patients, it is interesting to note that, out of the total number of patients for whom an alert is generated by a given mode, both Parzen windows and the BMN model generate the first alert for more than half of the patients 20 or more hours before the time of their last vital-sign observation before the emergency ICU admission. To elaborate, the Parzen windows generates alerts for 11 out of 17 patients at least 20 hours before the last observation, and the BMN model generates alerts for 13 out of 16 patients at least 20 hours before the last observation; whereas CEWS generates alerts for only 5 out of the 15 patients at least 20 hours before the last observation.

8.10 Discussion and Conclusions

In Chapter 2 of this thesis, it was stated that a patient monitoring system has a two-pronged function: (i) it should generate an alert for a patient who goes on to have an adverse event; and (ii) it should not generate alerts for patients who do not deteriorate. Thus the models considered in this chapter should be evaluated on both counts: sensitivity and false-alert rate. With this view in mind, although the Parzen windows model has the highest sensitivity (90.3%), it may be argued that in clinical practice, a system that generates unwanted alerts 26.1% of the time, is not truly beneficial. Based on this retrospective analyis, it may be argued that the BMN is better as it vastly minimises the false alert rate, while maintaining a sensitivity of 82.0%.

Secondly, it is recognised here that Table 8.6 reflects the results for a small number of test patients (19 EICU, and 19 NR patients). The 82.0% sensitivity of the BMN method, for example, means that it correctly generated an alert for 16 out of 19 patients who went on to have an emergency ICU admission. The 26.1% false alert rate of the Parzen windows

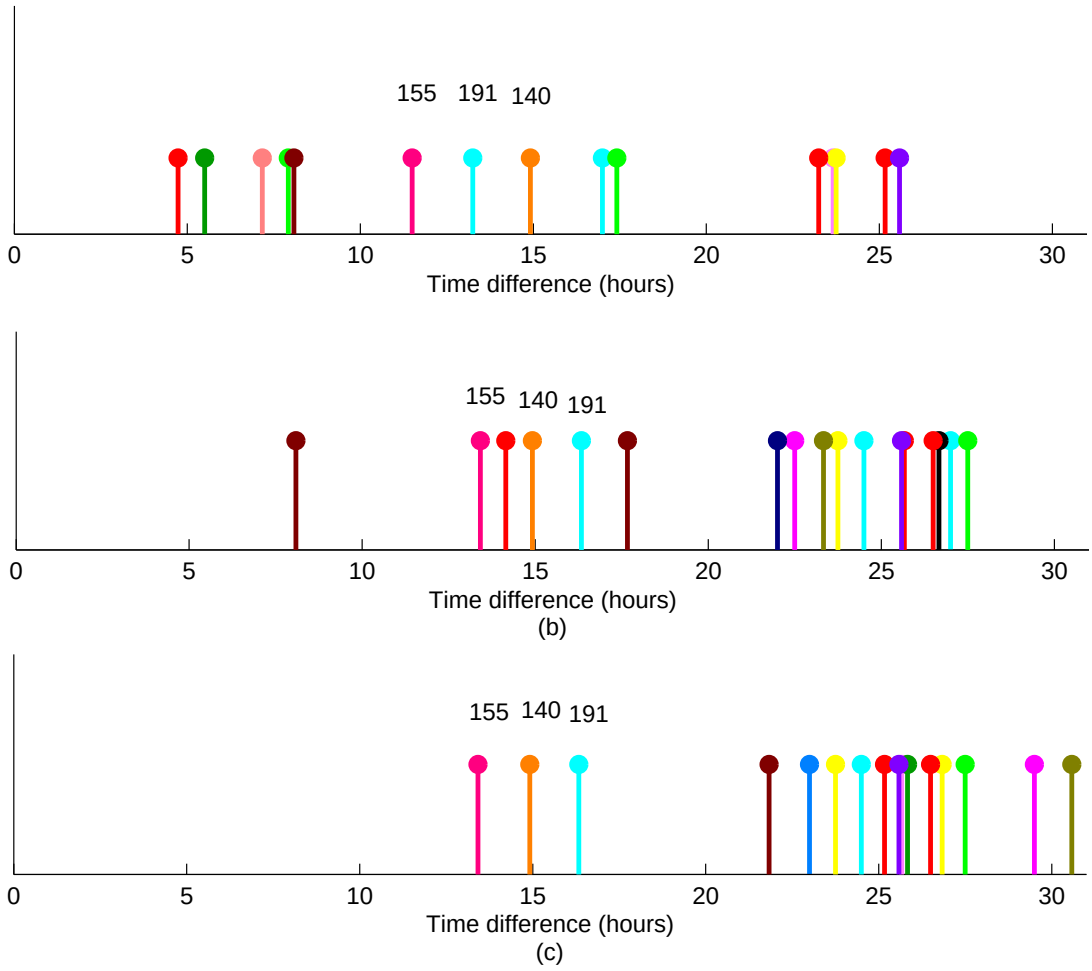


Figure 8.14: Timing of first alert for patients in the $\mathcal{CH}_2^{EICU}$ group, looking back from the time of the last observation, obtained using (a) CEWS, (b) Parzen windows, and (c) the BMN. The x-axis shows the difference, in hours, between the time of the first alert and the time at which the last observation was recorded before emergency ICU admission. Patients with study IDs 140, 155, and 191, who had an emergency ICU admission to the ICU within 24 hours of their admission to the upper GI ward, are indicated in orange, bright pink, and light blue.

model translates to false alerts in approximately 5 patients out of the 19 patients.

The EWS model with the highest sensitivity (CEWS) generated alerts for 78.8% of the patients in the EICU group, but also for 23.8% of the patients in the NR group (here considering the results for the 137 patients in $\mathcal{CH}_2^{NR_{train}}$), implying that it generated false alerts for almost a fifth of the patients. This may be explained by considering that none of the EWS models were developed specifically for the upper GI ward population. This is reflected in the scoring criteria used by these models. For instance, all three EWS models assign a score of only 1 to a HR of 110. From Figure 8.1(a) it can be seen that the higher mode of the HR of the $\mathcal{CH}_2^{EICU}$ group is at 110, implying that a HR of 110 is high enough to indicate abnormality in upper GI patients, and should therefore perhaps be assigned a higher score than 1 in an EWS model tuned to this population. Similarly, it can be seen from Figure 8.1 (b) that the $\mathcal{CH}_2^{EICU}$ group has a high BR mode at approximately 19 rpm, and further from Table 8.1 that the mean BR for the EICU group is 20.6 rpm. The CALMS-2 EWS assigns a score of 1, and both CEWS and NEWS assign a score of 0 to a BR of 20 rpm. The EWS systems are not calibrated for the upper GI ward population, which explains their relatively low sensitivity, compared to that of the Parzen windows and BMN models which were designed for this population. Further, it is reiterated that the results shown in Table 8.6 are based on data corresponding to the last day prior to ICU admission or discharge. Therefore it is not surprising that even the model with the lowest sensitivity (CALMS-2 EWS) generates alerts for 13 out of the 19 patients who had an emergency ICU admission.

Finally it is stressed again that in addition to accuracy (i.e. sensitivity and specificity), early detection is a desirable feature in a patient monitoring model. To this end, it was found that both automated models (Parzen windows, and the BMN model) generated alerts earlier than the EWS models, in more than half the patients considered. Further, they generated these alerts at least 20 hours before the adverse event. Although based on only 19 patients, these results are encouraging, and provide the basis for validating the early detection feature of these models using a larger dataset in the future.

In the final chapter, a future extension of the proposed BMN method towards patient-specific models is suggested, such that it would allow detection of deterioration from

patient-specific norms.

Chapter 9

Conclusion

9.1 Summary of Results

This thesis has described models for detection of deterioration in high-risk patients in two environments – (a) a step-down unit in which patients recovering from an ICU stay are being continuously monitored, and (b) a general ward where patients are recovering after having cancer surgery. The key findings are as follows.

- Improved classification of "normal" and "abnormal" vital signs can be performed by a "hybrid" method which combines a probabilistic approach with traditional classification approaches. The hybrid method not only results in fewer misclassifications of "normal" and "abnormal" vital signs, it also correctly identifies two rarer types of abnormalities, namely bradycardia and bradypnoea, which are usually misclassified by existing techniques.

- Abnormalities in vital signs can be modelled explicitly. Results showed that by constructing individual models using data describing tachycardia, bradycardia, tachypnoea, bradypnoea, and hypoxia, a patient status output can be generated which not only generates alerts when one more vital signs deviate from normal ranges, but also indicates which type of abnormality the patient is experiencing.

- Retrospective data analysis shows that a model designed to take into account the low-frequency of manually-acquired vital-sign data outperforms existing methods for (a) detecting deterioration in patients who go on to have an emergency ICU

admission, and (b) not generating false alerts for patients who recover from surgery without an adverse event.

- The automated methods considered in this thesis generate alerts for deteriorating patients earlier than early warning score systems which are currently in use in hospitals in the UK. The automated methods were found to generate alerts for the majority of the patients considered at least 20 hours prior to emergency ICU admission.

9.1.1 Classification of "Normal" and "Abnormal" Vital Signs

In the first half of this thesis, different approaches for classification of vital-sign data acquired from high-risk patients recovering from an ICU stay were described, with a view to developing a classification model that (a) best separated "abnormal" vital-sign data from "normal" data, (b) without misclassifying, in particular, "abnormal" data representing the rarer types of abnormalities, e.g. bradycardia and bradypnoea. A large dataset containing continuous-monitoring vital-sign data from 326 patients in the step-down unit of a hospital in the USA was considered. This dataset had been reviewed by clinical experts who labelled vital signs occurring within any period they considered to be associated with an adverse event, as being "abnormal".

Based on previous work on vital-sign novelty detection [67], in Chapter 3, a novelty detection (or one-class) approach was firstly adopted to construct a model of normality, without considering the clinical labels associated with "normal" and "abnormal" data. In novelty detection, a model of normality is bounded by some threshold of normality, such that data lying above that threshold are considered to be outside the boundary of normality, and hence "abnormal". An optimal threshold must be set. In Chapter 4, two approaches were considered for selecting such a threshold. First, a threshold was determined such that the distributions of the "normal" and "abnormal" classes estimated using the model of normality would best resemble the distributions of the "normal" and "abnormal" classes according to the clinical labelling. At the optimal threshold the Kullback-Leibler divergence [90] between the distributions of the estimated and the clinically labelled classes would be minimum. Second, a threshold was sought such that it

would result in the smallest classification error. Both approaches fail to correctly classify bradycardia- and bradypnoea-related data in the 4-D continuous vital-sign dataset. A two-class approach was then considered in Chapter 5 in which two-class classifiers were trained using "normal" *and* "abnormal" vital-sign data. Additionally, a "hybrid" technique was proposed in which classifiers were trained using only those clinically-labelled "normal" data which were within a threshold placed on the model of normality, and using only those clinically-labelled "abnormal" data which were above the threshold. Additional "hybrid" criteria were included based on the knowledge that the original clinical labels were imprecise leading to a large overlap between the "normal" and "abnormal" class, and causing the misclassification of bradycardia and bradypnoea data when the two-class classifiers were trained without the "hybrid" criteria. Using the proposed method, an optimal threshold was found such that (i) overall misclassification was reduced, and (ii) misclassification of bradycardia and bradypnoea data was reduced.

After an optimal threshold separating the "normal" and "abnormal" classes had been determined in Chapter 5, it became possible to construct models for sub-classes within the "abnormal" class. A multi-class output was developed in Chapter 6, based both on models of individual types of abnormality and a model of normality. The proposed multi-class output was shown not only to alert when one or more vital signs departed from normality, but also to indicate which type of abnormality was associated with the deterioration. Validation of the multi-class model requires an independent, and reliably labelled dataset, and is considered as future work.

9.1.2 Vital Signs in the SDU and Upper GI Ward

The latter half of this thesis has focused on models for detection of deterioration in patients recovering in a general ward after having cancer surgery. Following surgery these patients are admitted to the upper GI ward (unless they require a higher level of care), where they are monitored using a track-and-trigger system. Post-operative recovery usually involves an initial period lasting 4-5 days during which the patient recovers from the surgery, after which the patient becomes stable and is then discharged from the hospital, after a total of 8 to 10 days on the ward, on average. A few patients deteriorate on the ward such that an

emergency admission to the ICU has to take place. This adverse event may occur during the initial recovery period, or shortly thereafter, i.e. around day 4 or 5 on the ward.

Vital signs in the upper GI ward are collected manually by nurses at (on average) 4-5 hourly intervals. In Chapter 7 it was shown that their distribution is different to that of the vital signs collected via continuous monitoring in the step-down unit. In particular, there were considerably fewer examples of "abnormal" data in the upper GI dataset ($\mathcal{CH}_2$) than there were in the SDU dataset ($\mathcal{D}_1$). Further, the $\mathcal{D}_1$ dataset contains data for complete deterioration episodes, whereas in the $\mathcal{CH}_2$ dataset, only data up to a few hours before the emergency ICU admission are available.

9.1.3 Accuracy and Early Detection

Using the $\mathcal{CH}_2$ dataset, a density-based model was constructed in Chapter 8, and a Bayesian model was constructed in an attempt to deal with the low frequency of data collection. It was shown that the density-based model had the highest sensitivity, while the Bayesian model had the lowest false alert rate. More importantly, both these automated methods outperformed the early warning score systems that are currently used in hospitals in the UK. For patients who had emergency ICU admissions, both automated methods generated alerts earlier than the systems currently used.

The proposed model should be tested on a larger dataset containing more emergency ICU examples than the 19 used here. At the time of completion of this thesis, only the data from the 200 patients in the first phase of the CALMS-2 study were available. The proposed model can be validated in future by testing the performance using the data (from another 200 patients) collected during this second phase.

9.2 Future Work

9.2.1 Extension of BMN

The Bayesian model of normality proposed in Chapter 8 is based on the principles of Bayesian model averaging, in which hypotheses of varying order of complexity are traditionally weighted. In this thesis, only the Gaussian form was considered for all the

hypotheses. An exhaustive treatment would entail investigating the suitability of , for instance, a bi-modal Gaussian, or a Cauchy distribution of vital-sign data. It is also important that a one-sided distribution should be used for SpO_2.

The relative importance of vital signs in a model continues to be a matter of debate, which is why different early warning scores assign different weights to vital signs. Throughout this thesis, and as done in previous work, it is assumed that all vital signs carry equal weight in the model. In a future extension, a model that assigns, for example, more weight to the breathing rate, may be compared with one that assigns equal weights to all vital signs.

9.2.2 Patient-Specific Models

The BMN is described in Chapter 8 as a patient-based model, due to the fact that the hypotheses it contains are learned from individual patients; one hypothesis for each "normal" patient. However the overall model of normality is still a population-based model in the sense that a test patient's data will be compared with the model trained using the *population* of "normal" patients. A fully patient-specific model would be learned using only the patient's own data. Any incoming sample from the patient would be compared with historical data from the same patient. The rationale for such a patient-specific approach is based on the knowledge that for some patients, "normality" may be different from some population-specific "normality", in which case the population-specific model is not correctly calibrated for such patients. Previous work [17] has considered patient-specific models for individual dialysis patients, using data from periods with normal blood pressure as training data for a patient-specific model of normality.

A model of patient-specific normality for the upper GI patients needs to take into account the knowledge that vital signs during the first 4 to 5 days after surgery are "abnormal". Hence both a time-varying population-based model for the recovery period, and a complementary patient-specific model from day 5 onwards need to be considered.

Different approaches exist for performing online learning of time-varying data. Support vector regression, for example, has previously been proposed for performing online novelty detection in temporal data [98]. Several trend-based online pattern recognition and change

detection approaches for ICU data have been proposed [58, 142, 24]. The BMN model presented in this thesis could be adapted for on-line learning of track-and-trigger vital-sign data.

Instead of considering track-and trigger samples as being independent and identically distributed (i.i.d.), the model weights could be updated each time new data from the patient arrives. Recalling from Section 8.6.4, for a given patient in the training set, the distribution of the parameters Θ_i which describe the vital-sign distribution for the patient, are estimated by recursive updating:

$$p(\Theta_i \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, H_i) \propto p(\mathbf{x}_N \mid \Theta_i, H_i) p(\Theta_i \mid \mathbf{x}_1, \ldots, \mathbf{x}_{N-1}, H_i) \tag{9.1}$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are samples of vital-sign data belonging to the given patient. The weight of the hypothesis H_i for this patient, $P(H_i \mid X)$, was fixed after all the training data from the given training patient had been observed, and the probability, $P(\mathbf{x}_{new} \mid H_{i\ldots I})$, of any test data from a new patient, $\mathbf{x}_{new}$, was then computed using the fixed weights in the model.

$$P(\mathbf{x}_{new} \mid H_{i\ldots I}) = \sum_{i=1}^{I} P(H_i \mid \mathbf{x}_{new}). \tag{9.2}$$

In the patient-specific model, instead of using the fixed weights, $P(H_i \mid X)$, the weights are to be updated as the n^{th} sample from a test patient arrives. To simplify, Θ_i must continually be updated such that the recursive updating Equation 9.1 now becomes

$$p(\Theta_i \mid \mathbf{x}_{new_1}, \ldots, \mathbf{x}_{new_n}, H_i) \propto p(\mathbf{x}_{new_n} \mid \Theta_i, H_i) p(\Theta_i \mid \mathbf{x}_{new_1}, \ldots, \mathbf{x}_{new_{n-1}}, H_i) \tag{9.3}$$

As more data $\mathbf{x}_{new}$ from the new patient are observed, $p(\Theta_i \mid \mathbf{x}_{new_1}, \ldots, \mathbf{x}_{new_n}, H_i)$ and consequently the corresponding weight $P(\mathbf{x}_{new} \mid H_i)$, become more specific to this new patient. Eventually, when sufficient data from the new patient have been observed, $P(\mathbf{x}_{new} \mid H_{i\ldots I})$ will now reflect the probability that new data belong to this patient-specific model, and deviations in the corresponding novelty score $z(\mathbf{x}_{new})$ will represent departures from patient-specific normality.

9.2.3 Combined Continuous and Track-and-Trigger Monitoring

Phase 2 of the CALMS-2 trial, which involves continuous monitoring of ambulatory patients on the upper GI ward, is nearly complete. One of the biggest challenges in continuous monitoring of ambulatory patients is that of artefact and data dropouts caused by patient motion, or disconnection of probes. Therefore track-and-trigger remains the default monitoring system. A future model should consider combination of the high-frequency continuous data with the low-frequency track and trigger data in such a manner that when a continuous channel, e.g. SpO_2 drops out, the system is able to cope with it using the default lower frequency track-and-trigger signal.

The traditional approach for dealing with periods of artefactual or missing data is to complete the gap in a vital-sign signal by the population mean for that vital sign. However, doing so for a patient who is deteriorating during such a period would erroneously indicate that the patient is normal. This can be addressed by using regression techniques such as Gaussian processes (GP) [131], which estimate values of the missing data by considering a distribution over functions, and provide estimates along with bounds of uncertainty on the estimated values. In [176], a GP has been used for completing data gaps in Emergency Department patient-monitoring data.

A combination system can deal with two scenarios: first, continuous data are missing but track-and-trigger data are available; second, both continuous and track-and-trigger data are not available. This second case would occur if the gap in continuous data occurs between two successive track-and-trigger observations. In the first case, the GP should be used to estimate missing data from the last available continuous sample to the next track-and trigger observation, until the continuous data are available again. In the second case, where both the continuous and track-and-trigger data are not available, the GP-estimated value should be used throughout the gap period.

9.2.4 Vital-Sign Variability

In Chapter 7, it was found that deterioration may be indicated not only via abnormal changes in the amplitude of vital signs, but also via an increase in vital-sign variability (see Section 7.4.1). In addition to HR, BR, SpO_2, SBP, and temperature, future data

fusion models should also incorporate temporal changes in vital-sign variability so as to increase the discriminatory power of the model.

In chapter 7, 24-hour variability was calculated. Track-and-trigger samples in the upper GI ward are typically recorded every four to five hours. In order to calculate variability for shorter time intervals, the sampling rate of the vital-sign observations should be increased. To elaborate, if there are only two track-and-trigger observations in an interval of 8 hours, then 8-hourly variability cannot be estimated reliably. However as discussed in the previous section, combined continuous and track-and-trigger monitoring may be performed in wards where both modes of monitoring are implemented. Therefore, where available, continuous data (which are typically recorded as often as every 20 seconds) may be used, in order to include variability in the data fusion model. The vital signs can be resampled to obtain one sample per minute, for example, and the total minute-by-minute distance travelled by uni-dimensional and multi-dimensional vital-sign vectors in a 30-minute window can be computed, as described in Section 7.4.1. This will result in a variability signal with a frequency of 30 minutes. A sample-and-hold should be used to hold variability values for 30 minutes, such that all the signals (i.e., the vital signs and the variability) in the data fusion model have data during the same intervals. Alternatively a 30-minute sliding window may be used, in which case the latter sample-and-hold step is not required.

Appendix A

Normal-Gamma Conjugate Distributions for Uni-dimensional Data

A.1 Conjugate Prior and Posterior

Following Equation 8.11, the $\mathcal{NG}$ conjugate prior can be expanded as

$$p(\mu, \lambda) = \frac{1}{Z}\lambda^{\frac{1}{2}} exp\left(-\frac{\kappa_0 \lambda}{2}\left(\mu - \mu_0\right)^2\right)\lambda^{\alpha_0 - 1} exp\left(-\lambda\beta_0\right) \tag{A.1}$$

where $Z = \frac{\Gamma(a_0)}{b_0^{a_0}}\left(\frac{2\pi}{\kappa_0}\right)^{\frac{1}{2}}$[112].

By virtue of conjugacy, the $\mathcal{NG}$ posterior follows the same form as the prior in Equation 8.11, and may be expressed as

$$p(\mu, \lambda|\mathbf{x}_N)\ \mathcal{N}(\mu|\mu_N, \frac{1}{\kappa_N \lambda})\mathcal{G}(\lambda|\alpha_N, \beta_N) \tag{A.2}$$

where $\mu_N = \frac{\beta_0\mu_0 + N\bar{\mathbf{x}}}{\beta_0 + N}$,

$\kappa_N = \kappa_0 + N$,

$\alpha_N = \alpha_0 + N$,

$\beta_N = \beta_0 + \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})^2 + \frac{\beta_0 N(\bar{\mathbf{x}} - \mu_0)^2}{2(\beta_0 + N)}$

$\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$

Derivation of the above hyper-parameters may be found in [112].

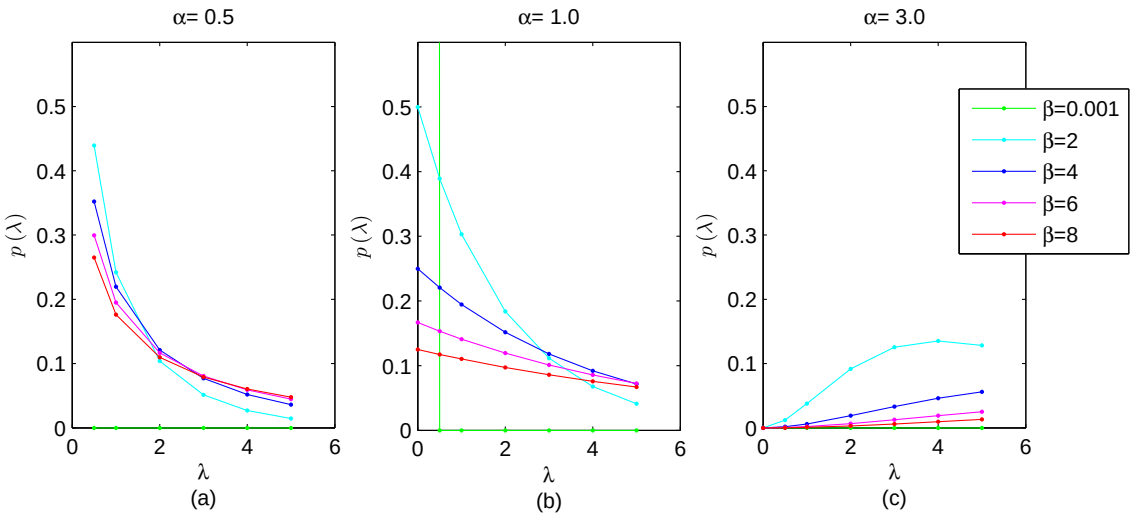


Figure A.1: $p(\lambda)$ for different combinations of α and β.

A.2 Hyper-Parameters

From Equation 8.10, the distribution over λ

$$p(\lambda) \sim \mathcal{G}(\lambda \mid \alpha, \beta) \tag{A.3}$$

$$p(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta} \tag{A.4}$$

where α describes the "shape" and β describes the "scale" of $p(\lambda)$. $p(\lambda)$ has the following properties [14]:

$\lambda_{mean} = \frac{\alpha}{\beta}$,

$\lambda_{mode} = \frac{\alpha-1}{\beta}$, and

$Var[\lambda] = \frac{\alpha}{\beta^2}$

Figure A.1 presents the relationship between α and β and their influence on λ. Note that the shape of the distribution changes with α, and the scale or the range covered changes with β.

To illustrate the influence of a_0 and b_0 on $p(\mu, \lambda)$, Figure A.2 illustrates sample models for different combinations of α and β. Note how the shape of the model, and of course its position along the y-axis, changes with the selection of α and β.

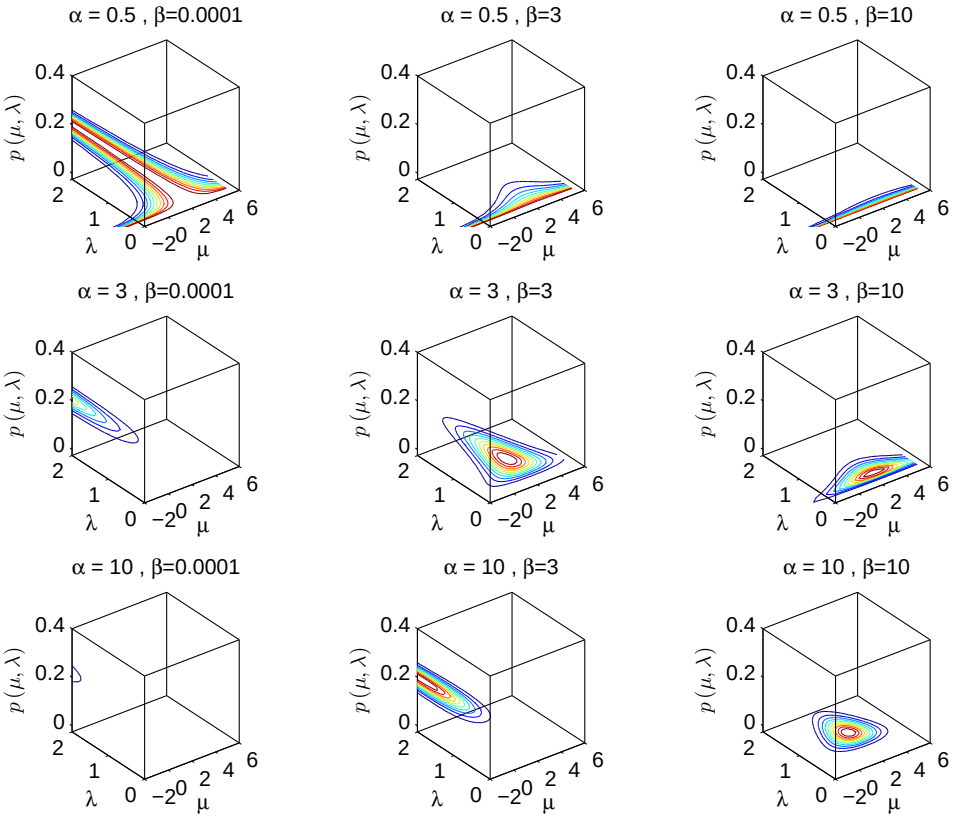


Figure A.2: Effect of a_0 and b_0 on $p(\mu, \lambda)$.

Appendix B

Normal-Wishart Conjugate Distributions for Multi-dimensional Data

B.1 Conjugate Prior and Posterior

The Bayesian modelling process for multi-dimensional data is the same as in the uni-dimensional case. The form of the distributions taken by candidate models for multi-dimensional data is the multi-dimensional analogue of the Normal-Gamma ($\mathcal{NG}$) distribution, called the Normal-Wishart ($\mathcal{NW}$) distribution, which is to define both the posterior and conjugate prior distributions.

Let $X = [\mathbf{x}_1 ... \mathbf{x}_N]$ be a d-dimensional vector where the parameters $\Theta = \{\boldsymbol{\mu}, \Lambda\}$ of the $\mathcal{NW}$ distribution are the mean $\boldsymbol{\mu}$ =$[\mu_1\ \mu_2]$ and the precision matrix Λ=$\begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}$.

$$p(\mathbf{x}|\mu, \Lambda, H_i) = (2\pi)^{-(Nd)/2} |\Lambda|^{N/2} \exp\left(-\frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}) \Lambda (\mathbf{x}_n - \boldsymbol{\mu})' \right) \tag{B.1}$$

$$p(\boldsymbol{\mu}, \Lambda) = \mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\kappa\Lambda)^{-1}\right) \mathcal{W}\left(\Lambda|T\right) \tag{B.2}$$

and the posterior $p(\boldsymbol{\mu}, \Lambda|\boldsymbol{D})$ is

$$p(\mu, \Lambda) = \mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{\mu}_N, (\kappa_N \Lambda)^{-1}\right) \mathcal{W}\left(\Lambda|T_N\right) \tag{B.3}$$

where

$$\boldsymbol{\mu}_N = \frac{\kappa_0 \boldsymbol{\mu}_0 + N\bar{\mathbf{x}}}{\kappa_0 + N} \tag{B.4}$$

$$T_n = T + S + \frac{\kappa_0 N}{\kappa_0 + N}\left(\boldsymbol{\mu}_0 - \bar{\mathbf{x}}\right)\left(\boldsymbol{\mu}_0 - \bar{\mathbf{x}}\right)' \tag{B.5}$$

$$S_n = \sum_{n=1}^{N}\left(x_n - \bar{\mathbf{x}}\right)\left(x_n - \bar{\mathbf{x}}\right)' \tag{B.6}$$

$$v_N = v_0 + N \tag{B.7}$$

$$\kappa_N = \kappa_0 + N \tag{B.8}$$

and the posterior marginals are given by the multi-dimensional t-distribution $p\left(\boldsymbol{\mu}|D\right)$

$$p\left(\boldsymbol{\mu}|D\right) = t_{v_N - d + 1}\left(\boldsymbol{\mu}|\boldsymbol{\mu}_n, \frac{T_N}{\kappa_N (v_N - d + 1)}\right) \tag{B.9}$$

and the Wishart distribution

$$p\left(\Lambda|D\right) = \mathcal{W}_{v_N}\left(T_N\right) \tag{B.10}$$

The Normal-Wishart formulae given above are adapted from [112].

B.2 Multi-dimensional Extension

Algorithm 8.1 can easily be extended to computing the evidence of a hypothesis for a multi-dimensional dataset, where Θ_i is correspondingly multi-dimensional. Step 3 of algorithm 8.1 is no longer simple to visualise or compute as there is now a five-dimensional parameter set $\Theta_i = \{\mu_1, \mu_2, \Lambda_{11}, \Lambda_{12} = \Lambda_{21,}\Lambda_{22}\}$.

In Section 8.7.1 it was shown that $P(\mathbf{x}_{source} \mid H_{source}) \approx 1$. Then, for some new data

$\mathbf{x}_{test}$, the evidence of the same hypothesis is $P(\mathbf{x}_{test} \mid H_{source})$, and one can express the ratio of the evidence of a model before and after observing this test data as

$$\frac{P(\mathbf{x}_{test} \mid H_{source})}{P(\mathbf{x}_{source} \mid H_{source})} = \frac{\sum_{j=1}^{J} \left[\left(p(\mathbf{x}_{test} \mid \Theta_i^j, H_i) p(\Theta_i^j \mid H_i) \right) \times A_j \right]}{\sum_{j=1}^{J} \left[\left(p(\mathbf{x}_{source} \mid \Theta_i^j, H_i) p(\Theta_i^j \mid H_i) \right) \times A_j \right]} \tag{B.11}$$

$$= \frac{A_j \times \sum_{j=1}^{J} p(\mathbf{x}_{test} \mid \Theta_i^j, H_i) p(\Theta_i^j \mid H_i)}{A_j \times \sum_{j=1}^{J} p(\mathbf{x}_{source} \mid \Theta_i^j, H_i) p(\Theta_i^j \mid H_i)} \tag{B.12}$$

and since $P(\mathbf{x}_{source} \mid H_{source}) \approx 1$, Equation B.12 can be simplified to

$$P(\mathbf{x}_{test} \mid H_{source}) = \frac{\sum_{j=1}^{J} p(\mathbf{x}_{test} \mid \Theta_i^j, H_i) p(\Theta_i^j \mid H_i)}{\sum_{j=1}^{J} p(\mathbf{x}_{source} \mid \Theta_i^j, H_i) p(\Theta_i^j \mid H_i)}. \tag{B.13}$$

In order to verify the above derivation of Equation B.13, a simple test was performed. Again, for ease of visualisation, the test was performed using uni-dimensional synthetic data. $P(\mathbf{x}_{test} \mid H_{source})$ was calculated using Equation B.13 for (a) $\mathbf{x}_{test} = [0]$, (b) $\mathbf{x}_{test} = [2]$, and (c) $\mathbf{x}_{test} = [10]$. The synthetic hypothesis $H_{source} \sim NG(\mu, \lambda \mid \mu_0 = 0, \lambda_0 = 1)$. As $\mu_{0_{source}} = 0$ and $\lambda_0 = 1$, between the three test cases, this hypothesis is most highly expected to be the "true" hypothesis for $\mathbf{x}_{test} = [0]$, and so it is expected that the evidence $p(\mathbf{x}_{test} \mid H_{source})$ will be the highest when $\mathbf{x}_{test} = [0]$. This is illustrated in Figure B.1 (repeated from Figure 8.9) which shows $p(\mathbf{x}_{test} \mid H_{source})$ for cases (a), (b), and (c). As expected, the hypothesis has the highest evidence (i.e. $P(\mathbf{x}_{test} \mid H_{source}) = 1$) when $\mathbf{x}_{test}$=0, and the evidence for H_{source} decreases to $P(\mathbf{x}_{test} \mid H_{source}) \approx 0.35$ when test data $\mathbf{x}_{test} = [2]$, and further to $P(\mathbf{x}_{source} \mid H_{source}) \approx 0$ when $\mathbf{x}_{test} = [10]$, indicating that $\mathbf{x}_{test} = [0]$ is the most likely, and $\mathbf{x}_{test} = [10]$ is the least less likely, to have been generated by the uni-dimensional Normal corresponding to hypothesis H_{source}.

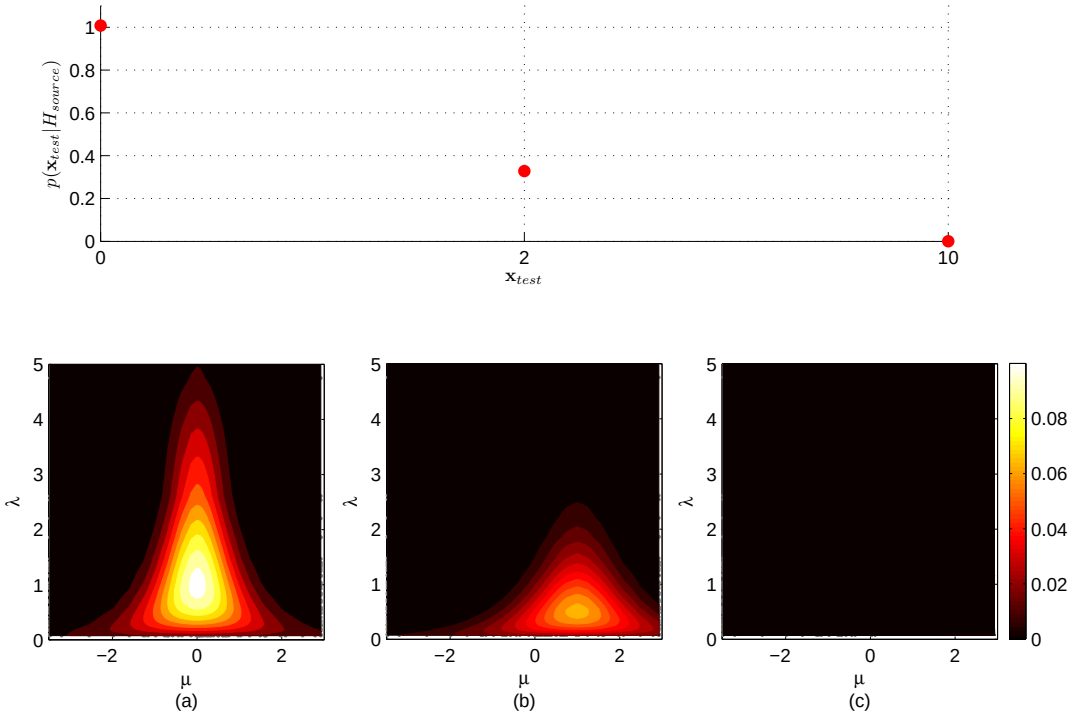


Figure B.1: Evidence for an example hypothesis after seeing example artificial data. The upper plot shows $p(\mathbf{x}_{test} \mid H_{source})$ at the corresponding value of $\mathbf{x}_{test}$. The lower plots show the corresponding posterior distribution of the hypothesis after observation of each test data, for (a) $\mathbf{x}_{test} = [0]$, (b) $\mathbf{x}_{test} = [2]$, and (c) $\mathbf{x}_{test} = [10]$.

Appendix C

Last Day of CALMS-2 EICU Patients

Here, vital-sign time-series for 6 of the 19 patients in the CALMS-2 dataset who had emergency ICU admissions are shown. In Figures C.1 to C.6, (a) shows the track-and-trigger values of the HR, BR, Temperature, and SBP on the left side y-axes, and SpO_2 on the right side y-axes, (b) the posterior probability $P(\mathbf{x}_{new} \mid H_i)$ for 5 randomly sampled hypothesis H_i. and (c) the corresponding BMN novelty score $z(\mathbf{x}_{new})$. The time-series for patients are identified by their CALMS-2 study IDs. The period shown corresponds to the last 24 hours, counting back from the time of the last observation before emergency admission to the ICU.

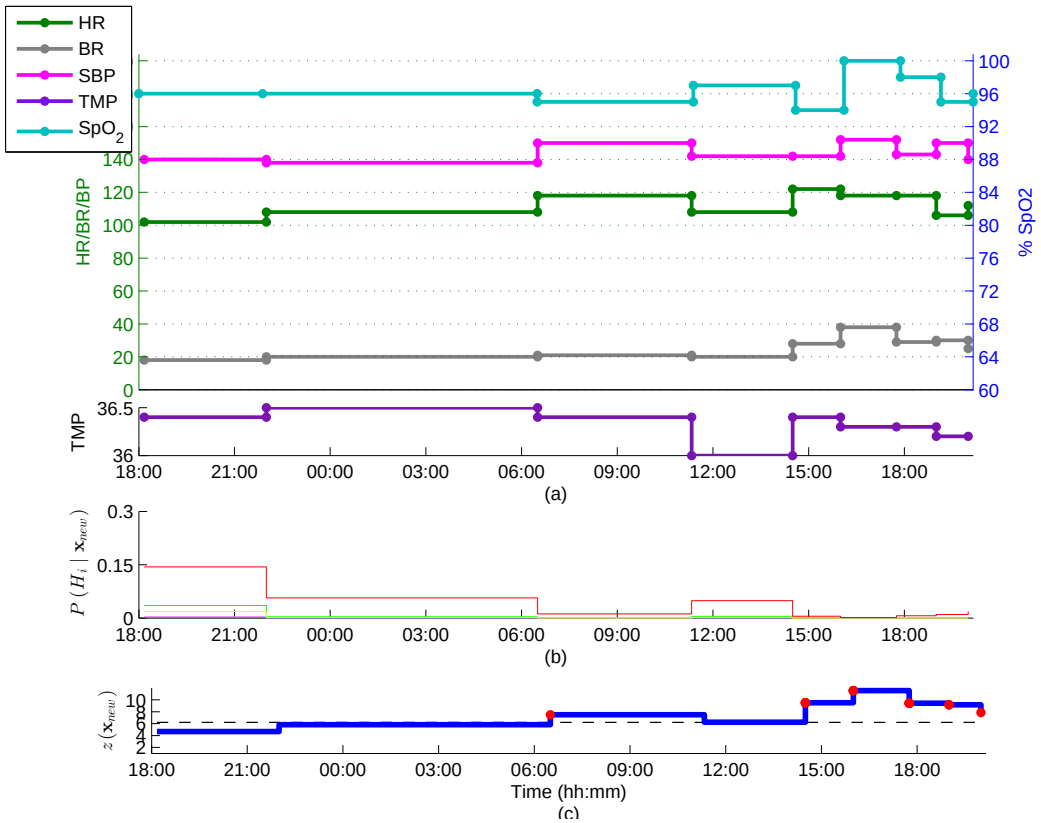


Figure C.1: Study ID 125

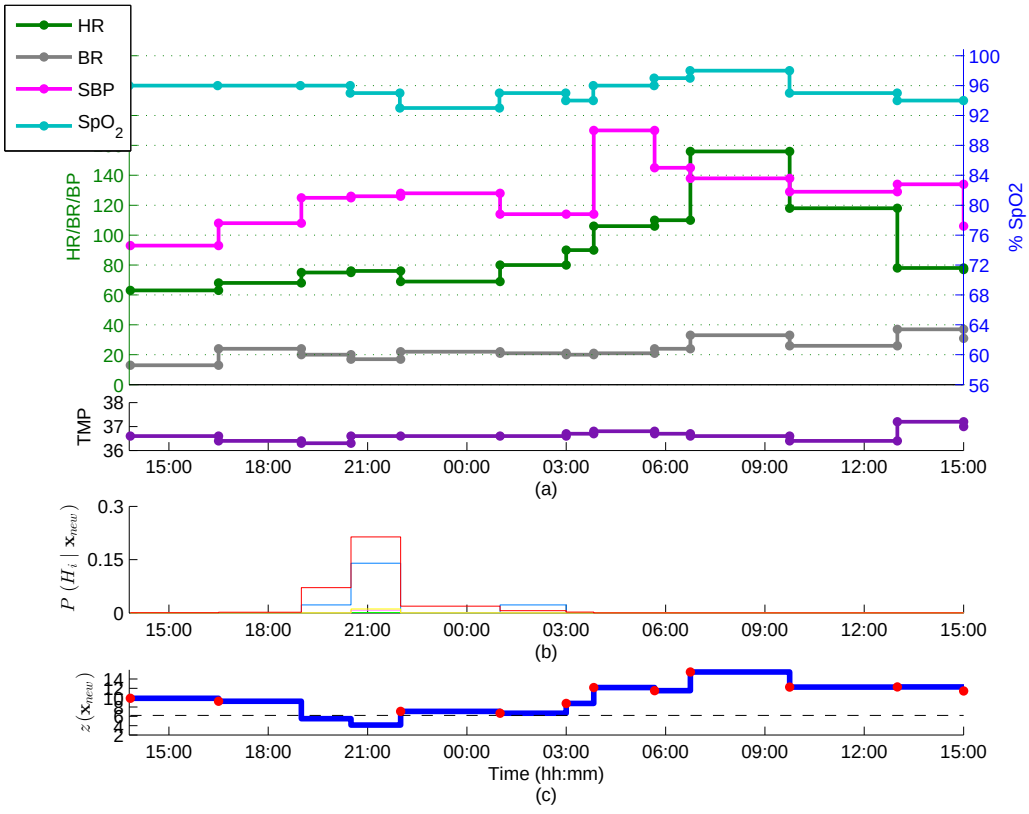


Figure C.2: Study ID 54

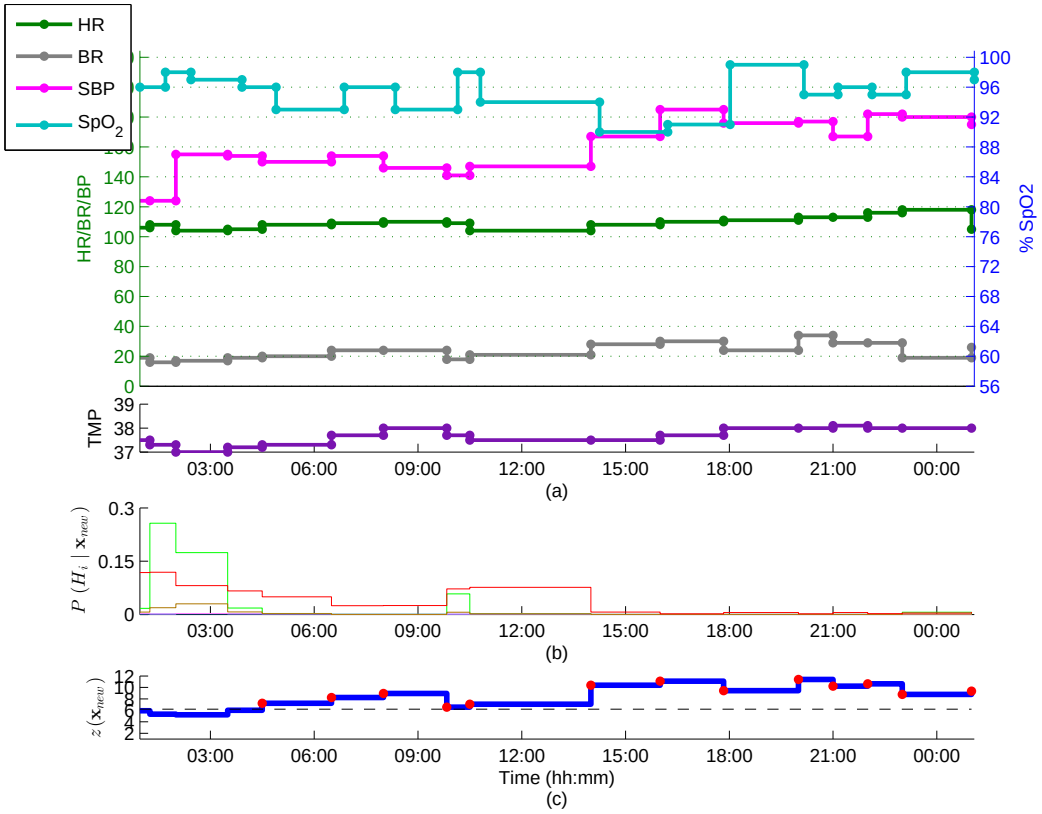


Figure C.3: Study ID 68

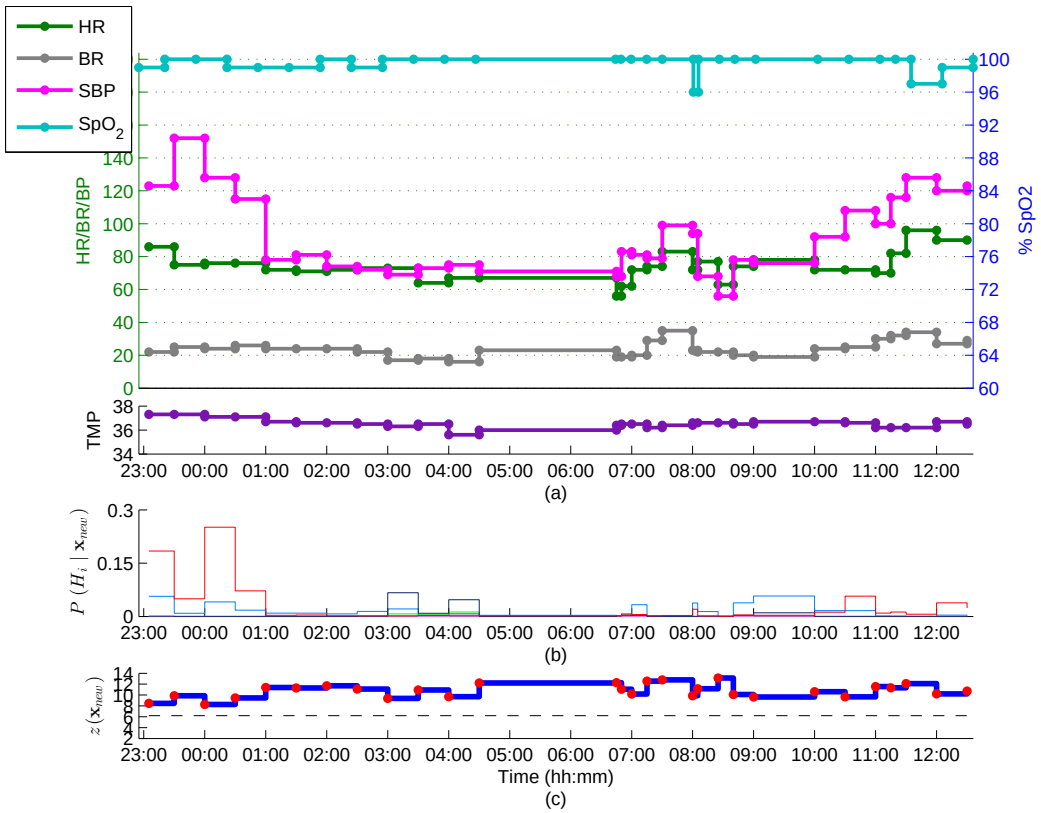


Figure C.4: Study ID 155

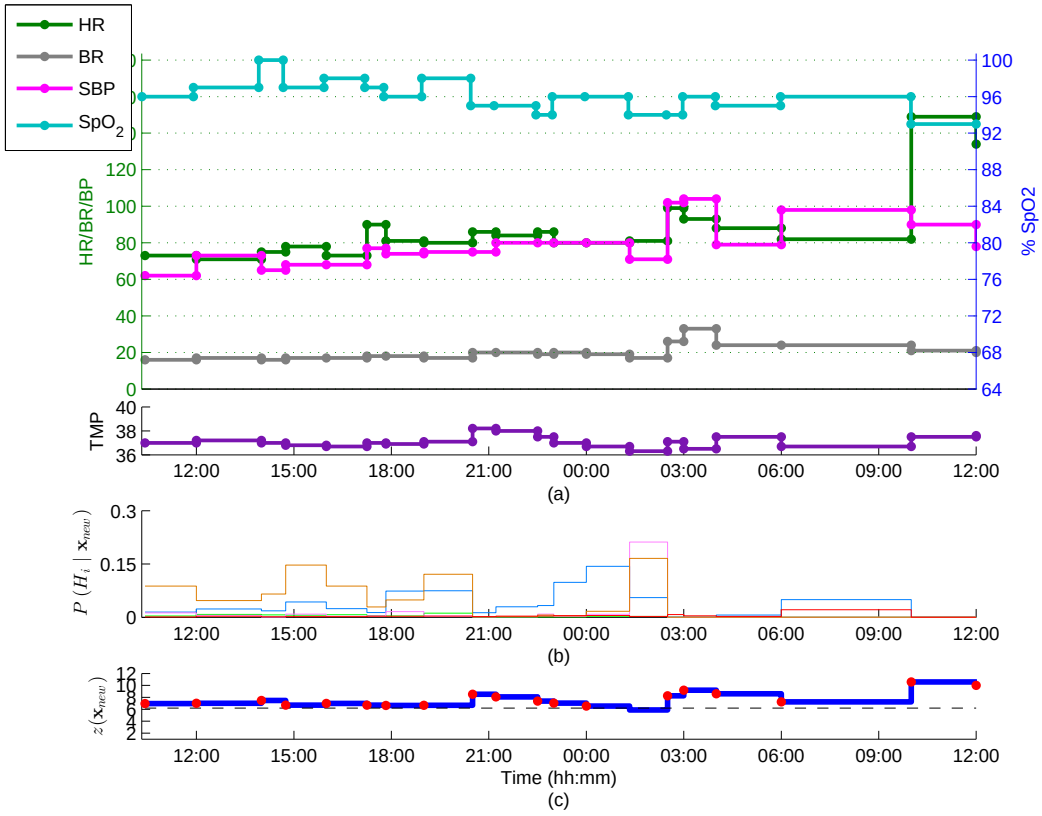


Figure C.5: Study ID 175

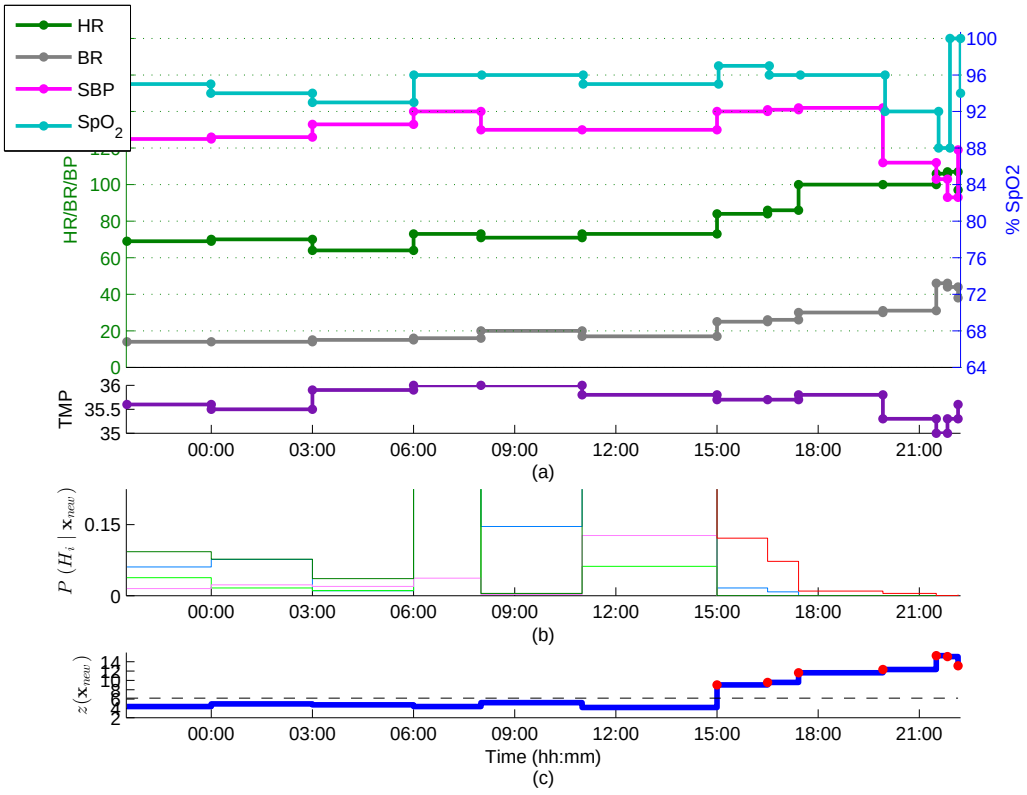


Figure C.6: Study ID 177

Bibliography

[1] J P Babu, G D Jindal, A C Bhuta, and G B Parulkar. Impedance plethysmography: basic principles. *Journal of Postgraduate Medicine*, 36(2):57–63, April 1990.

[2] O Badawi and M Breslow. Readmissions and death after ICU discharge: development and validation of two predictive models. *PloS One*, 7(11):e48758, 2012.

[3] C Ball, M Kirkby, and S Williams. Effect of the critical care outreach team on patient survival to discharge from hospital and readmission to critical care: non-randomised population based study. *BMJ*, 327(7422):1014, 2003.

[4] P Barach and S Small. Reporting and preventing medical mishaps: lessons from non-medical near miss reporting systems. *BMJ: British Medical Journal*, 320(7237):759, 2000.

[5] T Bardell, J F Legare, K J Buth, G M Hirsch, and I S Ali. ICU readmission after cardiac surgery. *European journal of cardio-thoracic surgery*, 23(3):354–359, 2003.

[6] K Beaumont and R Thomson. Deterioration in hospital patients: early signs and appropriate actions. *Nursing Standard*, 23(1):43–48, 2008.

[7] S E Bedell, D C Deitz, D Leeman, and T L Delbanco. Incidence and characteristics of preventable iatrogenic cardiac arrests. *JAMA*, 265(21):2815–2820, 1991.

[8] R Bellomo, D Goldsmith, S Uchino, J Buckmaster, G Hart, H Opdam, W Silvester, L Doolan, and G Gutteridge. Prospective controlled trial of effect of medical emergency team on postoperative morbidity and mortality rates*. *Critical Care Medicine*, 32(4):916–921, 2004.

[9] R Bellomo, D Goldsmith, S Uchino, J Buckmaster, G K Hart, H Opdam, W Silvester, L Doolan, and G Gutteridge. A prospective before-and-after trial of a medical emergency team. *The Medical Journal of Australia*, 179(6):283–7, September 2003.

[10] G Berlot, A Pangher, L Petrucci, R Bussani, and U Lucangelo. Anticipating events of in-hospital cardiac arrest. *European Journal of Emergency Medicine*, 11(1):24–28, 2004.

[11] J M Bernardo and A F M Smith. *Bayesian theory*, volume 405. Wiley. com, 2009.

[12] D M Berwick, D Calkins, C McCannon, and A Hackbarth. The 100 000 lives campaign. *JAMA*, 295(3):324–327, 2006.

[13] C Berzuini, R Bellazzi, S Quaglini, and D Spiegelhalter. Bayesian networks for patient monitoring. *Artificial intelligence in medicine*, 4(3):243–260, 1992.

[14] C M Bishop. *Pattern recognition and machine learning (information science and statistics)*. Springer, 2007.

[15] O Blatchford and S Capewell. Emergency medical admissions: taking stock and planning for winter. *BMJ*, 315(7119):1322–1323, November 1997.

[16] I Blunt, M Bardsley, and J Dixon. Trends in emergency admissions in England 2004 - 2009 : is greater efficiency breeding inefficiency? Technical Report July, 2010.

[17] Y Borhani. *Vital Sign Monitoring and Data Fusion in Haemodialysis*. PhD thesis, University of Oxford, 2013.

[18] S Brown, S Ratcliffe, J Kahn, and S Halpern. The epidemiology of intensive care unit readmissions in the United States. *American Journal of Respiratory and Critical Care Medicine*, 185(9):955–964, 2012.

[19] M Buist, E Jarmolowski, P Burton, S Bernard, B Waxman, and J Anderson. Recognising clinical instability in hospital patients before cardiac arrest or unplanned admission to intensive care. a pilot study in a tertiary-care hospital. *The Medical Journal of Australia*, 171(1):22, 1999.

[20] C Campbell. Kernel methods: a survey of current techniques. *Neurocomputing*, 48(1-4):63–84, October 2002.

[21] Centre for Clinical Practice. Nice guidelines: Acutely ill patients in hospital. Technical report, National Institute of Clinical Excellence, 2007.

[22] M Chambrin et al. Alarms in the intensive care unit: how can the number of false alarms be reduced? *Critical Care*, 5(4):184–188, 2001.

[23] M Chambrin, P Ravaux, D Calvelo-Aros, A Jaborska, C Chopin, and B Boniface. Multicentric study of monitoring alarms in the adult intensive care unit (ICU): a descriptive analysis. *Intensive Care Medicine*, 25(12):1360–1366, 1999.

[24] S Charbonnier, G Becq, and L Biot. On-line segmentation algorithm for continuously monitored data in intensive care units. *Biomedical Engineering, IEEE Transactions on*, 51(3):484–492, 2004.

[25] S Charbonnier and S Gentil. A trend-based alarm system to improve patient monitoring in intensive care units. *Control Engineering Practice*, 15(9):1039–1050, 2007.

[26] J Chen, R Bellomo, A Flabouris, K Hillman, S Finfer, et al. The relationship between early emergency team calls and serious adverse events*. *Critical Care Medicine*, 37(1):148–153, 2009.

[27] J Chen, R Bellomo, K Hillman, A Flabouris, and S Finfer. Triggers for emergency team activation: A multicenter assessment. *Journal of Critical Care*, 25(2):359–e1, 2010.

[28] J Chen, K Hillman, R Bellomo, A Flabouris, S Finfer, and M Cretikos. The impact of introducing medical emergency team system on the documentations of vital signs. *Resuscitation*, 80(1):35–43, 2009.

[29] L M Chen, C M Martin, S P Keenan, and W J Sibbald. Patients readmitted to the intensive care unit during the same hospitalization: clinical features and outcomes. *Critical Care medicine*, 26(11):1834–1841, 1998.

[30] C A Chrusch, K P Olafson, P M McMillan, D E Roberts, and P R Gray. High occupancy increases the risk of early death or readmission after transfer from intensive care*. *Critical Care Medicine*, 37(10):2753–2758, 2009.

[31] M Churpek, T Yuen, M Huber, S Park, J Hall, and D Edelson. Predicting cardiac arrest on the wardsin-hospital cardiac arrest predictiona nested case-control study. *CHEST Journal*, 141(5):1170–1176, 2012.

[32] G D Clifford, A Aboukhalil, J X Sun, W Zong, B A Janz, G B Moody, and R G Mark. Using the blood pressure waveform to reduce critical false ECG alarms. *Computers in Cardiology, 2006*, pages 829–832.

[33] D A Clifton. *Novelty detection with extreme value theory in jet engine vibration data*. PhD thesis, University of Oxford, 2009.

[34] D A Clifton, L A Clifton, and L Tarassenko. Patient-specific biomedical condition monitoring for post-operative cancer patients. Avoidable deaths in hospitals condition conitoring with novelty detection. *Biomedical Engineering*, pages 1–10.

[35] D A Clifton, J G Douglas, P S Addison, and J N Watson. Measurement of respiratory rate from the photoplethysmogram in chest clinic patients. *Journal of Clinical Monitoring and Computing*, 21(1):55–61, February 2007.

[36] D A Clifton, S Hugueny, and L Tarassenko. Novelty detection with multivariate extreme value statistics. *Journal of Signal Processing Systems*, 65(3):371–389, August 2010.

[37] D A Clifton, D Wong, L Clifton, R Pullinger, and L Tarassenko. A large-scale clinical validation of an integrated monitoring system in the emergency department. *IEEE Transactions on Information Technology in Biomedicine*, 14(4):835–842, 2013.

[38] L Clifton. *Multi-channel novelty detection and classifier combination*. PhD thesis, University of Manchester, 2007.

[39] L Clifton, D A Clifton, M A F Pimentel, P J Watkinson, and L Tarassenko. Gaussian process regression in vital-sign early warning systems. In *Engineering in Medicine*

and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, pages 6161–6164. IEEE, 2012.

[40] Joint Commission. National patient safety goals, 2009.

[41] C Cortes and V Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[42] C Crump, S Saxena, B Wilson, P Farrell, A Rafiq, and C Silvers. Using bayesian networks and rule-based trending to predict patient status in the intensive care unit. In *AMIA Annual Symposium Proceedings*, volume 2009, page 124. American Medical Informatics Association, 2009.

[43] C Crump, C Silvers, B Wilson, L Schlachta-Fairchild, C Lingley-Papadopoulos, and J Ashley. Using principal component analysis to aid bayesian network development for prediction of critical care patient outcomes. *The Journal of Trauma and Acute Care Surgery*, 71(6):1841–1849, 2011.

[44] M Cullinane, G Findlay, C Hargraves, and S Lucas. An acute problem. *A report of the National Confidential Enquiry into Patient Outcome and Death*, 2005.

[45] P L Davies, R Fried, and U Gather. Robust signal extraction for on-line monitoring data. *Journal of Statistical Planning and Inference*, 122(1):65–78, 2004.

[46] Health Department. Comprehensive critical care: a review of adult critical care services. Technical report, 2000.

[47] M DeVita, R Bellomo, K Hillman, J Kellum, A Rotondi, D Teres, A Auerbach, W Chen, K Duncan, G Kenward, et al. Findings of the first consensus conference on medical emergency teams*. *Critical Care Medicine*, 34(9):2463–2478, 2006.

[48] M DeVita, G Smith, S Adam, I Adams-Pizarro, M Buist, R Bellomo, R Bonello, E Cerchiari, B Farlow, D Goldsmith, et al. Identifying the hospitalised patient in crisis - a consensus conference on the afferent limb of rapid response systems. *Resuscitation*, 81(4):375–382, 2010.

[49] L Donaldson. An organisation with a memory: report of an expert working group on learning from adverse events in the NHS. *Clinical Medicine*, 2(5):452–457, 2002.

[50] R W Duckitt, R Buxton-Thomas, J Walker, E Cheek, V Bewick, R Venn, and L G Forni. Worthing physiological scoring system: derivation and validation of a physiological early-warning system for medical admissions. An observational, population-based single-centre study. *British Journal of Anaesthesia*, 98(6):769–74, June 2007.

[51] R O Duda, P E Hart, and D G Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[52] M Elliott. Readmission to intensive care: a review of the literature. *Australian critical care*, 19(3):96–104, 2006.

[53] L Esmonde, A McDonnell, C Ball, C Waskett, R Morgan, A Rashidian, K Bray, S Adam, and S Harvey. Investigating the effectiveness of critical care outreach services: a systematic review. *Intensive care medicine*, 32(11):1713–1721, 2006.

[54] J Feldman, M Ebrahim, and I Bar-Kana. Robust sensor fusion improves heart rate estimation: clinical evaluation. *Journal of Clinical Monitoring*, 13(6):379–384, 1997.

[55] J Fieselmann and C Helms. Respiratory rate predicts cardiopulmonary arrest for internal medicine inpatients. *Journal of General Internal Medicine*, 8(7):354–360, 1993.

[56] D Fotiadis, A Likas, and V Protopappas. Intelligent patient monitoring. *Wiley Encyclopedia of Biomedical Engineering*, 2006.

[57] C Franklin and J Mathew. Developing strategies to prevent inhospital cardiac arrest: analyzing responses of physicians and nurses in the hours before the event. *Critical Care Medicine*, 22(2):244–7, February 1994.

[58] R Fried, U Gather, and M Imhoff. Online pattern recognition in intensive care medicine. In *Proceedings of the AMIA Symposium*, page 184. American Medical Informatics Association, 2001.

[59] S A Frost, E Alexandrou, Y Bogdanovski, Tand Salamonson, P M Davidson, M J Parr, and K M Hillman. Severity of illness and risk of readmission to intensive care: a meta-analysis. *Resuscitation*, 80(5):505–510, 2009.

[60] H Gao, D Harrison, G Parry, K Daly, C Subbe, and K Rowan. The impact of the introduction of critical care outreach services in england: a multicentre interrupted time-series analysis. *Critical Care*, 11(5):R113, 2007.

[61] H Gao, A McDonnell, D Harrison, T Moore, S Adam, K Daly, L Esmonde, D Goldhill, G Parry, A Rashidian, et al. Systematic review and evaluation of physiological track and trigger warning systems for identifying at-risk patients on the ward. *Intensive Care Medicine*, 33(4):667–679, 2007.

[62] J Gardner-Thorpe, N Love, J Wrightson, S Walsh, and N Keeling. The value of modified early warning Score (MEWS) in surgical in-patients: a prospective observational study. *Annals of the Royal College of Surgeons of England*, 88(6):571–5, October 2006.

[63] A Gelman, J Carlin, H Stern, and D Rubin. *Bayesian data analysis*. Chapman&Hall, Boca Raton, FL, 2003.

[64] D R Goldhill, S A White, and A Sumner. Physiological values and procedures in the 24 h before icu admission from the ward. *Anaesthesia*, 54(6):529–534, 1999.

[65] C Gwinnutt, M Columb, and R Harris. Outcome after cardiac arrest in adults in uk hospitals: effect of the 1997 guidelines. *Resuscitation*, 47(2):125–135, 2000.

[66] D L Hall, S Member, and J Llinas. An introduction to multisensor data fusion. 85(1), 1997.

[67] A Hann. *Multi-parameter monitoring for early warning of patient deterioration*. DPhil thesis, University of Oxford, UK, 2008.

[68] D Harrison, H Gao, C Welch, and K Rowan. The effects of critical care outreach services before and after critical care: a matched-cohort analysis. *Journal of critical care*, 25(2):196–204, 2010.

[69] P Hayton, S Utete, D King, S King, P Anuzis, and L Tarassenko. Static and dynamic novelty detection methods for jet engine health monitoring. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1851):493–514, 2007.

[70] K Hillman, J Chen, R Cretikos, Miand Bellomo, D Brown, Go Doig, S Finfer, A Flabouris, et al. Introduction of the medical emergency team (MET) system: a cluster-randomised controlled trial. *Lancet*, 365(9477):2091–2097, 2005.

[71] K M Hillman, P J Bristow, T Chey, K Daffurn, T Jacques, S L Norman, G F Bishop, and G Simmons. Antecedents to hospital deaths. *Internal Medicine Journal*, 31(6):343–348, 2001.

[72] T Hodgetts, G Kenward, I Vlachonikolis, S Payne, and N Castle. The identification of risk factors for cardiac arrest and formulation of activation criteria to alert a medical emergency team. *Resuscitation*, 54(2):125–131, 2002.

[73] T J Hodgetts, G Kenward, I Vlackonikolis, S Payne, N Castle, R Crouch, N Ineson, and L Shaikh. Incidence, location and reasons for avoidable in-hospital cardiac arrest in a district general hospital. *Resuscitation*, 54(2):115–23, August 2002.

[74] J A Hoeting, D Madigan, A E Raftery, and C T Volinsky. Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401, 1999.

[75] M Hravnak, L Edwards, A Clontz, C Valenta, M Devita, and M Pinsky. Defining the incidence of cardiorespiratory instability in patients in step-down units using an electronic integrated monitoring system. *Archives of internal medicine*, 168(12):1300–1308, June 2008.

[76] C W Hsu, C Chang, and C J Lin. A practical guide to support vector classification. 2003.

[77] S Hugueny, D A Clifton, and L Tarassenko. Probabilistic patient monitoring using extreme value theory. A multivariate , multimodal methodology for detecting patient deterioration. *Critical Care*, 2009.

[78] M Imhoff and S Kuhls. Alarm algorithms in critical care monitoring. *Anesthesia and Analgesia*, 102(5):1525–1537, 2006.

[79] National Confidential Enquiry into Patient Outcome, Death, and GP Findley. *Knowing the risk: a review of the peri-operative care of surgical patients*. NCEPOD, 2011.

[80] National Confidential Enquiry into Patient Outcome, Death, and GP Findley. *Time to intervene? A review of patients who underwent cardiopulmonary resuscitation as a result of in-hospital cardiopulmonary arrest*. NCEPOD, 2012.

[81] A Kaben, F Corrêa, K Reinhart, U Settmacher, J Gummert, R Kalff, and Y Sakr. Readmission to a surgical intensive care unit: incidence, outcome and risk factors. *Critical Care*, 12(5):R123, 2008.

[82] N Kannathal, U Rajendra Acharya, E Y K Ng, S M Krishnan, Lim Choo Min, and Swamy Laxminarayan. Cardiac health diagnosis using data fusion of cardiovascular and haemodynamic signals. *Computer Methods and Programs in Biomedicine*, 82(2):87–96, May 2006.

[83] J Kause, G Smith, D Prytherch, M Parr, A Flabouris, and K Hillman. A comparison of antecedents to cardiac arrests, deaths and emergency intensive care admissions in Australia and New Zealand, and the United Kingdom - the ACADEMIA study. *Resuscitation*, 62(3):275–82, September 2004.

[84] J Kellett. Hospital medicine (part 2): What would improve acute hospital care? *European Journal of Internal Medicine*, 20(5):465–469, 2009.

[85] G Kenward, N Castle, T Hodgetts, and L Shaikh. Evaluation of a medical emergency team one year after implementation. *Resuscitation*, 61(3):257–263, 2004.

[86] W A Knaus, E A Draper, D P Wagner, and J E Zimmerman. APACHE II: a severity of disease classification system. *Critical care medicine*, 13(10):818–29, October 1985.

[87] M Köppen. The curse of dimensionality. In *5th Online World Conference on Soft Computing in Industrial Applications (WSC5)*, pages 4–8, 2000.

[88] E Koski, T Sukuvaara, A Mäkivirta, and A Kari. A knowledge-based alarm system for monitoring cardiac operated patients-assessment of clinical performance. *International Journal of Clinical Monitoring and Computing*, 11(2):79–83, 1994.

[89] A Kramer, T Higgins, and J Zimmerman. Intensive care unit readmissions in us hospitals: Patient characteristics, risk factors, and outcomes*. *Critical Care Medicine*, 40(1):3–10, 2012.

[90] S Kullback and R A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[91] C Langmead. Generalized queries and bayesian statistical model checking in dynamic bayesian networks: Application to personalized medicine. 2009.

[92] C Laramee, L Lesperance, D Gause, and K McLeod. Intelligent alarm processing into clinical knowledge. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6657–6659, 2006.

[93] N Laurens and T Dwyer. The impact of medical emergency teams on ICU admission rates, cardiopulmonary arrests and mortality in a regional hospital. *Resuscitation*, 82(6):707–712, 2011.

[94] P Laursen. Event detection on patient monitoring data using causal probabilistic networks. *Methods of Information in Medicine*, 33(1):111, 1994.

[95] S T Lawless. Crying wolf: false alarms in a pediatric intensive care unit. *Critical Care Medicine*, 22(6):981–985, 1994.

[96] P Lucas. Bayesian analysis, pattern analysis, and data mining in health care. *Current Ppinion in Critical Care*, 10(5):399–403, 2004.

[97] D J Lunn, A Thomas, N Best, and D Spiegelhalter. Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337, 2000.

[98] J Ma and S Perkins. Online novelty detection on temporal sequences. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–618. ACM, 2003.

[99] D C MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

[100] A Mäkivirta, E Koski, A Kari, and T Sukuvaara. The median filter as a preprocessor for a patient monitor limit alarm system in intensive care. *Computer Methods and Programs in Biomedicine*, 34(2):139–144, 1991.

[101] M Markou and S Singh. Novelty detection: a review - part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497, December 2003.

[102] M Markou and S Singh. Novelty detection: a review - part 2: neural network based approaches. *Signal Processing*, 83(12):2499–2521, December 2003.

[103] A McDonnell, L Esmonde, R Morgan, R Brown, K Bray, G Parry, S Adam, R Sinclair, S Harvey, N Mays, et al. The provision of critical care outreach services in england: findings from a national survey. *Journal of critical care*, 22(3):212–218, 2007.

[104] H McGloin, S K Adam, and M Singer. Unexpected deaths and referrals to intensive care of patients on general wards. Are some cases potentially avoidable? *Journal of the Royal College of Physicians of London*, 33(3):255–9.

[105] P McQuillan, S Pilkington, A Allan, B Taylor, A Short, G Morgan, M Nielsen, D Barrett, and G Smith. Confidential inquiry into quality of care before admission to intensive care. *BMJ*, 316(7148):1853–1858, 1998.

[106] W Melek, Z Lu, A Kapps, and W Fraser. Comparison of trend detection algorithms in the analysis of physiological time-series data. *Biomedical Engineering, IEEE Transactions on*, 52(4):639–651, 2005.

[107] P Metnitz, F Fieux, B Jordan, T Lang, R Moreno, and J Le Gall. Critically ill patients readmitted to intensive care units - lessons to learn? *Intensive Care Medicine*, 29(2):241–248, 2003.

[108] A Moon, J F Cosgrove, D Lea, A Fairs, and D M Cressey. An eight year audit before and after the introduction of modified early warning score (mews) charts, of patients admitted to a tertiary referral intensive care unit after cpr. *Resuscitation*, 82(2):150–154, 2011.

[109] F Mora, G Passariello, G Carrault, and J Le Pichon. Intelligent patient monitoring and management systems: a review. *Engineering in Medicine and Biology Magazine, IEEE*, 12(4):23–33, 1993.

[110] R Morgan, F Williams, and M Wright. An early warning scoring system for detecting developing critical illness. *Clinical Intensive Care*, 8(2):100, 1997.

[111] M Moya and D Hush. Network constraints and multi-objective optimization for one-class classification. *Neural Networks*, 9(3):463–474, 1996.

[112] K P Murphy. *Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series)*. The MIT Press, 2012.

[113] N Naeem and H Montenegro. Beyond the intensive care unit: a review of interventions aimed at anticipating and preventing in-hospital cardiopulmonary arrest. *Resuscitation*, 67(1):13–23, 2005.

[114] J Nolan, E Gallagher, L Llyod, and K Rowan. National cardiac arrest audit (NCAA). *The Intensive Care Society*, 10(4):313–315, 2009.

[115] J Nurmi, V Harjola, J Nolan, and M Castren. Observations and warning signs prior to cardiac arrest. should a medical emergency team intervene earlier? *Acta Anaesthesiologica Scandinavica*, 49(5):702–706, 2005.

[116] C Oberli, J Urzua, C Saez, M Guarini, A Ciprianio, B Garayar, G Lema, R Canessa, C Sacco, and M Irarrazaval. An expert system for monitor alarm integration. *Journal of Clinical Monitoring and Computing*, 15(1):29–35, January 1999.

[117] Royal College of Physicians. National early warning score (NEWS): standardising the assessment of acute-ill severity in the NHS. 2012.

[118] M Opio, G Nansubuga, and J Kellett. Validation of the VitalPAC early warning score (ViEWS) in acutely ill medical patients attending a resource-poor hospital in sub-saharan africa. *Resuscitation*, 2013.

[119] A Otero, P Félix, F Palacios, C Pérez-Gandía, and COS Sorzano. Intelligent alarms for patient supervision. In *Intelligent Signal Processing, 2007. WISP 2007. IEEE International Symposium on*, pages 1–6. IEEE, 2007.

[120] L Ott, M Pinsky, L Hoffman, S Clarke, D Clark, Sand Ren, and M Hravnak. Medical emergency team calls in the radiology department: patient characteristics and outcomes. *BMJ*, 21(6):509–518, 2012.

[121] J Pan and W J Tompkins. A real-time QRS detection algorithm. *IEEE transactions on bio-medical engineering*, 32(3):230–6, March 1985.

[122] D Parkinson and A R Liddle. Bayesian model averaging in astrophysics: a review. *Statistical Analysis and Data Mining*, 6(1):3–14, 2013.

[123] R M Pearse, D A Harrison, P James, D Watson, C Hinds, A Rhodes, R M Grounds, and E D Bennett. Identification and characterisation of the high-risk surgical population in the united kingdom. *Critical Care*, 10(3):R81, 2006.

[124] M Peberdy, W Kaye, J Ornato, G Larkin, V Nadkarni, M Mancini, R Berg, G Nichol, and T Lane-Trultt. Cardiopulmonary resuscitation of adults in the hospital: a report of 14 720 cardiac arrests from the national registry of cardiopulmonary resuscitation. *Resuscitation*, 58(3):297–308, 2003.

[125] M A F Pimentel, D A Clifton, L Clifton, P J Watkinson, and L Tarassenko. Modelling physiological deterioration in post-operative patient vital-sign data. *Medical & biological engineering & computing*, pages 1–9, 2013.

[126] M A F Pimentel, D A Clifton, P J Watkinson, L Clifton, and L Tarassenko. The observer effect when monitoring acutely-ill patients. *MEDSIP 2012. 5th International Conference on Medical Signal and Information Processing, Liverpool, UK*, 2012.

[127] G Priestley, W Watson, A Rashidian, C Mozley, D Russell, J Wilson, J Cope, D Hart, D Kay, K Cowley, et al. Introducing critical care outreach: a ward-randomised trial of phased introduction in a general hospital. *Intensive Care Medicine*, 30(7):1398–1404, 2004.

[128] D R Prytherch, G B Smith, P E Schmidt, and Peter I Featherstone. ViEWS - towards a national early warning score for detecting adult inpatient deterioration. *Resuscitation*, 81(8):932–937, 2010.

[129] D R Prytherch, G B Smith, P I Schmidt, Pand Featherstone, K Stewart, D Knight, and B Higgins. Calculating early warning scores–a classroom comparison of pen and paper and hand-held computer methods. *Resuscitation*, 70(2):173–8, August 2006.

[130] J A Quinn and C K I Williams. *Known unknowns: Novelty detection in condition monitoring*. Springer, 2007.

[131] Rasmussen, C E and Williams, C K. *Gaussian processes for machine learning*. The MIT Press, 2006.

[132] J Renton, D V Pilcher, J D Santamaria, P Stow, M Bailey, G Hart, and G Duke. Factors associated with increased risk of readmission to intensive care in australia. *Intensive care medicine*, 37(11):1800–1808, 2011.

[133] K Rich. Inhospital cardiac arrest: pre-event variables and nursing response. *Clinical Nurse Specialist*, 13(3):147–153, 1999.

[134] G Ritter and M Gallegos. Outliers in statistical pattern recognition and an application to automatic chromosome classification. *Pattern Recognition Letters*, 18(6):525–539, June 1997.

[135] J Roberts, T Parlikar, Tand Heldt, and G Verghese. Bayesian networks for cardiovascular monitoring. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pages 205–209. IEEE, 2006.

[136] S J Roberts and L Tarassenko. A probabilistic resource allocating network for novelty detection. *Neural Computation*, 6:270–284, 1994.

[137] M Roland and G Abel. Reducing emergency admissions: are we on the right track? Technical report, September 2012.

[138] A L Rosenberg, T P Hofer, R A Hayward, C Strachan, and C M Watts. Who bounces back? physiologic and other predictors of intensive care unit readmission. *Critical Care Medicine*, 29(3):511–518, 2001.

[139] A L Rosenberg and C Watts. Patients readmitted to ICUs* a systematic review of risk factors and outcomes. *CHEST Journal*, 118(2):492–502, 2000.

[140] R M Schein, N Hazday, M Pena, B H Ruben, and C L Sprung. Clinical antecedents to in-hospital cardiopulmonary arrest. *Chest*, 98(6):1388–92, December 1990.

[141] A P Schneider, D J Nelson, and D D Brown. In-hospital cardiopulmonary resuscitation: a 30-year review. *The Journal of the American Board of Family Practice/American Board of Family Practice*, 6(2):91, 1993.

[142] R Schoenberg, DZ Sands, and C Safran. Making ICU alarms meaningful: a comparison of traditional vs. trend-based algorithms. page 379, 1999.

[143] B Scholkopf, J C Platt, J Shawe-Taylor, A J Smola, and R C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–71, July 2001.

[144] E Seward, E Greig, S Preston, RA Harris, Z Borrill, TD Wardle, R Burnham, P Driscoll, BDW Harrison, DC Lowe, et al. A confidential study of deaths after emergency medical admission: issues relating to quality of care. *Clinical Medicine*, 3(5):425–434, 2003.

[145] J T Sharpley and J C Holden. Introducing an early warning scoring system in a district general hospital. *Nursing in Critical Care*, 9(3):98–103.

[146] A F Smith and R J Oakey. Incidence and significance of errors in a patient track and trigger system during an epidemic of Legionnaires' disease: retrospective casenote analysis. *Anaesthesia*, 61(3):222–8, March 2006.

[147] A F Smith and J Wood. Can some in-hospital cardio-respiratory arrests be prevented? a prospective survey. *Resuscitation*, 37(3):133–137, 1998.

[148] G B Smith, D Prytherch, P Meredith, P E Schmidt, and P I Featherstone. The ability of the national early warning score (news) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation*, 2013.

[149] G B Smith, D Prytherch, P Schmidt, P Featherstone, and B Higgins. A review, and performance evaluation, of single-parameter track and trigger systems. *Resuscitation*, 79(1):11–21, 2008.

[150] G B Smith, D R Prytherch, P Schmidt, P I Featherstone, D Knight, G Clements, and M Mohammed. Hospital-wide physiological surveillance–a new approach to the early identification and management of the sick patient. *Resuscitation*, 71(1):19–28, 2006.

[151] G B Smith, D R Prytherch, P E Schmidt, and P I Featherstone. Review and performance evaluation of aggregate weighted 'track and trigger' systems. *Resuscitation*, 77(2):170–9, May 2008.

[152] N Stevens, A Giannareas, V Kern, A Viesca, M Fortino-Mullen, A King, and I Lee. Smart alarms: multivariate medical alarm integration for post cabg surgery patients. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 533–542. ACM, 2012.

[153] C P Subbe, R G Davies, E Williams, P Rutherford, and L Gemmell. Effect of introducing the modified early warning score on clinical outcomes, cardio-pulmonary

arrests and intensive care utilisation in acute medical admissions. *Anaesthesia*, 58(8):797–802, August 2003.

[154] T Sukuvaara, E M Koski, A Makivirta, and A Kari. A knowledge-based alarm system for monitoring cardiac operated patients–technical construction and evaluation. *International Journal of Clinical Monitoring and Computing*, 10(2):117–26, May 1993.

[155] L Tarassenko. Novelty detection for identification of masses in mammograms. In *4th IEE International Conference on Artificial Neural Networks*, pages 442–447, Perth, Australia, 1995.

[156] L Tarassenko. *Guide to Neural Computing Applications*. Butterworth-Heinemann, 1998.

[157] L Tarassenko, D A Clifton, P R Bannister, S King, and D King. Novelty detection. Wiley Online Library, 2009.

[158] L Tarassenko, D A Clifton, M Pinsky, M Hravnak, J Woods, and P Watkinson. Centile-based early warning scores derived from statistical distributions of vital signs. *Resuscitation*, 82(8):1013–1018, 2011.

[159] L Tarassenko, A Hann, A Patterson, E Braithwaite, K Davidson, V Barber, and D Young. Biosign TM : Multi-parameter monitoring for early warning of patient deterioration. pages 71–76, 2005.

[160] L Tarassenko, A Hann, A Patterson, E Braithwaite, K Davidson, V Barber, and D Young. Biosign TM : Multi-parameter monitoring for early warning of patient deterioration. In *Medical applications of signal processing*, pages 71–76, 2005.

[161] L Tarassenko, A Hann, and D Young. Integrated monitoring and analysis for early warning of patient deterioration. *British Journal of Anaesthesia*, 97(1):64–68, July 2006.

[162] D M J Tax and R P W Duin. Data domain description using support vectors, 1999.

[163] M E Tipping and D Lowe. Shadow targets: a novel algorithm for topographic projections by radial basis functions. *Artificial Neural Networks, Fifth International Conference on (Conf. Publ. No. 440)*, pages 7–12.

[164] R Trinkle and A Flabouris. Documenting rapid response system afferent limb failure and associated patient outcomes. *Resuscitation*, 82(7):810–814, 2011.

[165] C L Tsien. Event discovery in medical time-series data. In *Proceedings of the AMIA Symposium*, page 858. American Medical Informatics Association, 2000.

[166] C L Tsien and J C Fackler. Poor prognosis for existing monitors in the intensive care unit. *Critical care medicine*, 25(4):614–9, April 1997.

[167] C L Tsien, I S Kohane, and N McIntosh. Multiple signal integration by decision tree induction to detect artifacts in the neonatal intensive care unit. *Artificial intelligence in medicine*, 19(3):189–202, July 2000.

[168] V Viallefont, A E Raftery, and S Richardson. Variable selection and bayesian model averaging in case-control studies. *Statistics in medicine*, 20(21):3215–3230, 2001.

[169] A Vlayen, S Verelst, G E Bekkering, W Schrooten, J Hellings, and N Claes. Incidence and preventability of adverse events requiring intensive care admission: a systematic review. *Journal of Evaluation in Clinical Practice*, 18(2):485–497, 2012.

[170] C Wallmuller, G Meron, I Kurkciyan, A Schober, P Stratil, and F Sterz. Causes of in-hospital cardiac arrest and influence on outcome. *Resuscitation*, 83(10):1206–1211, 2012.

[171] E W Weisstein. K-means clustering algorithm – from Wolfram MathWorld. http://mathworld.wolfram.com/K-MeansClusteringAlgorithm.htm.

[172] J G Williams, S E Roberts, M F Ali, W Cheung, D R Cohen, G Demery, A Edwards, M Greer, M Hellier, H Hutchings, et al. Gastroenterology services in the uk. the burden of disease, and the organisation and delivery of services for gastrointestinal and liver disorders: a review of the evidence. *Gut*, 56(suppl 1):1–113, 2007.

[173] B Winters, J Pham, E Hunt, E Guallar, S Berenholtz, and P Pronovost. Rapid response systems: A systematic review*. *Critical Care Medicine*, 35(5):1238–1243, 2007.

[174] M Wolf, M Keel, K von Siebenthal, H Bucher, K Geering, Y Lehareinger, and P Niederer. Improved monitoring of preterm infants by fuzzy logic. *Technology and Health care: Official Journal of the European Society for Engineering and Medicine*, 4(2):193, 1996.

[175] D Wong. *Identifying Vital-Sign Abnormality in Acutely-Ill Patients*. PhD thesis, University of Oxford, 2011.

[176] D Wong, Clifton D A, and L Tarassenko. Validation of a Gaussian Process framework for inference in missing vital-sign data. [In Review].

[177] M W Wukitsch, M T Petterson, D R Tobler, and J A Pologe. Pulse oximetry: analysis of theory, technology, and practice. *Journal of Clinical Monitoring*, 4(4):290–301, October 1988.

[178] P Yang, G Dumont, and J Ansermino. Adaptive change detection in heart rate trend monitoring in anesthetized children. *Biomedical Engineering, IEEE Transactions on*, 53(11):2211–2219, 2006.