Methods and Models for Acute Hypotensive Episode Prediction

Marzyeh Ghassemi

Linacre College

University of Oxford

A thesis submitted in partial fulfillment of the Master of Research in

Biomedical Engineering

November 25, 2011

To my husband and best friend Eric and my research baby Raziyeh; the two most important people in my life, without whom I would not be who I am.

Acknowledgements

In the process of completing this research, I have had the benefit of many supportive people. Listed here is a severely attenuated set of those whom I would like to thank. For all others, know that I am eternally grateful.

Prof. Lionel Tarassenko for being the world's most understanding advisor. Throughout my stay, there was not a single point at which you threw up your hands and gave up on me, which I believe you were totally entitled to. The academic guidance, emotional support and personal advice you gave me are greatly appreciated, and I hope that someday I will be as good a professor for others as you have been for me.

Dr. Gari Clifford for all of his assistance and advice. I have been consistently impressed at your kindness and ingenuity, and I hope that I will be working again with you in the near future.

The entire BioSignals Processing group at the IBME for making my stay a truly enjoyable one. Special recognition must also be made of Christina and Lei, who suffered the worst of research during pregnancy with me. I was happy to find that Madame Zaritska was wrong by 8 lbs about baby size!

Dr. Mic Craig, Dr. George Moody and Dr. Joon Lee of MIT for their support with the MIMIC II database and research in general.

My family and friends for their kind support and encouragement throughout my entire life.

My mother, Maryam Reitan Ghassemi, is my rock, and this experience was no exception to her general pattern of sweetness.

My husband, Eric Munson, has always been the best man for me, and was somehow able to detect when I needed to smile, laugh, cry or yell.

My baby, Razi, whose smile means more to me than anything I have written or could ever write.

Abstract

Methods and Models for Acute Hypotensive Episode Prediction

Marzyeh Ghassemi

Linacre College

Master of Research in Biomedical Engineering

Trinity Term, 2011

Hypotension is both a dangerous condition and a critical marker for the progression of certain conditions and diseases. Additional research has shown that the more quickly a hypotensive state is predicted, the better the probability of survival.

This thesis explores the creation of models to predict an acute hypotensive episode at a minimum of one hour before occurrence. Models were based on data acquired from MIMIC II, a large-scale ICU database. A variety of techniques including Parzen models of normality, logistic regression, and neural networks are used to analyse the relationship between vital signs and the occurrence of hypotension in the general ICU population.

Both univariate and multivariate models were evaluated, and multi-variate neural network models were found to be the best predictors of hypotension in our dataset. The best model was a multi-layer perceptron constructed using values and first-order statistical functions of a patient's mean arterial blood pressure, respiration rate, blood oxygenation level, and heart rate. Best performance occurred at a time of 60 minutes prior to the onset of a patient's first recorded hypotensive episode, and yielded an AUC of 0.84. This corresponded to a classification accuracy of 82% in a balanced dataset and 84% in an unbalanced dataset which more closely reflects the prevalence of hypotensive data in an ICU. These results indicate that there are early patterns underlying general hypotension detectable in patient physiology.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Hypotension (low blood pressure) is both a dangerous condition itself and a critical marker for the progression of certain conditions and diseases. Both isolated and sustained cases of hypotension are associated with higher mortality in patients in intensive care, and studies have shown that hypotension is one of the key predictors of mortality. Hypotension is also a prevalent precursor (and thus, indicator) of many other disorders (haemorrhage, dehydration, abnormal heart rhythms, heart attack, etc.), making its detection a critical first step in intensive care unit (ICU) monitoring. Hypotension has been successfully used by clinicians as a marker for the progression of sepsis (immune-system inflammatory state due to infection), the second most common cause of death in non-coronary ICU patients in the United States, making it an important state to identify early on. Additional research [1, 2] has shown that the more quickly a hypotensive state is predicted the better the probability of survival.

Several recent papers [3–6] have attempted to use standard vital signs recorded from patients in an ICU to predict acute hypotensive episodes (AHE), but most have focused on prediction up to 30 minutes prior to AHE onset. This is often too short for corrective action to be taken. Our goal is to produce a prediction of AHE occurrence of clinical usefulness within the ICU.

This thesis will examine a variety of approaches to attempt AHE prediction. In

the next section of this chapter, the nature of an AHE and the importance of AHE prediction in the ICU are reviewed. The second chapter describes work recently described in the literature related to AHE prediction. The third chapter describes the data sources used to construct our dataset and the dataset itself. The fourth chapter reproduces existing algorithms, benchmarking their results on the new dataset. The fifth chapter then explores long-term AHE prediction using only a patient's blood pressure signals. New variables other than the blood pressure are introduced in chapter six, and final results using the best of these variables are presented in chapter seven.

1.1 Hypotension Definition

Hypotension is a state characterized by low arterial blood pressure (ABP). Blood pressure itself is a measure of the pressure that blood exerts on blood vessel walls as it circulates, and is given in units of millimetres of mercury (mmHg). When blood pressure is measured, two measurements (systolic pressure and diastolic pressure) are recorded for each heartbeat. The systolic blood pressure is pressure at its highest: when the heart beats, pushing blood into arteries. Diastolic pressure is pressure at its lowest: when the heart is at rest between beats and blood is flowing back through the veins. Blood pressure readings are usually given with the systolic reading first, followed by the diastolic reading. Physicians often use a the mean arterial blood pressure during a single cardiac cycle as an indicator of the perfusion pressure seen by bodily organs. Mean ABP can be approximated using systolic and diastolic pressures, by two formulas: At normal resting heart rates the diastolic portion of the cardiac cycle is twice as long as the systolic (i.e. it takes twice as long for ventricles to fill with blood as it takes to pump blood out). Thus mean ABP is calculated as $\textit{mean ABP} = \frac{2}{3}\ \textit{Diastolic BP} + \frac{1}{3}\ \textit{Systolic BP}$.

For adults, the British Journal of Nursing, [7] and the United States National Library of Medicine and National Institute of Health [8] define individuals with a systolic/diastolic blood pressure of 90/60 mmHg or less as having low blood pressure. Normal blood

pressures are defined as those falling within the range of 90/60 mmHg - 130/80 mmHg. However, some individuals may have a lower blood pressure naturally, in which case the changes from the individual's normal state become the most important factor. Depending on the individual, systolic and diastolic blood pressure can vary by as much as 30 to 40 mmHg over 24 hours depending on factors such as time of day, stress levels, and temperature. Sensitivity to fluctuation in blood pressure can vary by individual: in individuals with naturally low blood pressures, a small drop in blood pressure (e.g. 20 mmHg) has been shown to cause transient hypotension. [9]

1.2 Hypotension Causes

Hypotension can be caused by a wide variety of conditions and disorders. In a recent diagnostic trial examining the efficacy of goal-directed ultrasound in determining the cause of non-traumatic emergency department hypotension in 200 patients, physicians used a list of 21 possible diagnoses for each hypotensive case. [10]

The most common root cause of reduced blood pressure is reduced blood volume, or hypovolaemia, stemming from blood loss or insufficient fluid within the body. Common causes of hypovolaemia include haemorrhage (excessive bleeding), excessive vomiting, and diarrhoea, which themselves can be triggered by a variety of circumstances and conditions. Disorders of the adrenal glands can also be problematic, as these glands are responsible for producing hormones like aldosterone, which control the amount of salt and water in the body. If the adrenal glands become damaged and aldosterone production is reduced, the loss of salt can cause hypovolaemia through dehydration. [11]

Cardiac output can also drop without an abnormal blood volume due to issues with the heart itself. Any cardiac anomaly that results in a reduced heart rate or decreased pump ability can cause hypotension. Serious cardiac emergencies like congestive heart failure, arrhythmias (abnormal cardiac electrical activity), myocardial infarction (heart attack), or bradycardia (low resting heart rate) are all known cardiac output depressors,

and can therefore lead to hypotension. [8] Additionally, some medications can slow heart rate and depress pumping ability.

Another major cause of hypotension is disproportionate vasodilation, which results in decreased vascular resistance and, therefore, decreased blood pressure. Vasodilation is caused through the relaxation of the smooth muscle surrounding blood vessels, which controls the widening and narrowing of the blood vessels through the autonomic nervous system (ANS). The ANS can be divided into two parts: the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS). The SNS mobilises the body's response to stress through functions such as accelerating heart rate, constricting blood vessels, and raising blood pressure. The PNS promotes the body's maintenance functions through actions such as decelerating heart rate and dilating the blood vessels. Vasodilation can be caused by brain or spinal cord injury that leads to increased parasympathetic activity or decreased SNS activity. Nervous system disorders such as Parkinson's disease and multiple system atrophy which are associated with decreased SNS activity also induce a side state of hypotension due to blood vessels remaining too wide. [11] Vasodilation is also a by-product of the progression of septic, toxic and anaphylactic shock. Both septic shock and toxic shock are caused by immune system insults through bacteria or other irritants; these bacteria attack the blood vessel walls causing fluid loss from the blood. Anaphylactic shock is caused by an allergic reaction during which the body produces a large amount of histamine, causing disproportionate vasodilation.

Conditions for which low blood pressure is considered to be less critical include postural or orthostatic hypotension (related to position of the body), age-related hypotension, and pregnancy-related hypotension. In other situations, persistent hypotension can be deadly and can move quickly from incidental to fatal. If instigated by reduced blood volume, death can be triggered by insufficient blood flow to critical organs. If caused by underlying heart conditions, hypotension can quickly move into cardiogenic shock (failure of the heart to pump). In a majority of the medical conditions mentioned, hypotension itself is not the underlying cause, but rather an early effect, of illness. The treatment for

hypotension is therefore linked to its cause, making an early diagnosis critical.

1.3 Hypotension as a Predictor of Mortality

Hypotension has been established previously as an important factor in mortality prediction for patients with heart disorders. For a cohort of over 2,400 patients with a recorded pulmonary embolism (blockage of the lung or one of its branches), systolic hypotension was the most significant prognostic factor in 3-month mortality prediction. [12] Similarly, for a cohort of over 41,000 patients with a recorded myocardial infarction, systolic hypotension was the second most significant 30-day mortality predictor following patient age. [13] In a cohort of 102 patients with a recorded return of spontaneous circulation (ROSC) after cardiac arrest (inability of the heart to contract), systolic hypotension was associated with a significantly higher mortality rate within 6 hours after ROSC (83% vs. 58%) and early exposure to hypotension was a strong, independent predictor of death. [14]

1.4 Hypotension as a Precursor to Septic Shock

Sepsis is the second most common cause of death in non-coronary ICU patients in the United States, and the tenth most common cause of death overall. [15] Traditionally sepsis has been used as an umbrella term to describe a state of inflammatory response caused by an infection by many possible agents (bacteria, virus, mould, etc.) [16]. The initial infection triggers a hyper-inflammatory state in which the immune system is oversensitive and, unable to control the infection, attacks body tissue in its over-zealousness. [17] The transition point from sepsis to severe sepsis occurs when the immune system's response causes hypotension and insufficient blood flow (hypoperfusion), leading to organ dysfunction.

In a study carrier out by Kumar et. al. [1], the case for early detection of serious sepsis was made stronger: they found that duration of hypotension before treatment was the

critical determinant of mortality in septic shock cases. Patients treated within the first hour of displayed hypotension had increased survival rates, and mortality rates increased with the progressing lengths of delay. This presents a strong case for the early detection of hypotension as a possible method of preventing septic shock deaths.

1.5 Physiological Basis for Early Hypotension Detection

To achieve early detection of hypotension, there must be early physiological signs which present. In the case of hypotension stemming from hypovolemia, this is unlikely given that the loss of fluid causing the episode is usually due to environmental conditions (e.g. blood loss stemming from an injury). [18] This is also true in cases of hypotension stemming from serious cardiac anomalies, where the decreased pumping action of the heart is sudden.

However, in cases of hypotension stemming from non-sudden cardiac anomalies (bradycardia), there may be longer-term (greater than 30 minutes prior to onset) physical signs of future hypotension. Bradycardia can occur as a result of abnormal activity from the heart's sinus node or blockage of the electrical impulses travelling from the atria to the ventricles. [19] Both of these causes can be experienced in varying degrees of severity, and the amount of time a condition may be measurably present in a patient before a hypotensive episode is experienced can vary from minutes to weeks.

In the case of hypotension due to disproportionate vasodilation, the timelines involved are often longer than 30 minutes as the root causes are often a result of nervous system disorders. Disproportionate vasodilation is a result of either the SNS not working to constrict blood vessels or the PNS over-dilating blood vessels. [11] When this is caused by brain or spinal cord injury, the associated time to hypotensive onset can vary from hours to a week. [20] In the case of hypotension caused by progression of septic, or toxic shock, the amount of time between immune system insult and the point of hypotension

varies based on the initial insult type. For example, gram-positive sepsis (sepsis caused by organisms which weigh a gram or more) usually progresses more slowly than gram-negative sepsis, which can be rapidly fatal in hours. [21] Finally, in hypotension stemming from anaphylactic shock, disproportionate vasodilation is a result of the body's reaction to large amounts of histamine. The body's reaction to histamine is fast and, while severity of symptoms can vary, the onset of cardiac arrest can occur within 1 minute to 6 hours depending on the allergen trigger. [22]

1.6 Problem Definition

Studies of possible hypotension predictor often focus on the short-term changes to patients just prior to onset. [4, 23] However, predictions made no more than 30 minutes prior to AHE onset are likely to simply confirm what medical staff already know, and provide little novel information. Models which require a great deal of data in order to make a prediction may also be unrealistic, as there are often short decision windows in an ICU setting. Additionally, in the case of hypotensive patients, treatment for the underlying causes can range from the administration of fluids and pressor medication to more critical interventions. [11] Finally, even in the most simple cases where pressor medication[1] may be all that is needed, the medications have full activation times ranging from 17 to 25 minutes for full effect. [24]

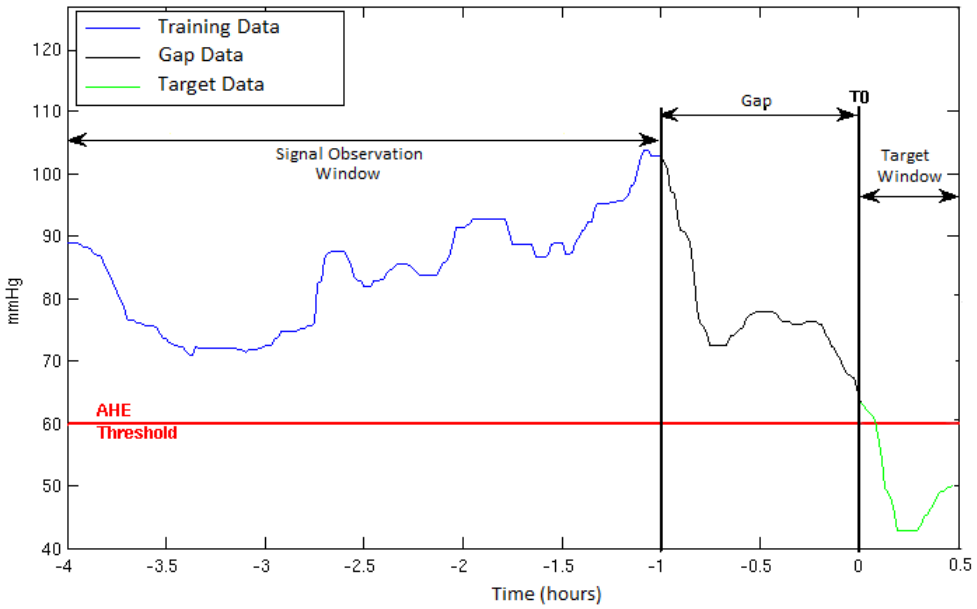
In order to create an AHE predictor with the earliest detection window possible, our aim is to identify data sources with at least six hours of mean ABP data prior to AHE onset.[2] The patient data is split into three segments for analysis, consisting of the following (see Figure 1.1):

- a 30-minute target window which identifies the patient as hypotensive or non-hypotensive

[1] "Pressor" is the general medical term for any medication that can be used to elevate blood pressure by increasing cardiac output and/or constricting systemic vasculature.

- a variable length gap in the time between prediction and target window onset (one hour to three hours prior to target)
- a 180-minute (3 hour) signal observation window preceding the prediction time (offset by the gap)

Figure 1.1: *Data Segmentation*



The data available in the signal observation window is the only data used to predict whether an AHE is likely to occur in the target window. Gap sizes are allowed to vary from one minute to three hours. Given that subsequent low blood pressure events are correlated with previous incidents, our efforts focus on identifying only the first AHE experienced by any patient.

In dataset construction, an AHE was defined as any segment of 30 minutes or greater in which the mean ABP was below 60 mmHg for at least 80% of readings. This ensures that any defined 30-minute AHE contains at least 24 minutes of hypotensive ABP measurements.

[2]Data sources and notation are more thoroughly covered in Chapter 3.

Chapter 2

Review of Previous Work

Possible systems for AHE detection include those which have been tailored to AHE prediction specifically, and those which are intended to provide detection of a supergroup that includes hypotension in general.

2.1 Commercial Monitoring

There are many bedside monitoring systems currently available which provide an alarm in response to the blood pressure falling below a pre-set level (e.g. Philips IntelliVue patient monitors), There are also systems which currently focus on the prediction of future clinical states (e.g. cardiovascular instability) based on vital signs such as blood pressure and respiration rate. [25] However, no system currently exists that can predict hypotensive onset or identify the causative conditions.

2.2 Short-term AHE Detection

Most attempts at hypotension prediction have focused on detection using information derived from the patient's blood pressure signal just prior to hypotensive onset (often a very short time before the onset).

In 2001, Bassale [23] examined the shape of ABP waveforms taken from invasive

beat-by-beat blood pressure measurement devices (e.g. an intra-arterial catheter). These measurements differ from the intermittent blood pressure measurements taken by non-invasive devices (e.g. an inflatable sphygmomanometer cuff on the upper arm) in that they are able to provide an indication of how the arterial blood pressure varies throughout the cardiac cycle rather than simply providing a single numeric measurement of ABP (such as the systolic, diastolic or mean pressure). Bassale found a significant change in the shape of ABP waveforms 1 minute and 5 minutes prior to hypotension as shown by the autocorrelation function of the blood pressure time series. Following this, Crespo et. al. [26] examined the variance of the ABP signal and the variance of the ABP wave's slope as relevant features to consider when attempting to predict an acute hypotensive episode. After AHEs had been identified, the five minutes of continuous ABP data preceding the episode were divided into five 20-second sub-segments and various metrics were computed including the beat-to-beat interval, the peak amplitude, the peak-to-valley distance, and the percussion wave slope which is the slope of the ABP waveform as it moves from valley to peak. They found that the variance of the ABP signal itself and the percussion wave slope were the most significant, predicting the occurrence of AHEs in 67% and 73% of the episodes, respectively. In both studies, the short time frame of the prediction window (1-5 minutes prior) would make early intervention impossible, thus providing little treatment value.

More recently, Lehman et. al. [27] proposed an algorithm using similarity-based searching and pattern matching in order to forecast AHEs in MIT's MIMIC II database [28]. The MIMIC II database is a large collection of clinical, numeric and waveform data obtained from ICU patients which is available for cross-sectional studies, longitudinal studies, and data-mining.[1] Lehman et. al. selected patients with a predisposition towards low blood pressure by determining who had been treated with pressor medication. Using physiological measurements (see below) from pressor-treated patients in the database, Lehman et. al. defined two groups: stable (further pressor medication was not needed)

[1]The MIMIC II Database is described in greater detail in Chapter 3.

and unstable (further pressor medication was needed). Heart rate (HR) and mean ABP measured every 60 seconds were processed with median filters, and feature vectors were constructed with one hour of data immediately following pressor withdrawal. Some of the features extracted were 15 minute and 60 minute regression slopes, auto-correlation coefficients, and MAP/HR cross correlation coefficients. A threshold-based K nearest neighbours (KNN) classification rule using a variable threshold $\rho(0 < \rho < 1)$ was used to classify patients as stable or unstable. The KNN algorithm specifies that any unknown object is classified by a majority vote of its K nearest neighbours. Patients were classified as unstable if at least ρK of the KNN were unstable and stable otherwise. The patterns were modelled with a Gaussian Mixture Model (GMM) and similarity was computed using a distance from centre metric. Models were evaluated using sensitivity and specificity, which rely upon the number of correct control predictions (true negatives) and the number of correct abnormal predictions (true positives) to assess model performance.[2] Lehman et. al. reported that their best model ($K = 23$ and $\rho = 0.41$) had a sensitivity of 0.74 and a specificity of 0.60 when using one hour of data immediately after pressor withdrawal.

In 2009, two papers were published by Stell et. al. describing the efforts of the EU funded Avert-IT project towards creating a hypotension alarm system via Bayesian neural networks. The goals of the Avert-IT project are to develop a system to monitor and predict the likelihood, and primary causes, of hypotension events within intensive and high-dependency care units. The system being developed will be based on an integrated, real-time data grid infrastructure drawing datasets from six European clinical centres. Currently, the group has come up with its own definition of a hypotensive event (less than 90 mmHg over 5 minutes), and has created 11 different types of event profiles characterised by the minute-to-minute measurements which occur just before hypotensive onset (e.g. whether an event is "triggered" and sustained by mean ABP only, or by some

[2]Sensitivity is the model's capability to correctly identify positives (abnormal patients), defined as $Sensitivity = \frac{TruePositives}{TruePositives+FalseNegatives}$, and specificity is the model's capability to correctly identify controls (normal patients), defined as $Specificity = \frac{TrueNegatives}{TrueNegatives+FalsePositives}$.

combination of mean ABP and systolic ABP). [29] Future steps include the completion of data collection from the European clinical centres, the construction of a Bayesian artificial neural network (BANN) system, and the physical implementation of this system into an interface that will trigger an alarm alerting care staff. [30] The AVERT-IT website (http://www.avert-it.org) was last updated in August of 2010, and has defined several specific objectives including the automatic and continuous monitoring of at least four patient parameters (ECG, ABP, oxygen saturation, and core temperature) in parallel with the generation of a continuous Hypotension Prediction Index (HPi). The most recent achievements listed on a December 2009 website press release include a defined patient parameter set, the development of functional design for their "HypoPredict Engine", and the creation of tested clinical user interfaces. The project is currently focusing on the finalization of the Avert-IT "Hypo-Net" application, and the distribution of the hardware and software systems to each project clinical trial centre. They note that patients are being actively recruited to the observational phase of the clinical trial in the lead centre. There has been no further update to the status of the project since February 2009 on the AVERT IT wiki (http://venus.nesc.gla.ac.uk/wordpress/), and no further papers since July 2009.

2.3 2009 Computers in Cardiology Competition

Automated hypotension prediction was the subject of the 2009 Computers in Cardiology competition. The goal was to predict whether a patient would develop an AHE given a gap time of 30 minutes before the hypotensive event, based on the vital sign data recorded for that patient. Papers from ten of the 2009 Computers in Cardiology contestants were available for public review, and represented a wide range of possible approaches to AHE prediction. Participants were guaranteed the availability of 10 hours of numerical mean, systolic and diastolic arterial blood pressure (ABPMean, ABPSys, ABPDias) which began 30 minutes prior to T0. The ABP waveform itself was also present in all but two of

the 60 patients. Waveform data are digitized time series of a patient's ABP waveforms acquired by an ICU bedside monitor, and sampled at 125 Hz. Numeric data are real-time ABP measurements derived from the waveforms as they are acquired by the same ICU bedside monitors, and are sampled at 1 Hz. The difference between numeric and waveform data is reviewed in greater detail in Chapter 3.

The goal of this thesis is to predict an acute hypotensive event more than 30 minutes before its onset. Methods which could be extended to meet this goal, even if they were originally designed for a shorter gap (our gaps vary from one to three hours,see Figure 1.1), are reviewed below.

2.3.1 Statistical Arterial Blood Pressure Based Indices

Five of the authors used some combination of statistical feature generation from the arterial blood pressure time series or other available data, and then applied a threshold or classification rule to the resulting features. [4, 6, 31–33] For example, one algorithm used the median of all ABP readings between 90 and 30 minutes prior to onset as the single value for AHE prediction, while another used a linear combination of ABP change in the 30 to 60 minutes and 60 to 90 minutes prior to T0.

The best-performing algorithms were submitted by Chen et. al. [4], and all indices were based on either the numerical or waveform arterial blood pressure signal. Below, each index is listed with the algorithm used to generate it. Another top-performer's algorithm (Mneimneh et. al.'s [6]) was exactly the same as that used in Chen et. al.'s Index I, with the average taken over a longer window of 20 minutes rather than five. A summary of the algorithms used is presented in Table 2.1.

Note that the calculations performed to derive Index I and II are the same, as the latter vary only in the type of data being selected (numeric vs. waveform). Accordingly, the scatterplot of these points (Figure 2.1) shows a clear linear relationship with a near 1:1 correlation ($y = 0.97x + 3.2$).

The four points that are some distance away from the line of identity (ordered from

Table 2.1: *Statistical Index Algorithms*

Label	Algorithm
Index I	Arithmetic mean of the five most recent minutes of numeric ABPMean data 30 minutes prior to T0
Index II	Arithmetic mean of the five most recent minutes of waveform ABP data 30 minutes prior to T0
Index III	Exponentially weighted moving average of the 10 hours of numeric ABPMean data prior to T1
Index IV	Predicted value of numeric ABPMean data at T0 via simple linear regression of the most recent hour of numeric ABPMean data 30 minutes prior to T0
Index V	Arithmetic mean of the five most recent minutes of numeric ABPDias data 30 minutes prior to T0
Index VI	Voting strategy between Index II and Index V; an AHE is predicted only if predicted by both Index II and Index V

smallest value of Index II to largest) may indicate that Index II is more sensitive to fluctuations in the waveform data or, more likely, that there are artefacts in the data.

Index III is the exponentially weighted moving average (EWMA) of 10 hours of ABP-Mean data 30 minutes before T0. Exponentially weighted averages are calculated by weighting the data so that values further back in the past with respect to the current one are given an exponentially decaying amount of importance. [34]

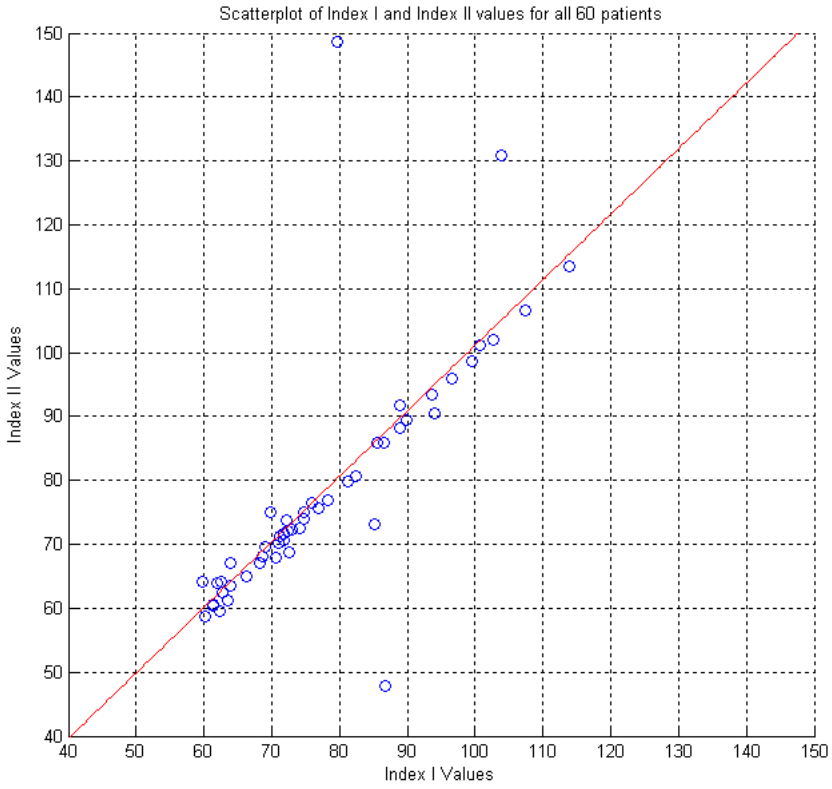
$$EWMA_t = \alpha Y_{t-1} + (1-\alpha) EWMA_{t-1}, t = 1, 2, \ldots, n \tag{2.1}$$

where

- $EWMA_0$ is the mean of the past 1.2 hours of data (see below for a discussion of the time constants used)
- Y_t is the observation at time t
- n is the number of observations to be monitored including $EWMA_0$
- $0 < \alpha <= 1$ is a constant that determines the depth of memory of the EWMA.

If we now substitute the equation for $EWMA_{t-1}$ (derived by setting $t = t - 1$ in

Figure 2.1: *Correlation Between Index I and II*



Equation 2.1) back into Equation 2.1, we obtain the result:

$$EWMA_t = \alpha Y_{t-1} + (1-\alpha)[\alpha Y_{t-2} + (1-\alpha)EWMA_{t-2}] \tag{2.2}$$

$$= \alpha Y_{t-1} + \alpha(1-\alpha)Y_{t-2} + (1-\alpha)^2 EWMA_{t-2} \tag{2.3}$$

Using this process we can substitute for $EWMA_{t-2}$, then for $EWMA_{t-3}$, etc. until we reach $EWMA_0$, which is just the mean over the initial 1.2 hours of data. It can be shown that this expanding equation can therefore be written as:

$$EWMA_{t-2} = \alpha \sum_{i=1}^{t-2} (1-\alpha)^{i-1} Y_{t-i} + (1-\alpha)^{t-2} EWMA_0, \quad t >= 2 \tag{2.4}$$

The exponential behaviour is characterised by the weights $\alpha(1-\alpha)^t$ in Equation 2.4 decreasing geometrically. Based on a property of geometric series, this means that the

summation of these weights must be 1. This is shown in Equation 2.5:

$$\alpha \sum_{i=0}^{t-1} (1-\alpha)^i = \alpha \frac{1-(1-\alpha)^t}{1-(1-\alpha)} = 1-(1-\alpha)^t \tag{2.5}$$

Given this fact, it can be shown that the contribution of each term to the final value $EWMA_t$ becomes smaller with each progressive step back in time, leading to a true exponential smoothing effect. In the exponential sum of (2.4), the time period is 1.2 hours. This is the time constant chosen by Chen et. al. [4], which was reported to give the optimal results in terms of classification performance on the training data. If α is expressed in terms of N time periods, where $\alpha = \frac{2}{N+1}$, N is 72 as we span over 1.2 hours and the time interval in the numerical data is 60 seconds. This gives a value for α of 0.027.

Index IV is the predicted value of ABPMean at T0 (hypotensive onset). This prediction is obtained from linear regression of the final hour of ABPMean numeric data 30 minutes prior to T0. The weights (β values in Equation 2.6 below) are then used to find a linear approximation of the value of the signal at any time after prediction onset. The value of the signal Y_{t+30} is represented by the following equation:

$$Y_{t+30} = \sum_{i=1}^{60} \beta_i * Y_{t-i} \tag{2.6}$$

Therefore, $Y_{T_{pred}+30} = \sum_{i=1}^{60} \beta_i * Y_{T_{pred}-i}$, where T_{pred} represents the time in minutes of prediction onset and i represents the sample number (at a rate of once per minute for the training data).

2.3.2 Probability Density Function and Curve Approximation

Three further papers used neural networks, histograms, and cubic splines to approximate the ABP signals or their probability density function (PDF), and then predict AHEs from the approximations.

One approach used neural network multimodels in order to predict the waveform which would occur in the hour after T0. [5] The author then used the definition of a hypotensive event provided by the challenge organisers to determine whether a patient would experience a hypotensive episode. The second approach involved cubic splines (piecewise polynomial curves), which were first used to approximate the mean ABP curves over intervals of 10.9 minutes. [35] Once curves were fitted, the corresponding coefficients were used as a discrete version of the continuous mean ABP signal: first each time interval's coefficients were evaluated across all training data to determine their ability to distinguish between hypotensive and control patients using a simple ranking algorithm. Then the best subset of these coefficients across 59 of the 60 patients in the training data were chosen for a logistic regression model by examining which values produced the highest number of correct classifications. In a final approach, histograms were generated from the mean ABP values every minute. [36] A 30-minute running window advancing by one minute was used to bin 10 hours of data from the training data set into histograms, producing a total of 600 running histograms for each patient. As a result, the authors observed that between 1.88% to 17.18% of hypotensive mean ABP training data was composed of values that were below 60 mmHg. An AHE was predicted for a patient in the test set based upon whether the percentage of ABP values in the abnormal pressure range (below 60 mmHg) was more than 17.18% of the entire record.

The neural network approach gave the best classification results in one of the Competition tasks, and represents a more complex approach than that which was taken by other competitors. The prediction of AHEs was based on generalized regression neural network (GRNN) [37] models, which are a variant of artificial neural networks (ANN) using radial basis functions (RBF) [38]. An ANN is a complex system comprised of non-linear processing units, neurons, and weights which allow the system to construct a non-linear mapping from input to output. ANNs are trained to learn the mappings using examples (training set), and are then evaluated on an independent test set. Artificial neural networks have now been applied to many classification problems, including fraud

detection [39], handwriting recognition [40], and mortality prediction [41].

Radial Basis Function Neural Networks (RBFNN) were introduced as a form of neural network by Powell [38] in the 1980s and have since been used in pattern recognition, clustering analysis, and other fields. The radial basis function (RBF) for each neuron in an RBFNN consists of as many dimensions as there are input variables, and is characterised by its localisation (centre) and radius (also referred to as a spread). RBF centre locations are chosen through a variety of methods, such as randomly sampling from amongst the training examples or using clustering techniques.

Generalized regression neural networks (GRNN) were proposed by Specht [37] in the 90s, specifically to address regression tasks. GRNNs are a type of radial basis function networks known for their fast learning and function-approximation capabilities. However, while GRNNs are quick to train, they tend to be large and slow in implementation. As with other pattern recognition methods, GRNNs suffer from the curse of dimensionality: as extra dimensions are added the complexity of the model scales exponentially. The GRNN is essentially a kernel-based probability density function (PDF) approximation method implemented in a neural network architecture. For regression problems, this means that the output of the network is regarded as the expected value of the model at a given point in input-space. This expected value is related to the joint PDF of the output and inputs. In order to perform PDF estimation through kernel-based approximation, functions (RBF units) are located at each available data point and added together to estimate the overall PDF. Kernel functions are typically Gaussians and, if sufficient training points are available, this method will yield an arbitrarily good approximation to the true PDF. The principal difference between an RBFNN and a GRNN is the number of neurons: RBFNN are usually composed of far fewer neurons than training points while GRNNs have one neuron for each training point.

Mathematically, GRNNs are an extension of general regression. Let $f(x, y)$ represent the known joint continuous PDF of a vector variable, $\mathbf{x}$, and a scalar variable, y. Let $\mathbf{X}$ and Y represent particular measured values of $\mathbf{x}$ and y. The regression of y on $\mathbf{X}$

(conditional mean of y given $\mathbf{X}$) is then given by:

$$E(y|\mathbf{X}) = \frac{\int_{-\infty}^{\infty} y\ f(\mathbf{X}, y)\ dy}{\int_{-\infty}^{\infty} f(\mathbf{X}, y)\ dy} \tag{2.7}$$

When $f(x, y)$ is unknown, it can be estimated from sample observations of $\mathbf{x}$ and y. The estimators proposed by Parzen [42] have been shown to provide good non-parametric estimates of the PDF given two assumptions: the underlying density is continuous and the first partial derivatives of the function are small for all $\mathbf{x}$. [43] The Gaussian-based probability estimator $\hat{f}(\mathbf{X}, Y)$ can then be specified for given sample values $\mathbf{X}_i$ and Y_i of the random variables $\mathbf{x}$ and y, where M is the number of sample observations and P is the dimension of the vector variable $\mathbf{x}$. (Equation 2.8) The probability estimate is the sum of all sample probabilities, where each sample probability is centred on a sample $\mathbf{X}_i$, Y_i at a chosen width σ.

$$\begin{aligned}\hat{f}(\mathbf{X}, Y) =& \frac{1}{(2\pi)^{P/2}\sigma^P} \\ & \cdot \frac{1}{M} \sum_{i=1}^{M} e^{-\frac{(\mathbf{X}-\mathbf{X}_i)^T(\mathbf{X}-\mathbf{X}_i)}{2\sigma^2}} \\ & \cdot e^{-\frac{(Y-Y_i)^2}{2\sigma^2}}\end{aligned} \tag{2.8}$$

For simplification, the scalar function D_i^2 is then defined as the squared distance between training points and the point of prediction. (Equation 2.9) D_i^2 is used as a measure of how well each training sample ($\mathbf{X}$) can represent the prediction point, $\mathbf{X}_i$.

$$D_i^2 = (\mathbf{X} - \mathbf{X}_i)^T(\mathbf{X} - \mathbf{X}_i) \tag{2.9}$$

Combining the joint probability estimate $\hat{f}$ (Equation 2.8) and the scalar distance metric D_i^2 (Equation 2.9) for substitution into the conditional mean (Equation 2.7) yields

the estimated conditional mean of y given $\mathbf{X}$, $(\hat{Y}(\mathbf{X}))$. [3]

$$\hat{Y}(\mathbf{X}) = \frac{\sum_{i=1}^{M} Y_i \; e^{-\frac{D_i^2}{2\sigma^2}}}{\sum_{i=1}^{M} \; e^{-\frac{D_i^2}{2\sigma^2}}} \tag{2.10}$$

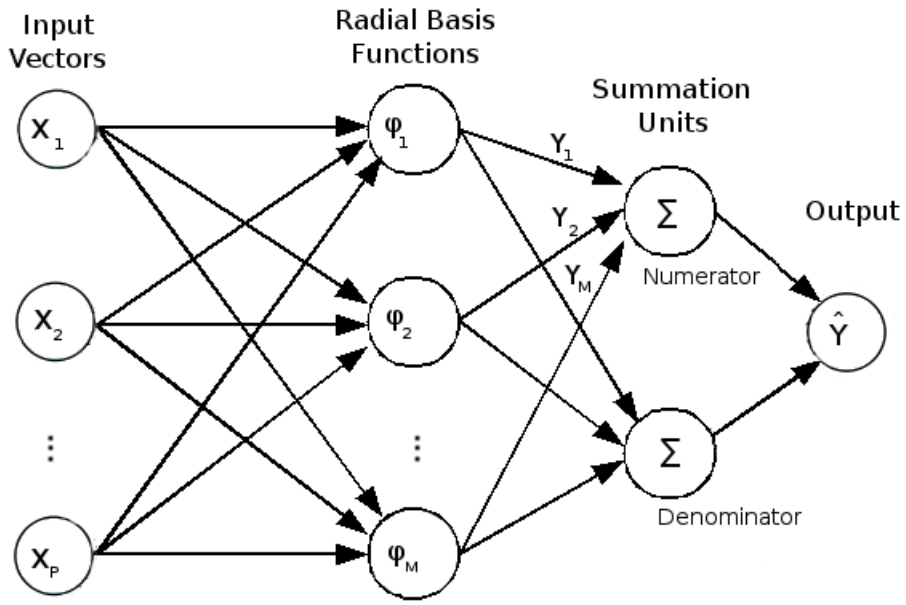
Functionally, GRNNs create predictions by locating Gaussian kernel functions at each training point; in this way, GRNNs can be thought of as RBFNNs in which a hidden unit is centred at every training input. GRNNs consist of a total of 4 layers: input layer, first hidden layer, second hidden layer, and output layer. (See Figure 2.2) The first hidden layer in the GRNN contains the radial units and operates just as the radial basis layer for RBF networks. The second hidden layer consists of units with the network's targets, $Y_1, \ldots, Y_M$, and estimates its output as a weighted average of the first hidden layer outputs, where weighting is proportional to the closeness of the training point to the point which is being estimated. In regression problems, typically only a single output is estimated, and so the second hidden layer usually has two units: the numerator and denominator summations from Equation 2.10. The weighted average is obtained by first calculating a weighted sum for each output of the first hidden layer and then dividing by the sum of the weighting factors. Finally, the output layer performs the actual divisions and reports the result as the predicted value.

2.4 Long-term AHE Prediction

Most recently, a paper by Lee et. al. [3] explored an AHE predictor for intensive care units based on heart rate and blood pressure time series. Lee's predictor differs from the previously described attempts in that it used a much larger cohort of data in order to attempt to provide predictions up to four hours prior to AHE onset. A total of 1,357 records were compiled for analysis, creating 130,325 control and 3,953 hypotensive examples consisting of a 30 minute target window, a gap between prediction time and

[3]Note that in Equation 2.10 the order of integration and summation have been interchanged from Equation 2.7.

Figure 2.2: *Generalised Regression Neural Network*



the target window ranging from one to four hours, and a 30 minute signal observation window. Only the information in the observation window was available to the predictor as input.

The dataset was constructed by using a sliding window within each patient's data to examine whether a given segment was suitable for inclusion as a control or hypotensive example. Therefore, several examples (both control and hypotensive) could come from a single patient. However, distinct examples from the same record were not allowed to be split across the training and testing set, but had to belong exclusively to one or the other. This was done in order to test whether a classifier trained on a sub-set of patients was able to generalise and correctly classify a new, unknown patient.

A total of 45 features were extracted from the HR, mean ABP, and pulse pressure (pulse pressure is the difference between systolic and diastolic blood pressure) time series in the observation window using functions such as the mean, variance, skewness, kurtosis, linear regression slope, and relative energies in different spectral bands. The feature space dimensionality was then reduced via principal component analysis (PCA) to 15 - 16 features. Feed-forward 3-layer artificial neural networks (ANNs) with a single hidden

layer of 20 hidden units and a log-sigmoid activation function were trained to perform non-linear binary classification. Classification performance was evaluated using 5-fold cross-validation, where a distinct ANN was trained for each gap size and fold.

Given the large skew (only 3% of available data was hypotensive) in the available data, training data was balanced by using a randomly sampled subset of the control examples matching the size of the hypotensive examples. However, test data was left unbalanced to reflect the true prevalence of AHEs discovered (this would give a sensitivity of 0 and specificity of 1 if all controls were predicted). Weighted posterior probabilities based on different gap sizes were also investigated, but did not significantly outperform the non-weighted predictor, suggesting that consulting previous predictions was unnecessary. While the positive predictive value[4] of the classifier was low, the ANN performed well on the shortest time-gap, with a best mean area under ROC curve[5] of 0.934 (sensitivity of 0.85 and specificity of 0.86). The performance diminished significantly as the time gap itself increased, suggesting that the autocorrelations of HR and ABP decreases further from the hypotensive event. The authors noted that this algorithm was designed to identify patients at risk for an AHE, rather than to recommend clinical attention through alarm generation. This ensures less disruption to clinical staff, and less likelihood of warnings being ignored. One further concept identified is the fundamental problem with a binary approach to the identification of hypotensive patients. The authors note that constant AHE thresholds (i.e. 60 mmHg) do not take borderline cases or individual variation into account, and suggest that misclassification costs could vary across patients from a clinical perspective.

[5]Positive predictive value measures the proportion of classified positives which have been correctly identified. See Section 4.1.1 for more detail.

[5]ROC curves plot index sensitivity versus 1-specificity as the classification threshold is varied. See Section 4.1.1 for more detail.

2.5 Summary

Table 2.2 lists the previously mentioned methods for hypotensive episode prediction, along with their timeline and performance. Only the two top performers from the 2009 Computers in Cardiology competition are listed (both attributed to Chen et al. [4]). Note that Lee's method is the only method evaluated on a timeline longer than 30 minutes, and has the best prediction outcomes regardless of prediction timeline. Lee's paper was published during the invesitgation of this thesis, and is therfore the most recent result to use as comparison. Thus, Lee's results should be noted as the benchmark in the following Chapters, and will used as a comarison for the final outcomes.

Table 2.2: *Summary of Algorithms*

Author	Algorithm	Timeline	Reported Outcome	Sens.	Spec.
Bassale	ABP shape variance	1-5 min	Significant difference in ABP waveform shape		
Crespo et al.	ABP-derived feature variance	5 min	73% Correct classifications		
Chen et al.	Five-min mean of numeric ABPMean	30 min	82.5% Correct classifications		
Chen et al.	10 hour EWMA of numeric ABPMean	30 min	80% Correct classifications		
Lehman et al.	KNN classification using GMM	1 hour	Prediction of stable/unstable patients	0.74	0.60
Stell et al.	Baysian Neural Networks		Prediction of hypotension likelihood and cause		
Lee et al.	Multi-layer Neural Networks on 45 features	1 hour	Prediction of hypotension likelihood and cause	0.85	0.86

Chapter 3

Data

3.1 Introduction

While there has been progress toward the recognition of AHEs as a critical determinant of patient success, there is currently no AHE prediction system which would work to provide alerts more than an hour prior to hypotensive onset. In this thesis, we define long-term prediction as prediction of the event more than an hour before its onset. The most relevant work is the recent paper by Lee et. al. [3], reviewed in section 2.4, which was able to provide reasonable AHE prediction for a time-gap of one hour prior to hypotensive onset.

Building on this understanding, our goal is to produce an AHE predictor which is capable of producing results within a time frame of one to three hours prior to AHE onset. As noted by Lee et. al., such a system could be the basis for a real-time AHE prediction system applicable to ICU conditions. Given this goal, our efforts have been to create a robust dataset, replicate the most successful prediction methods upon this dataset, and provide new methods based on our own analysis of the results.

Timeseries data of ABP was extracted from MIT's MIMIC II Database in order to include both patients with a recorded AHE and patients without. As of December 2008, the MIMIC II Database contained 2320 complete adult patient records (including

both recorded physiological signals and time series, and accompanying clinical data - see below). Each such record covers at least one entire ICU stay, typically including signals and time series with durations of several days with few or no interruptions. [28]

3.2 Data Selection Criteria

Each patient within the assembled datasets was assigned to group H or group C depending on whether they experienced an AHE (H) or not (C). As noted previously, an AHE was defined as any segment of 30 minutes or greater in which ABP is below 60 mmHg for at least 80% of the data samples. Note that this is different from the definition used for the 2009 Computers in Cardiology competition (see previous chapter) which required at least 90% of data samples below 60 mmHg in any 30-minute segment. Thus AHEs may contain intervals of signal loss or MAP in the normotensive range for up to six minutes in any 30-minute segment. In other words, an AHE contains at least 24 minutes of MAP measurements in the acute hypotensive range.

Data was only taken from patients with at least 6 hours of data before their initial AHE, and with at least 70% of the ABPMean data within the feasible ABP measurement range (10 - 220 mmHg). The exact time that an AHE occurs was set to be the beginning of the target window, and a time T0 was specified such that T0 marks the beginning of the target window. (Note: this means that the AHE, if present, will begin within six minutes of T0.)

3.2.1 Data Collection

For the MIMIC II data, there are three possible types of data available for every record: waveform, numeric and clinical. The process for acquiring this data was as follows:

- The numeric data is acquired by ICU bedside monitors in real-time, from waveforms that are acquired and digitized by these same monitors.
- Some of the digitized waveforms, and some of the derived numerics, are transferred

in real-time to a device that the manufacturer (Philips) calls a "DServer" (data server). The "DServer" collects waveforms and numerics from as many as 64 patients continuously and simultaneously. It is queried by one or more ICU central stations to provide real-time displays of waveforms, vital signs, and detected events.

- For the MIMIC II project, Philips provided a "Data Archiving Agent" (another networked computer) that continuously interrogated the "DServer" in order to capture all of the data streaming through it. This "Data Archiving Agent" is equipped with large removable disk drives that are swapped out on a regular basis. The swapped-out disks, containing 2-3 weeks of data from several ICUs, are mounted on another computer, where the collected data are reformatted from undocumented proprietary formats into documented PhysioBank-compatible[1]formats and de-identified.
- Many recordings span multiple disks, making the matching and joining of these subsystems a necessity; frequently this is not possible because of the technical limitations of the monitors, the "DServer", and/or the "Data Archiving Agent":
  - The monitors are not capable of transferring all of the digitized waveform data and all of the derived numerics simultaneously to the "DServer", however, and the "Data Archiving Agent" is intermittently unable to capture all of the data streamed to the "DServer".
  - As data packets sent from the monitors to the "DServer" do not contain timestamps or sequence numbers, it is very difficult to determine the existence, location, or duration of gaps in the data streams.

A consequence of this complex data collection procedure is that there are often numerics that are unattached to the waveforms from which they were derived (and conversely, waveforms unattached to numerics). Furthermore, the time and amplitude resolutions of the waveforms transmitted from the monitors are coarser than those available within the monitors and used to derive the numerics.

[1]Physiobank is a large collection of digital physiologic signals and related data for use by the biomedical research community. http://www.physionet.org/physiobank/.

3.2.2 Data Types

A patient's waveform data is meant to contain, at a minimum, a time series of the patient's electrocardiogram (ECG) and waveform ABP signals sampled at 125 Hz (1 sample every 8 ms). There can be as many as six additional signals available. The most commonly occurring additional signals are one to eight ECG leads (labelled I, II, III, AVL, AVR, AVF, V, MCL1). ECG leads are defined by where they are placed on the patient's body: I is the left-side chest, II is the left-side upper quadrant, III is the right-side upper quadrant, AVR is the right-side lateral arm, AVL is the left-side lateral arm, AVF is the right-side lateral lower leg, V is the first precordial (chest) lead, and MCL1 is the first modified chest lead.[2] Additional signals include the pulmonary artery pressure (PAP), the photoplethysmogram waveform recorded by a pulse oximiter (PLETH), and the central venous pressure (CVP).

The numeric data contains a time series of the patient's heart rate and the mean, systolic, and diastolic ABP; most records also contain several additional time series, such as respiration rate, a measure of blood oxygen saturation (SpO_2), etc. The vital signs are sampled between 0.0167 Hz (1 sample every 60 seconds) and 1 Hz (1 sample every second).

The clinical data contains information entered into the ICU medical information systems automatically or manually by the ICU staff. It includes records of the medications administered, measurements taken, interventions performed, and clinical outcomes (via ICD9 codes).[3] Of note is that the associated timestamps may be imprecise due to manual entry.

[2]MCL1 is an ECG lead obtained by modifying the lead arrangement for the standard six precordial leads V1-V6. It is commonly used to view the heart on the horizontal plane when specific issues need to be observed, rather than to monitor cardiac rhythm.

[3]The International Classification of Diseases (ICD) is published by the World Health Organisation (WHO) and provides a list of codes for disease classification. The ICD9 codes (established in 1975) map each health condition to a unique code of up to six characters.

3.3 Patient Groups

Two patient groups were created and subsequently divided into training and testing sets. The two groups, or classes, are labelled Hypotensive and Control, and will be abbreviated as H and C. Each group is described below in detail:

Hypotensive Patients (84 in total)

- Patients display their first recorded AHE in the target window (beginning within 6 minutes of the T0 mark)
- Patients may or may not display subsequent AHEs after the target window

Control Patients (84 in total)

- Patients display no recorded AHE in the entirety of their ICU data

Only a single six-hour data window was used from each patient record: either six hours of non-hypotensive data from a patient who never experienced an AHE or six hours of data just prior to the first AHE a patient experienced. As mentioned in Section 1.6, the initial AHE was the target of prediction as subsequent hypotensive events are correlated to the first event experienced. Data was only taken from patients with at least six hours of data before the T0 point (see Figure 1.1), and who demonstrated at least 70% of their ABPMean data within the feasible ABP measurement range (10 - 220 mmHg).The exact time that an AHE occurs (T0) was called the beginning of the 'target window', and all data prior to the target window and added gap was considered to be valid training data. Gap sizes varied from one hour to 3 hours, which allows for just 3 hours of training data when using the largest gap size.

3.4 Data Validation

As summarised in Table 3.1, the new dataset's proportion of ICD9 diagnostic codes and pressor administrations are not completely correlated with the AHE definition used.

3.4.1 Time Prior to T0

While all patients had a minimum of 6 hours of data prior to the T0 mark, some patients did have more recorded data prior to this point. Note that this data was not used, as it could not be gathered for all patients. The mean and median time in hours prior to T0 is specified for both subgroups of patients in Table 3.1.

3.4.2 Pressor Administration

Pressor drugs are sympathomimetic agents that mimic the effects of sympathetic adrenergic activation on the heart (cardiostimulatory drugs) and blood vessels (vasoconstrictor drugs). Some of these drugs increase heart rate and cardiac contractility by stimulating cardiac beta1-adrenoceptors (beta-agonists). Other sympathomimetic drugs increase systemic vascular resistance by stimulating vascular alpha-adrenoceptors (alpha-agonists). Finally, there are non-sympathomimetic, vasoconstrictor drugs, such as vasopressin analogs, that have proven to be effective pressor agents. Pressor admission in the patient population was determined by a MIMIC II database query for records which contained a list of standard vasopressive medications described in the Appendix.

As noted in Table 3.1, the number of patients with pressor medications is higher in the H group than the C group (82.1% vs. 51.2%). This would tend to support our determination of the H group being at a higher risk for hypotensive intervention.

3.4.3 ICD9 Diagnostics

There are seven sub-categories of hypotension which fall under the umbrella ICD-9 diagnosis code for hypotension (458.0 - 458.9). ICD9 code correlation in the patient population was determined by a MIMIC II database query for records which contained at least one ICD9 code within the hypotensive category described in the Appendix.

As seen in Table 3.1, hypotensive ICD9 diagnostics were more prevalent in the H group (13.1% vs. 4.8%), but were not present at a level which might be expected for a

group of patients with identified AHEs. In general, ICD9 codes tend to correlate well with important clinical syndromes and diagnoses but not with detailed physiologic data. ICD9 codes are also created by staff other than the caring clinician. If hypotension is a significant part of the ultimate clinical picture, it is likely that an appropriate ICD9 code would have been assigned. However, the chosen definition for an AHE might be considered insignificant in some clinical settings for coding purposes.

3.4.4 Care Unit Differentiation

In certain cases, patients experiencing hypotension as a result of hypovolaemia (e.g. blood loss or insufficient fluids) versus other causes may need to be excluded from consideration in model generation. Cases of hypovolaemia induced hypotension are easily predictable given a patient's history or appearance, and prediction is therefore of little medical use.

Determining the exact cause of hypotension during an admission would require a combination of examining the ICD-9 codes, nursing notes, and discharge summaries. Patient data could also be examined to determine if there were blood transfusions or fluid administration during the care record. If a patient were to receive a large volume of blood or fluids, it could be assumed that the hypotension was caused by hypovolemia. However, this was difficult to determine given that the first clinical intervention after a drop in blood pressure is fluid resuscitation.

Therefore, the care unit to which each patient was admitted was examined in an attempt to determine the reason why patients in the MIMIC II database were experiencing hypotension. It was assumed that patients admitted to the cardiac surgery care unit (CSRU) and coronary care unit (CCU) would not suffer from hypovolemia as their primary cause of admission is cardiac/vasodilation related. The number of patients in each category with admission to these units are listed in Table 3.1.

The proportion of patients who were admitted to the CSRU or CCU in the H group (70.2%) was much higher than that seen in the overall patient population (9%), which could imply that our chosen patients did not tend to suffer from hypovolemia as their

primary cause of admission.

Table 3.1: *Comparative Dataset Statistics*

	C	H	All MIMIC II Patients[4]
Number of Patients	84	84	25,328
Patients with Hypotensive ICD9 Codes	4 (4.8%)	11 (13.1%)	1,554 (6.1%)
Patients with Pressor Administration	43 (51.2%)	69 (82.1%)	8,693 (34.3%)
Number of In-Hospital Deaths	5 (6.0%)	26 (31.0%)	2,666 (10.5%)
Patients Admitted to CCU/CSRU	62 (73.8%)	59 (70.2%)	2302 (9%)
Mean Time Prior to T0	10	6.4	N/A
Median Time Prior to T0	7	6	N/A

[4]Collective MIMIC II statistics taken from Summary Statistics given for the 2.5 release of MIMIC. (http://mimic.physionet.org/database/releases/70-version-25.html)

3.5 Experimental Method

3.5.1 Data Download

The MIMIC II data can be read in to Matlab files for processing using one of the following methods:

- Use the Physiobank ATM to download .mat files and then use the provided rdsamp function to examine specific sections of each patient record. This requires that the data is downloaded prior to code execution.
- Use the wfdb2mat function to convert segments of wfdb formatted data into .mat files. This also requires that the data is downloaded prior to code execution.
- Use the Matlab WFDB Toolkit to specify which sections from each patient should be downloaded and automatically converted into a Matlab compatible data format.

Given the large size of the files involved, download of the entire record was not deemed suitable. Instead, the Matlab WFDB Toolkit was used to specify which sections of data were needed for any given patient.

3.5.2 Data Pre-processing

Once acquired, each signal is stored as a single row of data. The corresponding .info files contain information on a signal-by-signal basis on the gain, baseline, and units. These are used to convert the signal values from their recorded values of analog-to-digital converter units into their nominal values. The gain for each signal is the number of ADC units that correspond to a single physical unit, and the baseline value is the value of the ADC units that maps to zero in the physical unit space. In Matlab, these ADC values are converted to physical units using the base and gain constants.

In our implementation, basic data pre-processing was performed by an N-point median filter of the data, where N corresponds to a 10 minute window (see below for a more

detailed explanation). The Matlab function "medfilt1" was modified to use the "nanmedian" function rather than the median function. The "median" function is a simple median operator which calculates the median value of a signal over a given number of points. Usage of the "nanmedian" function ensures that the median operation treats non-numerical values as missing values. The "medfilt1" function performs a one-dimensional median N-point filtering of the signal. If N is odd, the new output signal is the nanmedian of the valid $\frac{N-1}{2}$ points on either side of each point. After filtering, data whose values are outside physiologically plausible maxima or minimawas removed.

3.5.2.1 Numeric Data

In the case of the numeric data (sampled at once a minute) we use N = 11. Finally, data which lies outside of the physiologically plausible range for mean arterial blood pressure, systolic arterial blood pressure, and diastolic arterial blood pressure is coded as a non-numeric value. In this case, the valid data ranges presented by Saeed [44] were used as the bounds for data which are given as 10 mmHg $\leq$ ABPMean $\leq$ 220 mmHg, 10 mmHg $\leq$ ABPDias $\leq$ 200 mmHg, and 30 mmHg $\leq$ ABPSys $\leq$ 250 mmHg. An example of numeric filtering can be seen in Figure 3.1.

3.5.2.2 Waveform Data

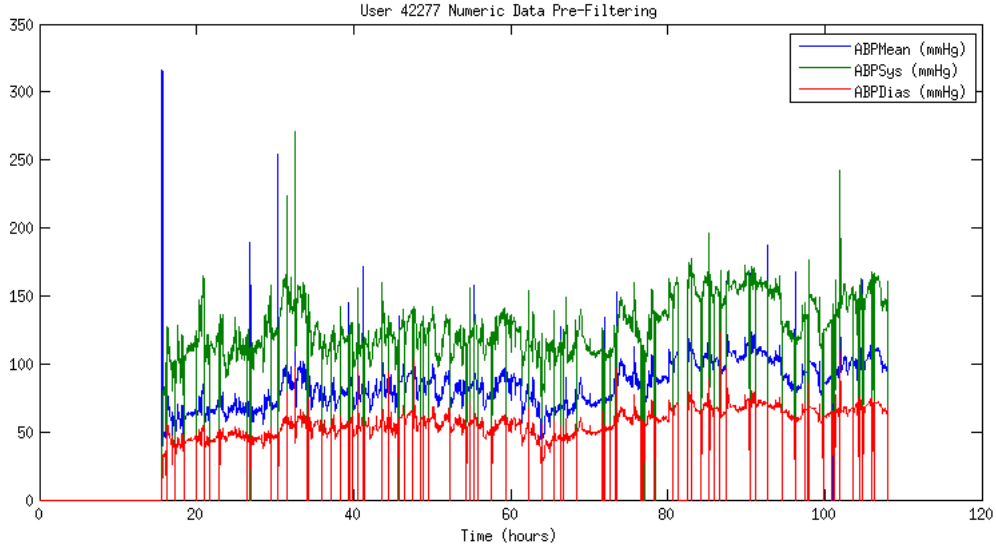
For the waveform data used to generate indices, the original sampling rate was 125 Hz. The data was first downsampled by a factor of five to reduce the number of points for analysis. This results in a sampling rate of 25 Hz, or once every .04 seconds. To prevent aliasing, a low-pass filter is used to reduce the bandwidth of the signal before downsampling. The lowpass filter is a 256-point, Hamming-window based, linear-phase FIR digital filter with a cut-off frequency of 12.5 Hz. Zero-phase digital filtering was used to preserve features in the filtered waveform by processing the input in both the forward and reverse directions.

There are specific high-frequency components (faster than 12.5 Hz) of the ABP wave-

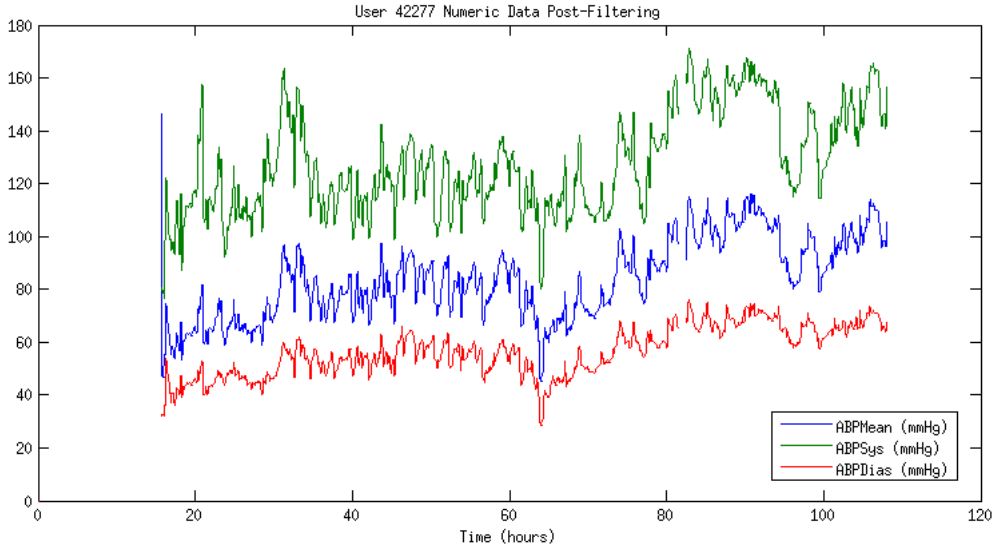
form which will be removed by this filtering. However, given that the signal is not going to be analysed beyond performing an arithmetic mean, the effect of this filtering should be negligible. To confirm this, 100 five-minute segments of ABP waveform data were selected at random from amongst the training data. The mean of these segments was first calculated without any modification to the values, and then after the signal had been low-pass filtered and downsampled as specified above. The difference between the filtered and unfiltered means was never greater than $4.3 * 10^{-04}$ mmHg for any of the 100 samples, which is an insignificant change.

Figure 3.1: *Numeric Data Filtering*

(a) *Pre-filtering*



(b) *Post-filtering*



Chapter 4

Long-term AHE Prediction Using Existing Algorithms

4.1 Introduction

The problem under consideration is the prediction of an acute hypotensive event more than one hour before its onset. As previously shown in Figure 1.1, each patient's data has been segmented into a 30-minute target window in which an AHE may or may not occur, a one to three hour gap time, and a 3 hour signal observation window.

In this Chapter, algorithms which have previously been successfully implemented for other, similar problems are evaluated. Each algorithm is first reviewed in detail independently, and the final section compares the results generated from all three.

In order to assess the validity of algorithms previously reviewed, standard classifier evaluation methods were used and testing/training sets were generated using a data resampling method.

4.1.1 Classifier Evaluation

Our interest is in identifying the best AHE prediction method for an ICU setting, in which data is being gathered and processed as close to real-time as possible. As such, the

performance of the statistical indices and classification methods will be evaluated based on their performance in such a context.

As we are considering a two-class classification problem, each proposed classifier was trained using both normal (control) and abnormal (hypotensive) data from a training dataset. The classifiers' free parameters are then optimised to distinguish between the two classes by the selection of an appropriate threshold on the classifier's output.

All indices were evaluated based on their ability to distinguish between individuals in the H and C groups. Several measures of classifier performance are calculated based on the ability of the classifier's ability to distinguish between the two given classes. In this regard, there are four types of possible classifications: true positives, false positives, true negatives, and false negatives. In our problem definition, hypotension is the event to be predicted, thus the four classifications can be summarised as follows:

- **True Positives** occur when hypotension is predicted for a hypotensive patient (H patient classified correctly)
- **False Positives** occur when hypotension is predicted for a control patient (C patient classified incorrectly)
- **True Negatives** occur when hypotension is not predicted for a control patient (C patient classified correctly)
- **False Negatives** occur when hypotension is not predicted for a hypotensive patient (H patient classified incorrectly)

The total number of correct classifications (CC), can then be defined as $CC = True\ Positives + True\ Negatives$. Using these definitions, further performance metrics can be specified for a classifier.

4.1.1.1 Sensitivity and Specificity

Sensitivity and specificity measure the ability of a classifier to identify a specific subgroup of patient. Sensitivity is a statistical measure of a classifier's ability to correctly identify

positives (AHE patients), defined as the proportion of correctly classified positives to all positives:

$$Sensitivity = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{4.1}$$

Specificity measures a classifier's ability to correctly identify negatives (non-AHE patients), defined as the proportion of correctly classified negatives in a population:

$$Specificity = \frac{TrueNegatives}{TrueNegatives + FalsePositives} \tag{4.2}$$

4.1.1.2 Positive Predictive Value and Negative Predictive Value

Positive predictive value (PPV) and negative predictive value (NPV) measure the proportion of patients with a positive or negative prediction who are correctly identified. Thus PPV is the probability that a patient will experience an AHE given that one was predicted, and NPV is the probability that a patient will not experience an AHE given that control status was predicted.

$$PPV = \frac{TruePositives}{TruePositives + FalsePositives} \tag{4.3}$$

$$NPV = \frac{TrueNegatives}{TrueNegatives + FalseNegatives} \tag{4.4}$$

Note that both measurements are dependent on the prevalence of positives in the dataset.

4.1.1.3 Receiver Operating Characteristic Curve and the Area Under Curve

A standard method of assessing classifier performance is the generation of Receiver Operating Characteristic (ROC) curves and the calculation of the area under the curve (AUC). Non-binary classifiers output a numerical value, referred to here as an index, for a given set of inputs. Output indices are obtained from training data; these are then sorted, and

each is considered as a possible threshold value for classification. The optimal threshold would then be used to distinguish between positive and control patients in the testing data. ROC curves are used to determine what threshold value is most successful in distinguishing between the index values of the two classes in the training data. ROC curves are plots of sensitivity versus 1-specificity as the threshold is varied as follows:

For each unique index value obtained from training data:

1. Set the index value to be the classification threshold. Patients with an index value lower than the threshold are classified as positives[1], and all others as controls.

2. A count is made of the total number of:

   (a) AHE patients with an index lower than this value (true positives)

   (b) AHE patients with an index greater than or equal to this value (false negatives)

   (c) Non-AHE patients with an index lower than this value (false positives)

   (d) Non-AHE patients with an index greater than or equal to this value (true negatives)

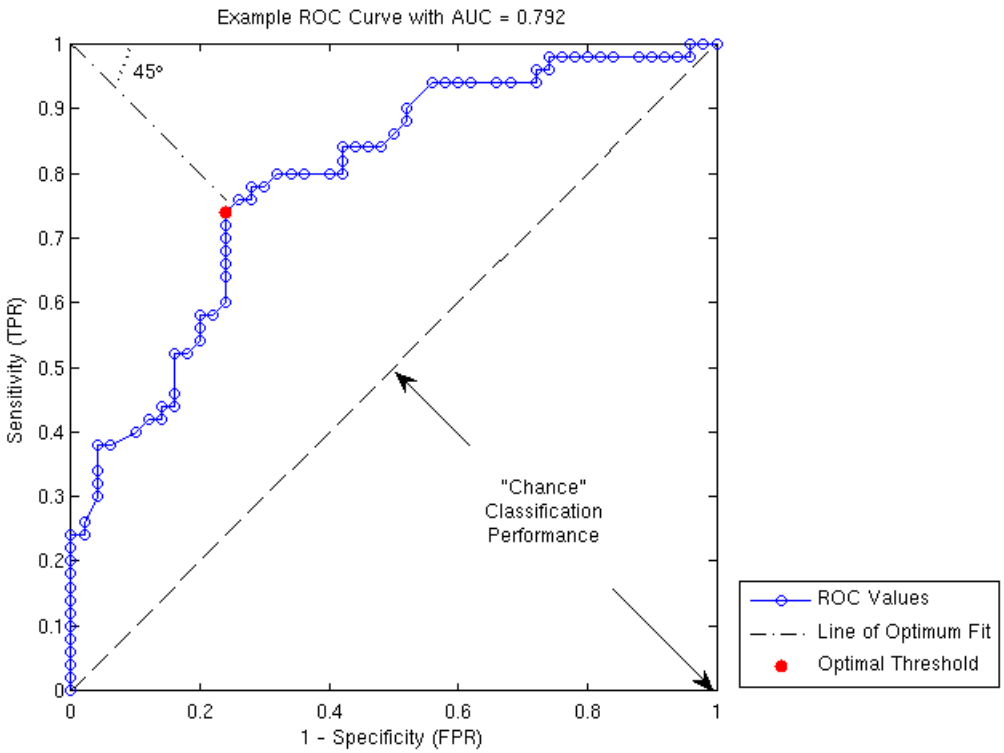
3. Calculate the sensitivity for this threshold

4. Calculate the specificity for this threshold

The optimal classifier, is found when the threshold gives the best compromise between failing to detect true positives and generating false positives. Note that we have assumed that the cost of misclassifying positives and negatives is equal, thus the best classifier is given by the threshold which generates the ROC space point lying on a 45° line closest to perfect classification (the point [0 , 1] in ROC space). (See Figure 4.1)

[1]This assumes that we are considering an index of normality, not abnormality. In the case of an index of abnormality, positive classifications are assigned to index values higher than the threshold.

Figure 4.1: *ROC Example Curve*



4.1.2 Data Resampling

In order to use ROC curve analysis in the determination of the optimal classifier, the curve itself should be computed using training data which is independent of the test data. Otherwise there is the danger that the classifier is simply learning the particulars of that specific data rather than gathering a larger understanding about the condition in general. However, given the limited size of our dataset, this means that an ROC curve with a relatively low number of points will be generated (84 points if half of data is assigned to train and half is assigned to test). Therefore, in order to produce the best ROC curve possible (i.e. with the largest number of points), a data resampling technique was used.

Data resampling can refer to a number of methods for expanding the available number of testing points for a classifier. The three most popular methods are cross-validation, bootstrapping, and jack-knifing. [45] Cross-validation is a way of partitioning available data into subsamples such that parameters estimated in one subsample are then applied

to another subsample. In the assessment of a classifier, this means that the classifier is evaluated by using random subsets of the testing and training data. Bootstrapping is similar, except that any single measurement could be used multiple times in the test population.

Given the equal number of positive and control patients in our dataset of size N, a balanced distribution was used to evaluate classifiers. Performance metrics were generated by randomly selecting a training set of size $m = \frac{3*N}{4} = 126$ and a testing set of size $n = \frac{N}{4}$ from the available data, ensuring that there is no intersection between the two sets. This was performed 10 times, and the method proposed by Xie et. al. [46] was then applied to create a smooth ROC by using all the output values (m*10) as part of the classifier output. Having obtained the optimal threshold on the training data, we then report classification performance using that threshold on the test data patterns. Thus the sensitivity and specificity will be calculated for each threshold value in the entire output range rather than in each of the subsets independently. For each gap time (60 - 180 minutes prior to T0), the smooth ROC curves and other metrics were calculated 100 times in order to generate a reliable estimate of the variance of the performance metrics. This method is summarised in the following algorithm:

for each g = 60 : 180
  for each i = 1 : 100
    for each j = 1 : 10
      1. randomly select **m** training and **n** testing points
      2. create/train classifier using training subset
      3. test classifier on testing subset to get **n** output values
    end
    4. create a single ROC curve using **j x m** training values
    5. select optimal threshold from ROC curve
    6. calculate performance metrics on **j x n** test values
  end
  7. tabulate the median performance metrics from the **i** ROC curves
end
8. plot the performance metrics for each gap time, **g**

4.2 Statistical Indices

A sub-set of the statistical indices which were shown to have performed well on the 2009 Computers in Cardiology datasets were investigated on our problem. Indices which evaluated to a non-numeric value were not classified. In this case, Index I through VI as described below in Table 4.1 were implemented. Note that as a minimum of 3 hours of data (as opposed to 10 hours of data) was available for use in the calculation of Index III, a time constant of the same proportion was used in the Exponentially Weighted Moving Average formula. As a time constant of 1.2 hours had been selected for 10 hours of data, a time constant of 22 minutes was used instead.

Table 4.1: *Statistical Index Algorithms*

Label	Algorithm
Index I	Arithmetic mean of the five most recent minutes of numeric ABPMean data
Index II	Arithmetic mean of the five most recent minutes of waveform ABP data
Index III	Exponentially weighted moving average of the 3 hours of numeric ABPMean data
Index IV	Predicted value of numeric ABPMean data at T0 via simple linear regression of the most recent hour of numeric ABPMean data
Index V	Arithmetic mean of the five most recent minutes of numeric ABPDias data
Index VI	Voting strategy between Index II and Index V; an AHE is predicted only if predicted by both Index II and Index V

Performance evaluation as specified by the AUC, percentage of correct classifications, sensitivity, specificity, PPV and NPV is shown below for the entirety of the time range using the balanced test set (Figure 4.2). Numbers shown represent the median performance seen across all 100 iterations of the algorithm, with whiskers representing the 10th and 90th percentile values.

As shown in Figure 4.2, while performance varied across the metrics, there was a general trend towards performance reduction as the gap size was increased. In all cases except for Index III, AUC scores tended to follow approximately the pattern as the length of the gap is increased.

When evaluating the balanced test set, most indices had a roughly balanced sensitivity

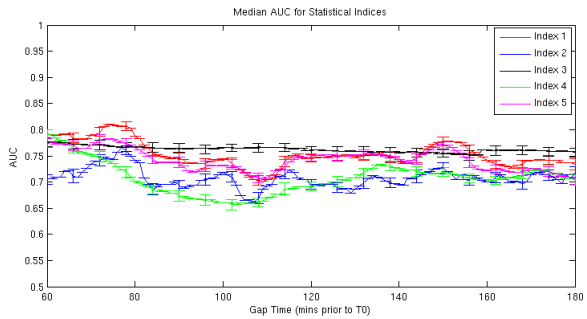
and specificity over time, but sensitivity tended to decrease as gap size increased while specificity either increased or remained roughly the same. Similarly, the PPVs and NPVs of indices were clustered near 0.75 and 0.7, with a gradual decline in performance over time.

The percentage of correct classifications tended to be similar in the balanced test data set, with top performance from Index I using the smallest gap size (75% correct) and Index III using the longest gap size (70% correct). Index III also had less variance in the number of correct classifications as gap size was varied, and did not demonstrate the decrease in performance that other indices did in the 80 to 120 minutes gap times.

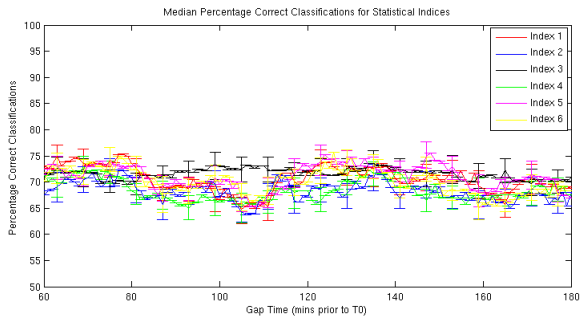
The purpose of re-implementing statistical indices was to identify the best performing metrics for AHE prediction based on the fact that simple statistical indices had been shown to be good predictors of AHE onset within 30 minutes. ABP values are significantly lower in patients just prior to AHE onset, as demonstrated by the fact that a simple five-minute average of one or more ABP signals is always amongst the best predictors of AHE onset within the next 30 minutes, but this does not hold for longer-term prediction. The best metrics (Index I and Index III) may be useful as metrics for more advanced prediction rules, but are incapable of predicting an AHE more than an hour ahead.

Figure 4.2: *Statistical Index Performance on Balanced Dataset*

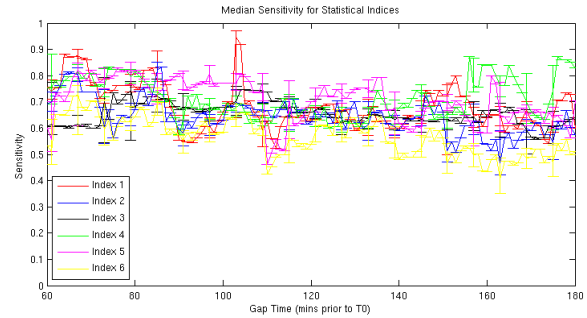
(a) *Area Under Curve*



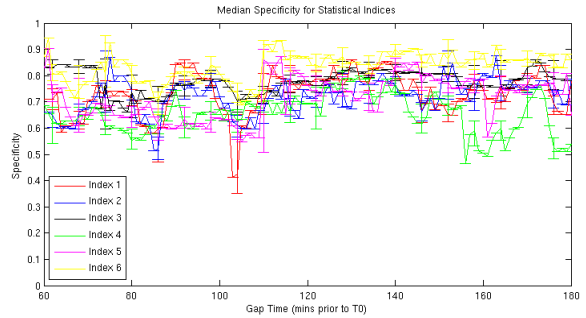
(b) *Percent Correctly Classified*



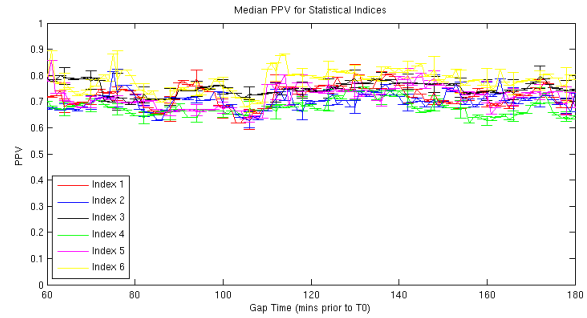
(c) *Sensitivity*



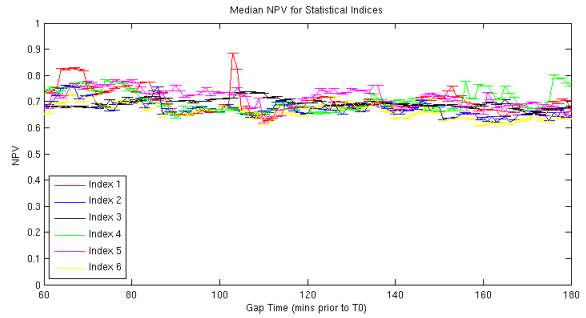
(d) *Specificity*



(e) *PPV*



(f) *NPV*



4.3 Models of Normality

One possible reason for the good performance of the GRNN-based model in the 2009 Computers in Cardiology Competition despite over fitting is its use of template matching at the core of the method. GRNNs create input nodes for each of the training vectors and calculate a distance metric between these nodes and new test data in order to predict outcomes. This ensures that only the templates closest to the input vector are used in the prediction of a new output. Henriques et. al. take this approach further by correlating the input vector with each training vector, before using a weighted combination of the mostly highly correlated GRNNs to calculate the output. Simplifying this approach, we examine the use of Parzen windows [42] initially to construct models of normality (control patients).

With Parzen windows, the data observed is assumed to be drawn from an underlying probability density function (PDF). The Parzen window method uses the normalized linear combination of kernel functions placed at each point represented by the training data to estimate the PDF. This allows for an estimate of the probability that a new data pattern $\mathbf{x}$ is a part of the training data set of P-dimensional patterns. As explained previously in Chapter 2, an unknown function $f(x)$ which represents the known joint continuous PDF of a vector variable, $\mathbf{x}$ can be estimated from sample observations. The Gaussian-based probability estimator $\hat{f}(\mathbf{X})$ can then be specified for given sample values of the random variable $\mathbf{x}$ where M is the number of sample observations and P is the dimension of the vector variable $\mathbf{x}$. This can be understood as a sum of spherical Gaussians centred at every training data point, normalized by the number of total training samples, M.

We consider a region of volume V_m, containing k_m of the M total training points. As shown in Equation 4.5, the probability density p_m in the region containing $\mathbf{x}$ can be estimated as the fraction of samples that fall into the region normalized by the region's total volume.

$$p_m(\mathbf{x}) = \frac{k_m/M}{V_m} \tag{4.5}$$

Depending upon what function is chosen, the total number of samples which fall in the window, k_m can be expressed as:

$$k_m = \sum_{i=1}^{M} \varphi(\mathbf{x}_i) \tag{4.6}$$

By substituting Equation 4.6 into Equation 4.5, we can create an equation for $p_m(\mathbf{x})$ which gives an estimate of the probability density at the location of the pattern $\mathbf{x}$ from the training patterns $\mathbf{x}_1, \ldots, \mathbf{x}_M$:

$$p_m(\mathbf{x}) = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{V_m} \varphi(\mathbf{x}_i) \tag{4.7}$$

To evaluate k_m itself, we must choose a window function φ. In order to ensure that the estimated probability density function p_m satisfies the properties of a PDF (non-negative and integrate to one), the window function itself must also be non-negative and $\int \varphi(\mathbf{x}) dx = 1$. φ is therefore usually chosen to be a PDF itself, and often a Gaussian PDF as it is infinitely differentiable. The equation for a Gaussian distribution of zero-mean and unit-variance, which is centred at the origin can be expressed in Equation 4.8.

$$\varphi(\mathbf{z}) = \frac{1}{(2\pi)^{P/2}} \, e^{-\frac{\|\mathbf{z}\|}{2}} \tag{4.8}$$

Substituting the specified φ function, and the volume of a spherical Gaussian surface of radius σ into the probability density estimate in Equation 4.7 yields:

$$p_m(\mathbf{x}) = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{\sigma^P} \, \frac{1}{(2\pi)^{P/2}} \, e^{-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2\sigma^2}} \tag{4.9}$$

After re-arrangement, the Parzen Windows estimator can be expressed as a simple sum of spherical Gaussian distributions centred on each data point, and normalized by M:

$$p_m(\mathbf{x}) = \frac{1}{M\ (2\pi)^{P/2}\sigma^P} \sum_{i=1}^{M} e^{-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2\sigma^2}} \tag{4.10}$$

where $\| \mathbf{x} - \mathbf{x}_i \|^2$ is the squared distance between training points and the point of prediction, and σ is the spread of Gaussian function or the Parzen Window width.

4.3.1 Unconditional PDF Estimation via Data Fusion

In an effort to create models which recognize that different thresholds of ABPMean may represent normality for different patients, a real-time automated data fusion model was explored. Tarassenko et. al. [47–49] describe a data fusion model of normality in which vital signs (heart rate, breathing rate, oxygen levels, temperature and blood pressure) are the input vector x in a Parzen windows estimator, the output of which is a patient status index of novelty (or abnormality). The inspiration for the model was the evidence that many in-hospital cardiac arrest patients have previously had abnormal vital signs six to eight hours before the arrest. [50–53]

The data fusion model uses vital sign data acquired from acutely-ill hospital patients to generate a five-dimensional probabilistic model of normality as an approximation to the data's unconditional PDF. This previously generated model is then used to evaluate the probability that incoming vital signs are normal. Data was collected from 150 general-ward patients at Oxford's John Radcliffe Hospital between 2001 and 2003, creating a training set of approximately 3,500 hours of data. Specific groups from which patients were drawn include, but were not limited to, post myocardial infarction, severe heart failure, acute respiratory problems, and elderly hip fracture patients.

The Patient Status Index (PSI) increases with abnormality, and an alert is generated when the PSI crosses a clinically-validated threshold. In this way, the PSI alerts can replace single-channel alert systems by notifying nursing staff if either a single vital sign deviates greatly from its normal value or two or more vital signs deviate from their normal values by smaller amounts.

The model of normality is the training data's unconditional PDF, $p_m(\mathbf{x})$, where $\mathbf{x} = x_1, x_2, \ldots, x_5$ is a data vector composed of vital sign parameters x_1 = heart rate (HR), x_2 = breathing rate (BR), x_3 = systolic/diastolic mean blood pressure (SDA), x_4 = peripheral oxygen saturation (SpO_2), and x_5 = temperature. Each of the five parameters is normalised (zero-mean, unit-variance transformation) to give equal weighting to each parameter. With 400 5-D prototype data patterns, Equation 4.10 becomes:

$$p_m(\mathbf{x}) = \frac{1}{400\ (2\pi)^{5/2}\sigma^5} \sum_{i=1}^{400} e^{-\frac{(\|x-x_i\|)^2}{2\sigma^2}} \tag{4.11}$$

The Patient Status Index is a measure of novelty, i.e. departure from normality:

$$novelty = \log_e \frac{1}{p(\mathbf{x})} = -\log_e p(\mathbf{x}) \tag{4.12}$$

New incoming data can then be compared to the model's probability distribution, whereby areas of low probability will correspond to abnormal vital signs and areas of high probability will correspond to normal vital signs.

For our problem, each of the 84 control patient's data was considered for four of the five variables. Temperature was excluded as most of the MIMIC II data contains only intermittent temperature readings recorded by the nursing staff rather than any continuously measured data.

All data which would be used in model construction for any gap size (60 - 240 minutes prior to T0) was collected from the 84 control patients in the population. The resulting means and standard deviations were then used to normalise all patient data. The results of the normalisation are given in Table 4.2 below, and compared against the means and standard deviations obtained by Hann et. al [49] in the calculation of the PSI. Note that our MIMIC II data makes use of mean ABP while the previous PSI data used the arithmetic mean of the systolic and diastolic blood pressures (SDA).

A model of normality was then constructed for the shortest gap time using the hour of data 60 minutes prior to T0 for 63 ($\frac{3}{4}$) of the control patients. 21 patients were held

Table 4.2: *Patient Data Normalisation Boundaries*

	MIMIC II		PSI Data	
	Mean	STD	Mean	STD
ABPMean / SDA	84.72	13.76	94.68	12
HR	85.18	15.06	83.77	17.48
RESP (BR)	17.61	5.84	18.3	5.06
SpO_2	97.57	3.52	95.2	3.49

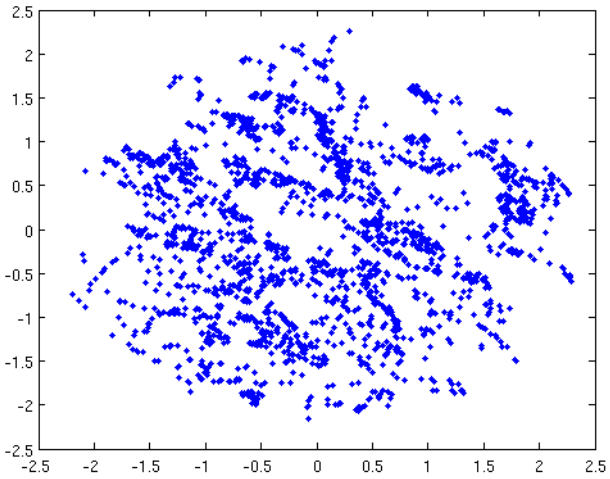
out to use in the test population. This amounts to 3780 samples for the 63 control patients. Before building a model of normality, the 10% most abnormal patterns (those furthest away from the (0,0,0,0) mean) were removed from consideration. The removal of this data left 3402 total samples with which to construct the model of normality. After normalisation, data which was more than 5 standard deviations from the mean was removed as artefactual.

In order to determine if the 3402 samples of patient data in the chosen four variables diverged significantly enough for PDF classification, the Neuroscale algorithm was then used in order to produce a mapping from 4-dimensional space to a visualisable 2-dimensional space. This algorithm uses a Radial Basis Function (RBF) neural network to create a non-linear mapping between the network input (4-dimensional data) and output (2-dimensional data). As seen in Figure 4.3 below, control patients appear to have a closely clustered distribution in the 2-D visualisation space despite being drawn from a more diverse population than the tightly defined hypotensive patients.

In order to test performance, the most recent hour of data was gathered for the test patient population consisting of the remaining 21 control patients and 21 randomly selected hypotensive patients, and the probability density $p(x)$ was evaluated at each point. This was done four times, using an independent set of control and hypotensive patients in the test population each time. Plotted below are the histogram values for novelty as evaluated from the testing data. Histograms were normalised so that the area under the curve for both C and H patients was the same.

Three different thresholds for differentiation between the two classes were then eval-

Figure 4.3: *Neuroscale Mapping of Control Points*



uated using that novelty value for a given centile of the control patient data. This corresponds to selecting the classification threshold to be a pattern in the constructed 3402 sample model of normality which is more abnormal than most of the others. For example, setting a threshold at the 99th centile is equivalent to determining that any value further from the model's centre than the least normal pattern amongst the control patients is abnormal.

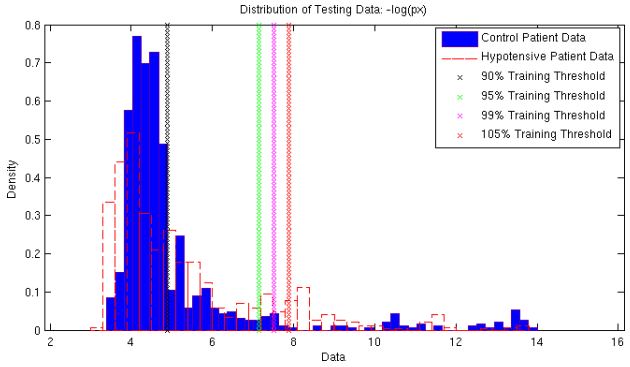
1. 90th centile of normality (novelty value of 90th least normal pattern)
2. 95th centile of normality
3. 99th centile of normality
4. 105th centile of normality (threshold is 5% beyond the most abnormal pattern value seen in the control data)

Using these values as thresholds, patterns in the test data were classified. Any control patterns at or below this threshold were determined to be true negatives, while hypotensive patterns above this threshold were true positives. Shown in Figure 4.4 are two different histograms for different selections of test and train patients.

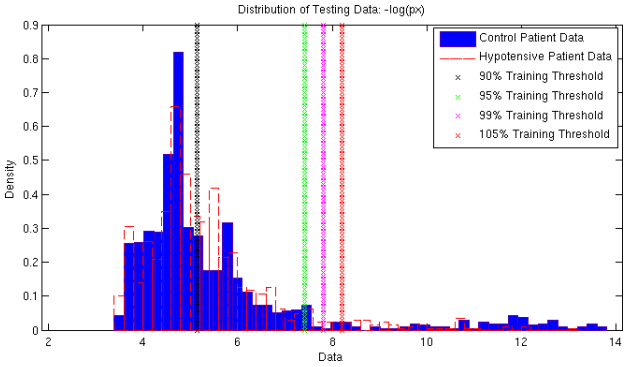
From previously examining the distributions, our assumption was that there would be far more false negatives than false positives generated with the more stringent thresholds,

Figure 4.4: *Novelty Score Histograms*

(a) *First Novelty Distribution*



(b) *Second Novelty Distribution*



and this would became more balanced as the threshold was expanded. This is due to a large number of the hypotensive patterns falling into the "normal" range. Actual performance results reflect this, with mean classification rates, sensitivity, specificity, PPV and NPV listed in Table 4.3.

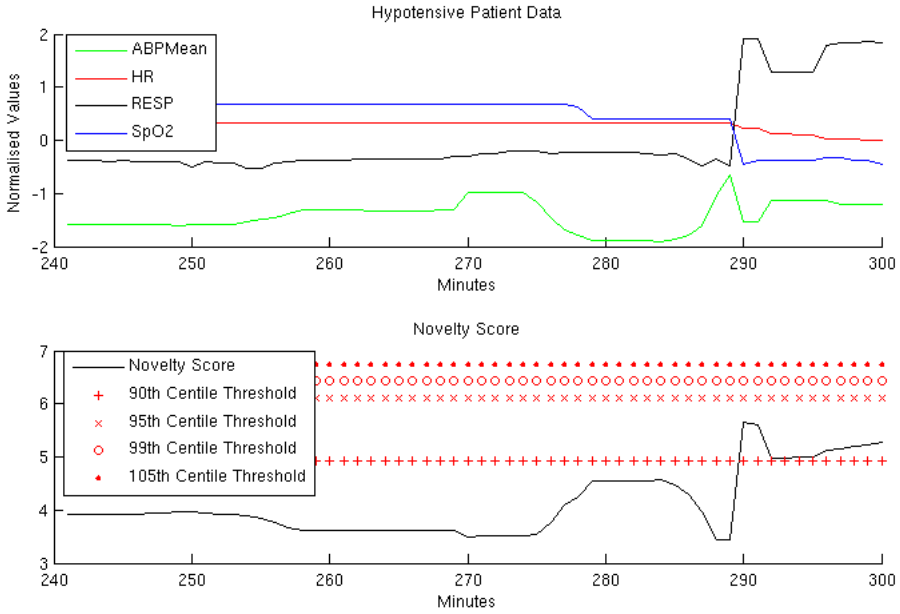
Table 4.3: *Performance with Varying Thresholds*

Threshold	90th Centile	95th Centile	99th Centile	105th Centile
Correct Classifications	51.46%	52.16%	52.26%	52.27%
Sensitivity	0.46	0.16	0.13	0.11
Specificity	0.57	0.88	0.90	0.92
PPV	0.51	0.50	0.50	0.50
NPV	0.53	0.53	0.52	0.53

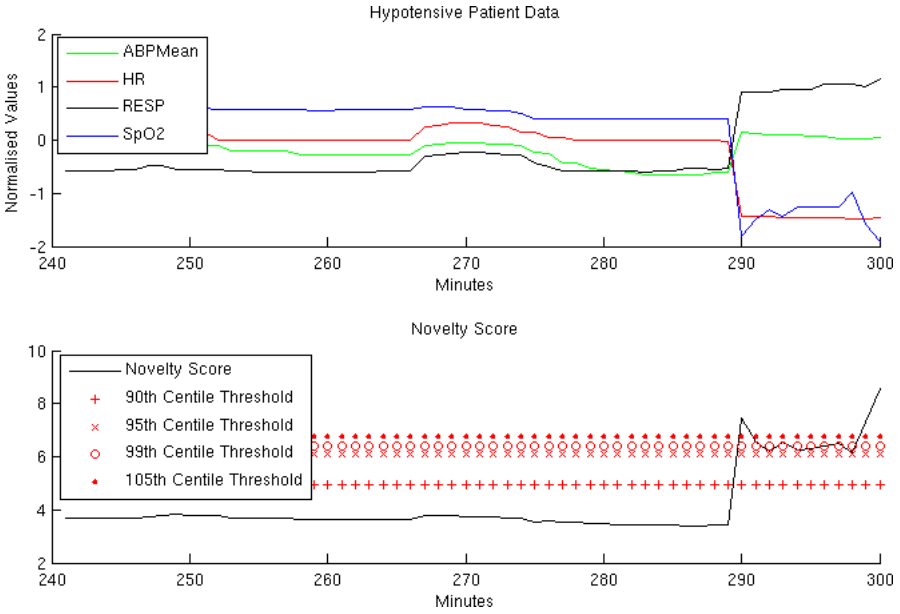
Interpreting these results, it is not surprising that amongst an hour of patient data (in this case the hour of data between 2 hours and 1 hour prior to T0) there are many normal patterns in the hypotensive cases. This explains why our classification scores are only slightly higher than random. An example of the novelty score seen for two individual patients over the entire hour of data is shown in Figure 4.5 for comparison.

Figure 4.5: *Patient Novelty Scores*

(a) *First Sample Patient*



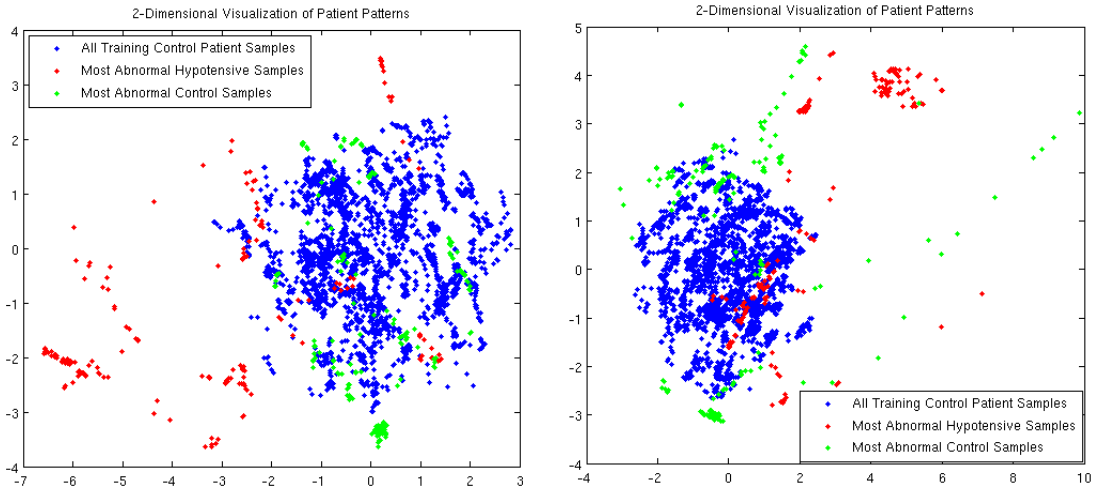
(b) *Second Sample Patient*



In order to determine if there was a consistent underlying cause for the most abnormal hypotensive patterns separable from "normal" control data and "abnormal" control data, the 20% of patterns lying furthest from the model of normality were plotted on the Neuroscale map of normality for both control and hypotensive test patterns. Examples of two plots are shown in Figure 4.6.

Figure 4.6: *Neuroscale Patient Data Visualisation*

(a) *First Data Distribution Example* **(b)** *Second Data Distribution Example*



The corresponding normalised value ranges (median ± std) of the four vital signs for these most abnormal hypotensive and control patterns are listed as the average across four independent executions of the algorithm, each consisting of a different random training/testing partition of data.

Table 4.4: *Abnormal Data Ranges*

Variable	Hypotensive	Control
ABPMean	-0.68 ± 1.02	0.47 ± 1.14
HR	0.41 ± 1.60	0.28 ± 1.60
RESP	0.66 ± 1.91	0.34 ± 1.70
SpO_2	-0.98 ± 1.36	-0.25 ± 0.95

Examining these values, those hypotensive patients with the most abnormal scores seem to have a consistently low ABPMean, elevated HR, elevated RESP and lowered SpO_2 level. These attributes are distinct from those seen in the most abnormal non-AHE

patients as well, which are closer to the model mean in all variables.

4.4 Discussion

In this chapter, statistical indices were evaluated against a balanced test set using prediction times between one and three hours prior to AHE onset. These results are to be used as a benchmark against other methods evaluated in this thesis. Amongst the statistical indices, correct classifications centred on 70% for top performers. Given these results, Statistical Indices I and III were chosen as possible parameters for inclusion in a future classifier. The use of the Patient Status Index (PSI) as a simple classifier produced very poor results, often resulting in scores which were little better than chance. This is largely due to the fact that vital signs in our control population an hour or more before an acute hypotensive event are not necessarily very different from those of the patients who eventually experience the event.

Chapter 5

Long-term AHE Prediction Using Classification Methods

5.1 Introduction

The previously evaluated indices and model have centred upon the calculation of a single value which is then thresholded in order to generate the best possible classification performance on a balanced test set. In contrast to these methods, we explore in this chapter classification methods which involve the generation of a model. The model describes the relationship between input parameters generated from a patient's vital signs, and a binary output parameter: does acute hypotension (as previously described) occur or not?

We believe that classification methods may be reasonable for AHE prediction given that there are specific physiological conditions which trigger its onset. As reviewed in 1.5, many of the conditions which cause hypotension tend to have a gradual effect on the body, resulting in possible features for a classification model to use as determinants. Additionally, prior work using classifiers for similar problems have met with some success. [3]

In this chapter, Linear Classifiers, Multi-layer Perceptron Neural Network Classifiers and Models of Normality (Bayesian Classifiers) are explored as possible methods of clas-

sifying the patient data. Each method is reviewed independently, and results are then presented. Finally, all of the methods are compared in order to determine which type of classifier may be optimal.

5.1.1 Dataset Distribution

As noted by Lee et al. [3], the occurrence of hypotensive episodes in ICU data is far rarer than the statistics of our training data would suggest. This is partially due to the prompt treatment of hypotension with pressors and fluid in ICUs. Lee's findings indicated that acute hypotensive episodes only represented 3% of the data were hypotensive. Similarly, our analysis of MIMIC II [28] suggests that clinically significant hypotension actually occurs in 6% of the ICU population. (Table 3.1)

To evaluate classifier performance in a real-life setting, it is important that its performance is evaluated on a test set which reflects the prior probabilities of positives (hypotension) and controls (normal). As our target condition is uncommon, it is important that a prediction method generates few false positives, as the difference between positives and controls could lead to more false than true positives being predicted.

Classifier performance was therefore evaluated on both a balanced dataset, and one which reflected the prior probability of hypotensive data in an ICU situation. In the secondary method of performance evaluation, the training set was composed using balanced training data (equal ratios from each class), whereas the testing set included 6% of patients with hypotension. This leaves fewer examples available for testing purposes, and models were evaluated in the following way:

As with the balanced data set, 10 iterations were completed in which a group of $m = \frac{3*N}{4} = 126$ patients was randomly selected for the training set and the remaining $n = \frac{N}{4}$ patients were used as the testing set. All the patient data in the training set was used, creating a balanced training set consisting of 63 randomly selected control examples and 63 randomly selected hypotensive examples. To produce the required distribution for the testing set, the remaining 21 control samples and one of the 21 hypotensive samples

were used in each of the 10 iterations. This created a test set in which 4.5% of examples (10/220) were hypotensive, which lies between the 3% and 6% incidence rates mentioned previously.

5.2 Classification Using Linear and MLP Classifiers

5.2.1 Linear Classifier - The Single Layer Perceptron

A linear classifier is the simplest form of a classifier, and performs classification based on a linear combination of input values. [54] The single layer perceptron (SLP) is the simplest form of a linear classifier: it consists of a single binary discriminant with adjustable weights w_N and a bias weight b. (See Figure 5.1) A linear classifier splits a multi-dimensional input space with a hyperplane. The perceptron algorithm was proposed in the 1950's by Rosenblatt, [55] and was shown to converge for classification problems which are linearly separable.

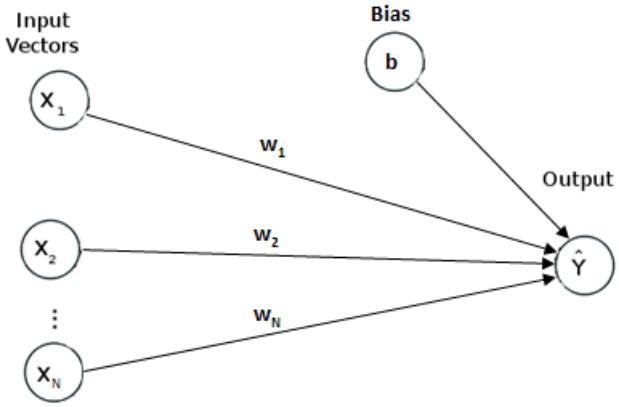


Figure 5.1: *SLP Network Diagram: Inputs $x_{1...N}$ represent patient data with corresponding weights $w_{1..N}$ applied to each. The bias value is denoted as b, and the final output after applying the activation function to Y_{IN} is $\hat{Y}$.*

The perceptron algorithm begins with a set of random values for w and b, and a learning rate α which is set to a value between 0 and 1. The algorithm then iterates to reduce the error between the output value and target value (desired classification label). During training, the input values $x_1 \ldots x_N$ are presented and the perceptron output Y_{IN}

is calculated according to the Equation:

$$Y_{IN} = \sum_i x_i * w_i + b \tag{5.1}$$

An output value $\hat{Y}$ is generated by applying an activation function to the Y_{IN} value. The activation function for a two class problem is usually a binary threshold function. The output value, $\hat{Y}$ is then compared to the desired target value, t. The weights and bias are then updated using the following equations:

$$w_i = w_i + \alpha * (t - \hat{Y}) * x_i \tag{5.2}$$

$$b = b + \alpha * (t - \hat{Y}) \tag{5.3}$$

The training process continues until a stopping criterion is met. The trained SLP can then be used on the test set to classify each pattern in this set.

5.2.2 Multi-Layer Perceptron Classification

As the classification of the vital sign data sets "normal" or "hypotensive" may be a non-linear problem, a Multi-layer Perceptron (MLP) classifier was also explored. The Multi-Layer Perceptron (MLP) uses one or more hidden layers to achieve more complex associations and non-linear mappings. [55]

An MLP has two or more layers of nodes connected such that the input is fed forward through each subsequent layer to the output layer. As with the SLP, each node in a given layer is connected to every node in the subsequent layer by a weight $w_{i,j}$. The MLP nodes use a non-linear activation function to create better approximations of non-linear relationships in data. Usually the non-linear activation function is a sigmoid, which is described by:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{5.4}$$

The simplest supervised learning technique for the MLP is called error backpropagation. [56] The goal is to modify the weights w and biases b so that the output y moves closer to the target output t. See Appendix 4 for more details.

5.2.3 Implementation

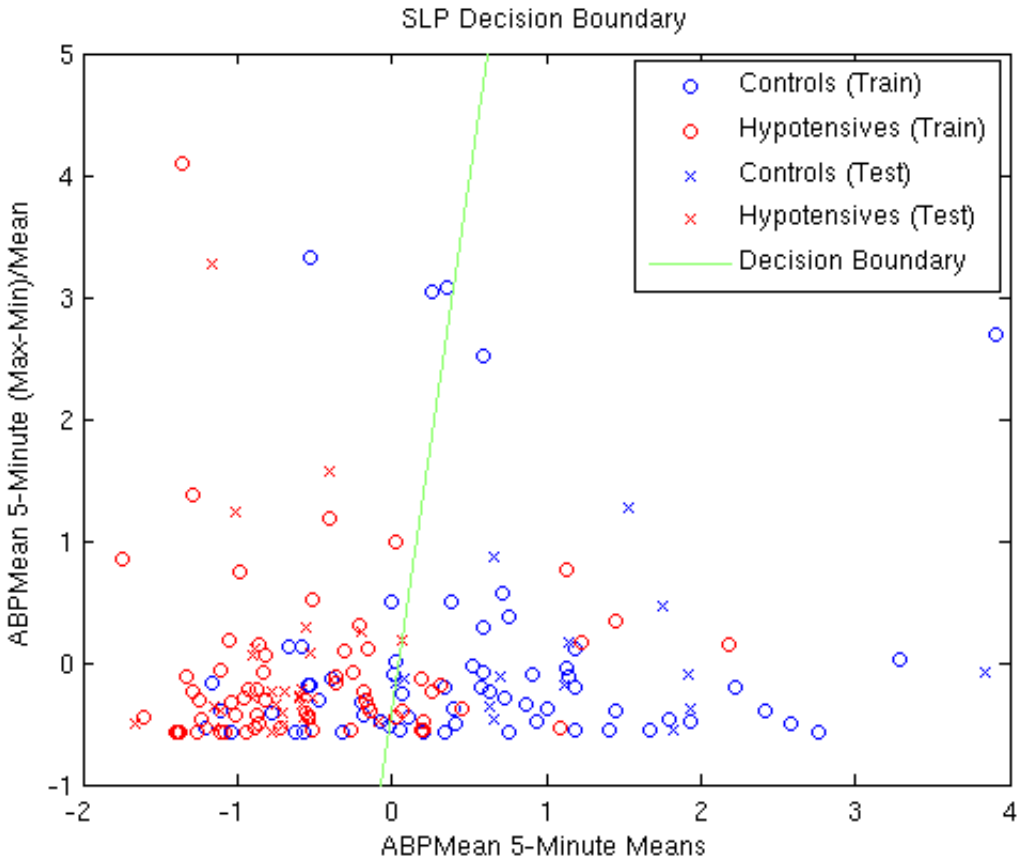
Features which were extracted for both the SLP and MLP models from the ABPMean signal were the individual values themselves, the percent change from consecutive readings, the 5-minute mean, the 5-minute variation ((max-min)/mean), the 5-minute median, and the 3-hour exponentially weighted mean. This led to a total of 6 variables for the networks to choose from at each given gap time. Models were run for every fifth gap time, leading to a total of 25 points of evaluation. Each feature was normalised to a zero-mean, unit-variance distribution to help avoid poor local minima.

As an initial demonstration of classification capacity, a linear SLP was constructed using two of the features (the 5-minute variation and 5 minute means), and the chosen classification boundary is shown in Figure 5.2.

An SLP with a linear output function (linear regression) and an SLP with a logistic output function (logistic regression) were both tested on a balanced and unbalanced testing dataset (two SLP models for evaluation). Models were trained using the perceptron update algorithm until there was either less than 5% of training data classified incorrectly or 100 training passes had been completed.

The MLP also consisted of six inputs, and the number of hidden nodes used was varied from 10 to 30 in increments of two nodes at each step, each of which used log-sigmoid activation functions. This created at total of 11 models for evaluation. This was done to evaluate the impact of additional nodes on classification performance. Note that the classifier by Lee et. al. [3] consisted of an MLP constructed with a single hidden layer of

Figure 5.2: *2-Dimensional SLP Classification*



20 nodes and log-sigmoid activation functions. Models were trained using the gradient descent algorithm until there was either less than 1% of training data classified incorrectly or 1000 training passes had been completed.

Input node weights were drawn at random from a zero mean, isotropic Gaussian with variance of 0.01. Distinct SLPs and MLPs were trained for each gap size and iteration and were tested on both balanced and unbalanced testing datasets.

5.2.4 Results

As seen below in Figure 5.3, Figure 5.4, Figure 5.5 and Figure 5.6, the unbalanced test case demonstrates that the model still is not able to give a high value of positive predicting on the given low incidence rate. Results for both the linear and logistic SLP models were very similar, producing AUCs near 0.6, with an average performance of 62%

correct classifications in the balanced case. Performance on the unbalanced dataset was similar, averaging near 65%, with the best single classification performance near 80%. In general, sensitivity and NPV were higher than specificity and PPV, meaning that the models generated fewer false positives than false negatives. This tends to be true of most of the models we have evaluated thus far, simply because even hypotensive patients will have some non-hypotensive data over the course of three hours. PPV performance in the balanced dataset was never above 0.1, which indicates that this model would not be as successful at identifying sparse hypotensive events in a real-time ICU situation. MLP results were better than the SLP outcomes, with AUCs ranging from 0.81 to 0.68. Higher AUC scores yielded similarly improved classification performances of 73% and 68% average correct classifications in the balanced and unbalanced cases.

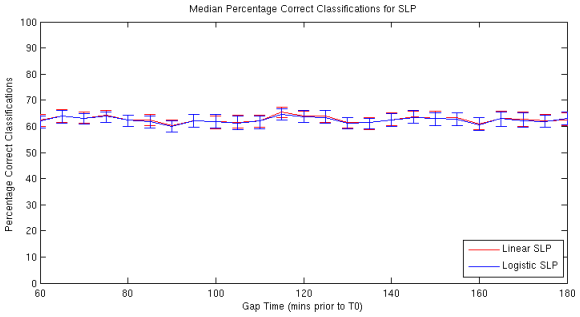
To create a benchmark for performance of classification algorithms in the unbalanced case, statistical Indices 1, 3, 4, and 5 were recalculated using the unbalanced dataset. (Figure 5.7) Classification performance in this case tended to be near 60% for most of the indices, with Index 3 yielding the most consistent performance. Sensitivity and specificity were roughly balanced, and tended to have a more gradual decline as time gaps extended. As might be expected from an unbalanced distribution, the NPV was high across the gap times, but PPV was low. In this case, PPV did not exceed 0.175 for any of the median values across all indices.

Given that the SLP performance indicators (Figure 5.3) had very little variance as the gap was varied from 60 to 180 minutes, the performance of the SLP models outside of this timespan in the balanced test set was also explored. As shown in Figure 5.8, after a top classification performance of 93% in the minute prior to AHE onset, the SLP's performance reduces rapidly, dropping to nearly 65% by the 35 minute gap time. AUC descends more gradually, reducing from 0.96 at the lowest gap time to 0.66 at the 55 minute gap. As the gap is increased beyond 180 minutes AUC and classification performance fall off gradually, ending near the 50% mark at 240 minutes prior to onset. By the 240 minute gap time, the features derived from blood pressure alone are not

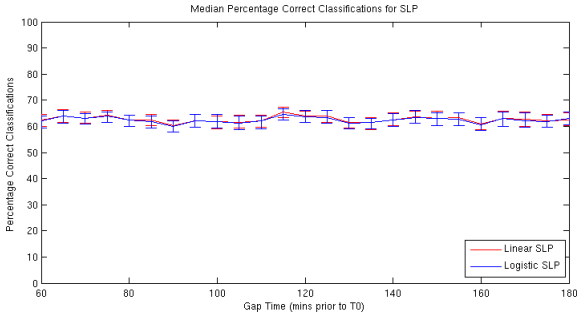
impacted significantly enough by the future hypotensive event to separate positives from controls. However, by using features generated from more variables than just the ABP, we may be able to increase performance to a more acceptable rate across the longer gap times.

Figure 5.3: *SLP Performance on Balanced Dataset*

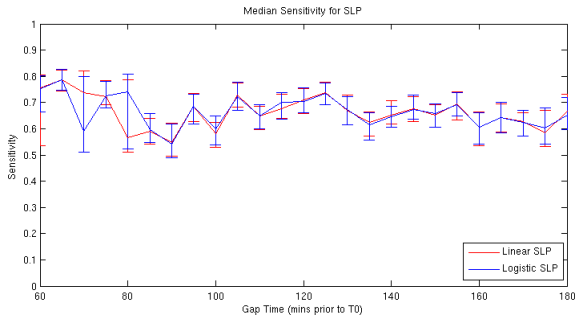
(a) *AUC*



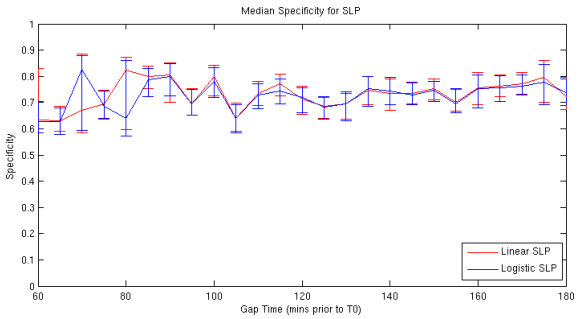
(b) *Percent Correctly Classified*



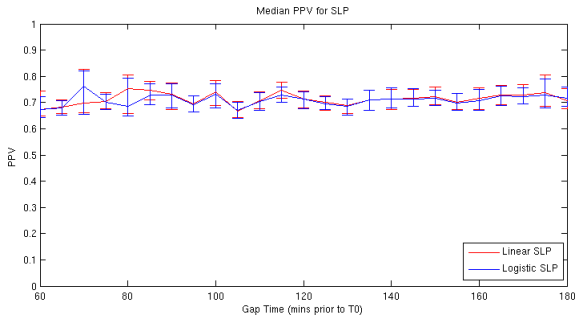
(c) *Sensitivity*



(d) *Specificity*



(e) *PPV*



(f) *NPV*

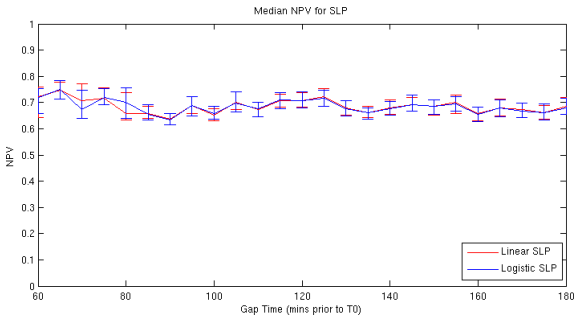
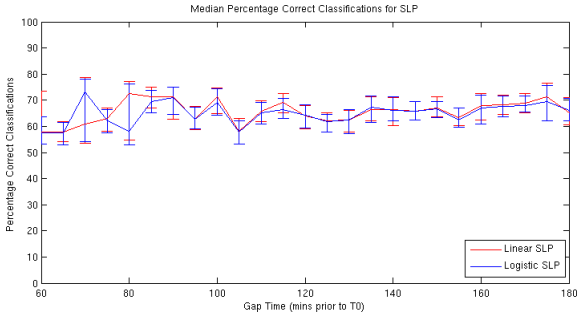
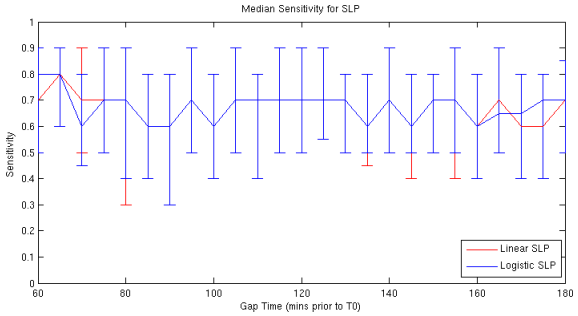


Figure 5.4: *SLP Performance on Unbalanced Dataset*

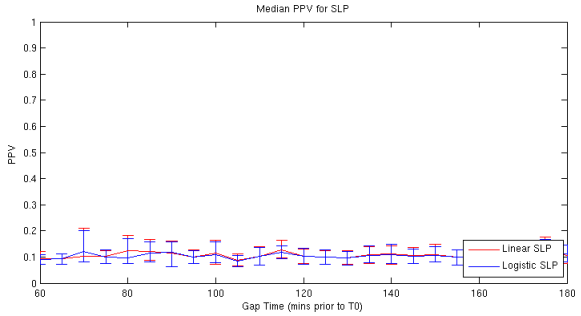
(a) *Percent Correctly Classified*



(b) *Sensitivity*



(c) *PPV*



(d) *NPV*

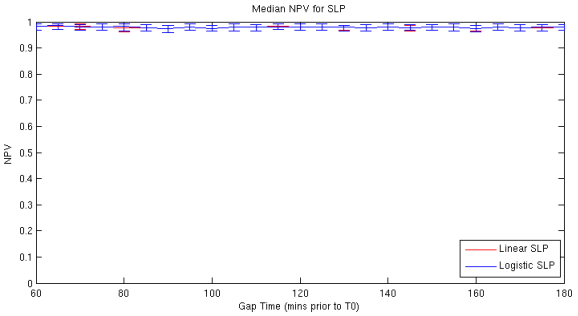
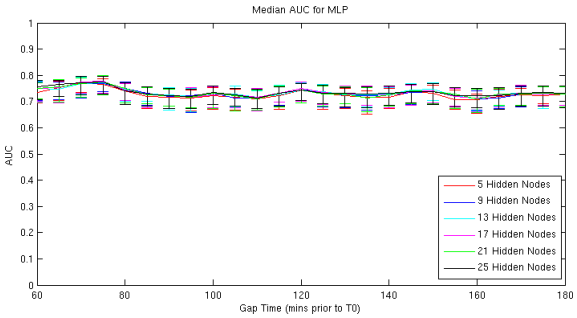
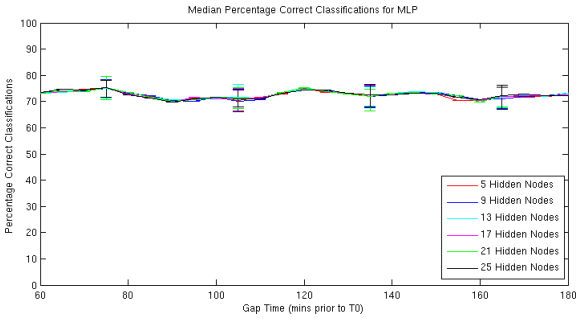


Figure 5.5: *MLP Performance on Balanced Dataset*

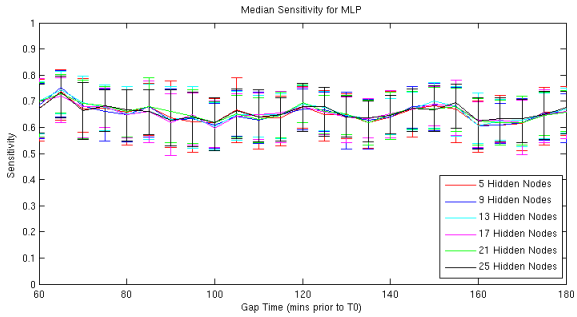
(a) *AUC*



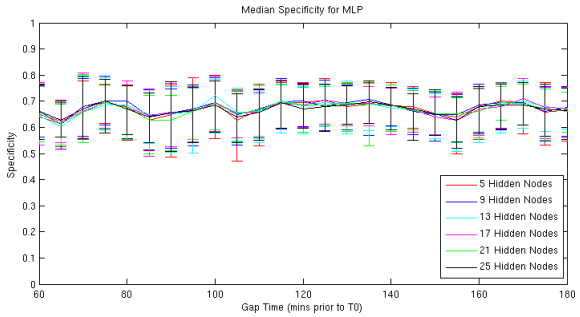
(b) *Percent Correctly Classified*



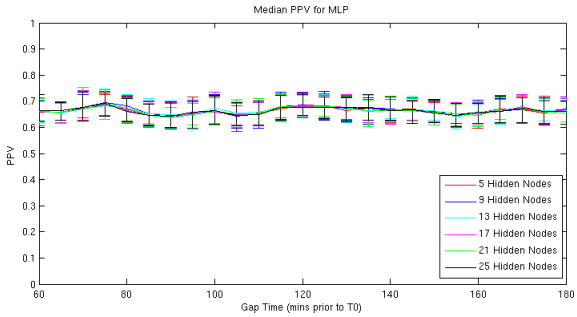
(c) *Sensitivity*



(d) *Specificity*



(e) *PPV*



(f) *NPV*

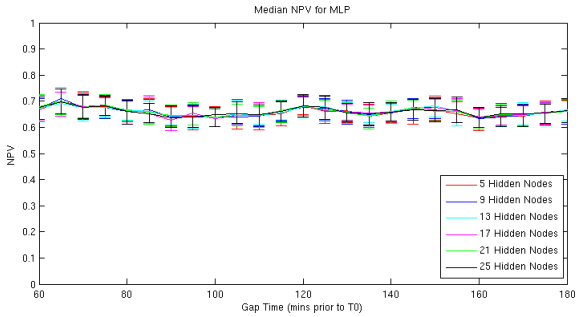
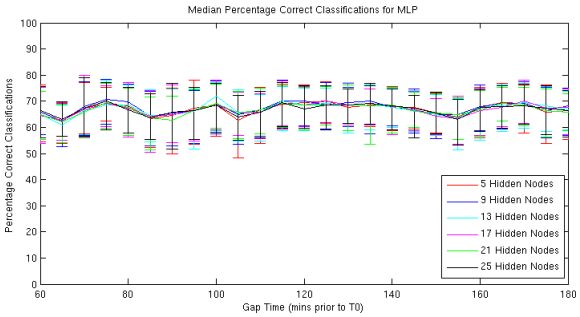
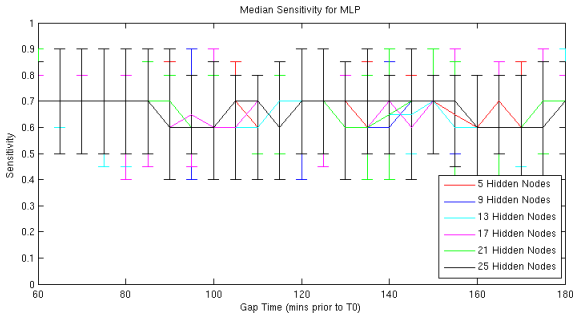


Figure 5.6: *MLP Performance on Unbalanced Dataset*

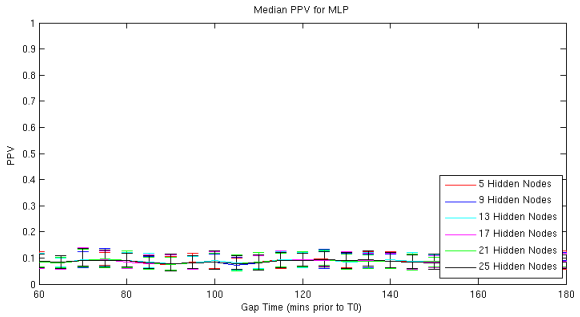
(a) *Percent Correctly Classified*



(b) *Sensitivity*



(c) *PPV*



(d) *NPV*

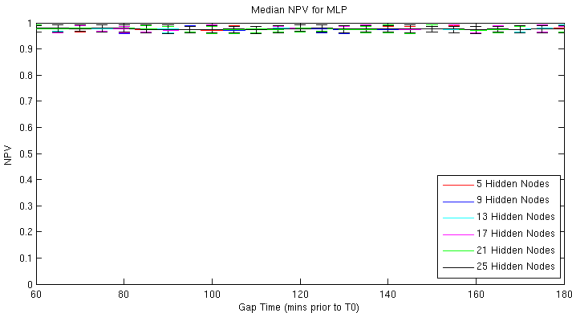
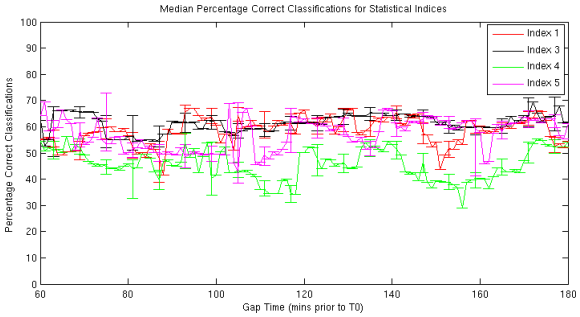
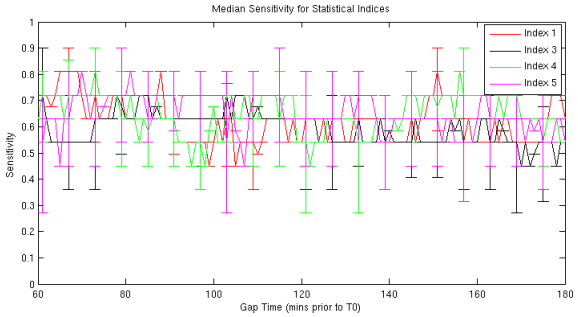


Figure 5.7: *Statistical Index Performance on Unbalanced Dataset*

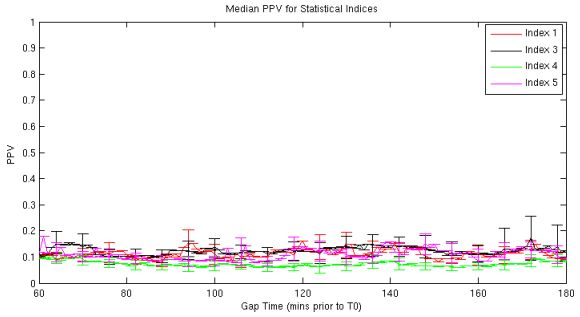
(a) *Percent Correctly Classified*



(b) *Sensitivity*



(c) *PPV*



(d) *NPV*

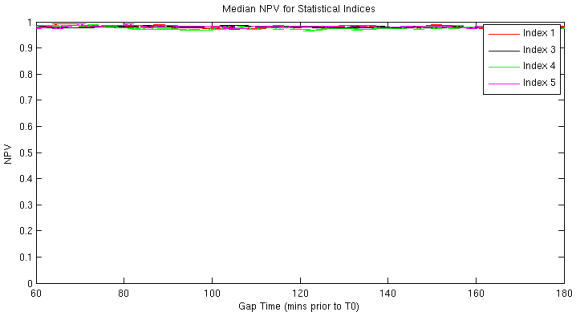
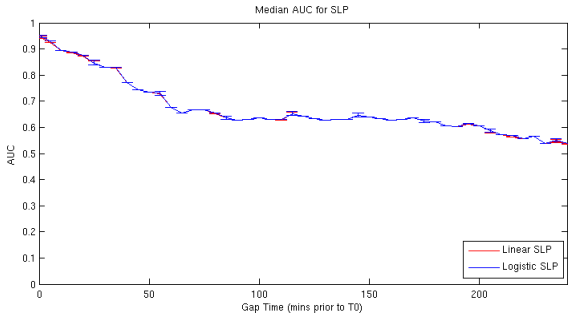
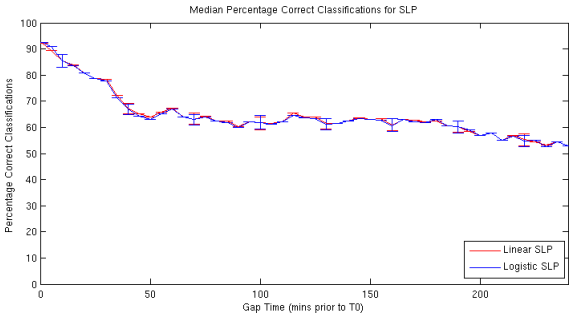


Figure 5.8: *SLP Expanded Gap-Time Performance on Balanced Dataset*

(a) *AUC*



(b) *Percent Correctly Classified*



5.3 Classification Using Parzen Models of Normality

An attempt was made to construct a Parzen Windows model which utilized a dataset specific to our target patient group. The data fusion model evaluated in the previous chapter used cluster centres and variances indicated by the analysis of general-ward patients drawn from specific clinical groups, rather than a more targeted control and positive population for acute hypotension.

Initially, Parzen windows were used on the training data to estimate the PDF of our random variable (ABPMean), allowing for an estimate of the probability that a new data pattern from the test set belongs to the normal or abnormal model. The abnormal model was constructed using training data taken from hypotensive patients (H), while the normal model was constructed using data taken from the control patients (C). As a rule of thumb, the number of examples required to train a model should be at least 10 times the dimensionality of the input patterns. [57] Given 126 training examples in each fold (63 hypotensive and 63 control) the input vectors for the hypotensive and control models could typically have five parameters. A 5-dimensional vector was constructed from each training record in the ABPMean data in three separate ways:

- Model 1: using every data point over 5 minutes
- Model 2: using every third data point over 15 minutes, after median filtering
- Model 3: using every sixth data point over 30 minutes, after median filtering

The Parzen model equation can be expressed as follows:

$$p_m(\mathbf{x}) = \frac{1}{63\ (2\pi)^{5/2}\sigma^5} \sum_{i=1}^{63} e^{-\frac{D_i^2}{2\sigma^2}} \tag{5.5}$$

First, all the training data (hypotensive and control) in each gap size is normalised to have a zero mean and unit variance. Then, the kernel width, σ, can be derived analytically for each dataset as follows. Given a discrete distribution of the variable $\mathbf{x}$ with unknown underlying distribution, we calculate the variance of k samples using:

$$\sigma^2_{sample} = \frac{1}{k}\sum_{i=1}^{k} \parallel x_i - \bar{x} \parallel^2 \tag{5.6}$$

where k is the number of samples of x, and $\bar{x}$ is the sample mean. To calculate the entire training set's variance, we first identify each datapoint's k nearest neighbours (N_i), and then use the set's mean squared distance to provide an estimate of the variance local to each datapoint. The mean of these M local variances is then used as an estimate of the overall variance.

$$\sigma^2 = \frac{1}{M}\sum_{i=1}^{M}(\frac{1}{k}\sum_{j\in N_i} \parallel x_i - x_j \parallel^2) \tag{5.7}$$

Thus the overall average of the mean squared distance between each training pattern and its k nearest neighbours. The variance (σ^2) is calculated every fold for each Parzen model, where k was set to 10. [54] This gives the median (and 10th-90th percentile) values seen in Table 5.1 as a result for the three Parzen models. The appropriate value for σ is then used in Equation 5.5.

Table 5.1: *KNN Sigma Values, K = 10*

5 Minute Sigma	10 Minute Sigma	15 Minute Sigma
1.60 (1.12 - 2.4)	1.69 (1.34 - 2.11)	1.70 (1.42 - 2.18)

Test samples are normalised according to the means and standard deviations found in the training set for each particular gap size. Samples are then evaluated against each of the Parzen models to determine the probability of membership. The probability that a sample is a member of the hypotensive group is denoted as $p_h(x)$, and the probability that it is a member of the control group is denoted $p_c(x)$. If $p_h(x) > p_c(x)$, hypotensive status is predicted. Otherwise, the sample is assumed to be from a control patient.

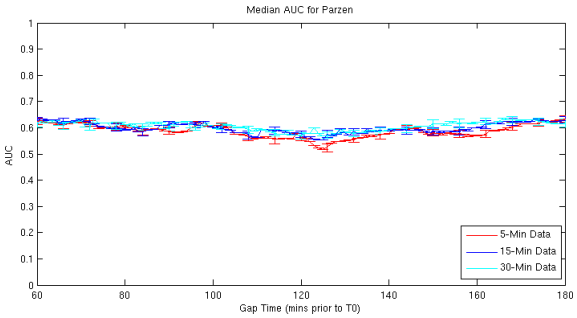
As seen in Figures 5.9 and 5.10, the number of correct classifications in both the test set is near 60% for the three models over time. This is improved from the results obtained in the previous chapter using model parameters obtained from a general hospital

population (50% correct classifications).

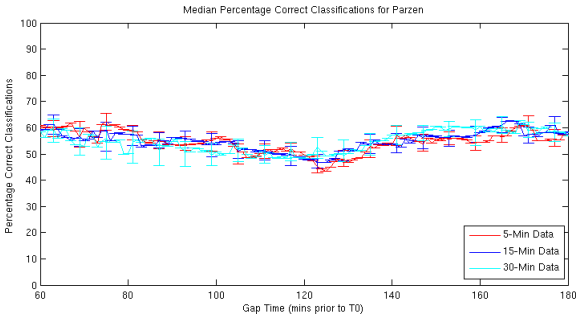
Sensitivity was approximately equal to 0.65 for the models while specificity decreased over time, gradually decreasing to 0.5. Balanced NPV remained near 0.6, while NPV for the unbalanced dataset maintained a value near 0.85. PPV was significantly lower in the unbalanced dataset compared to the balanced dataset (0.3 vs. 0.6), which is expected given the proportion of controls in the unbalanced dataset.

Figure 5.9: *Parzen Model Performance on Balanced Dataset*

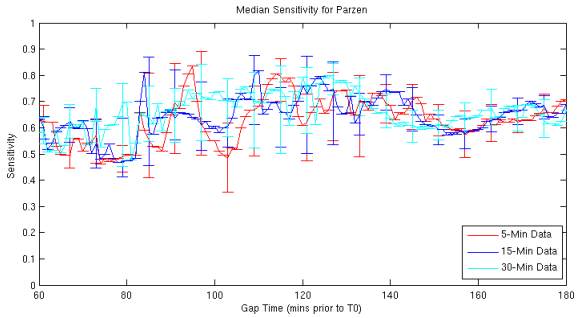
(a) *AUC*



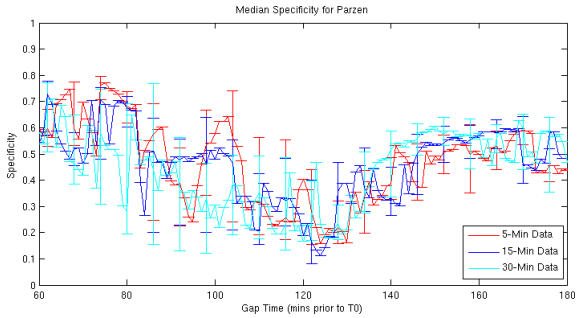
(b) *Percent Correctly Classified*



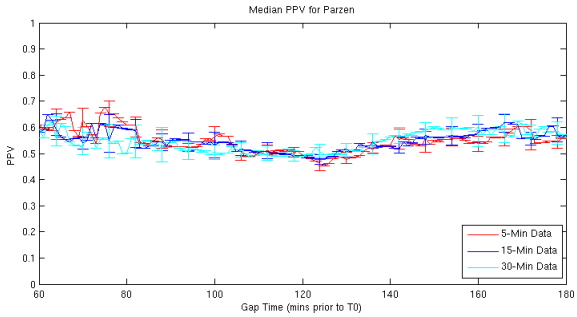
(c) *Sensitivity*



(d) *Specificity*



(e) *PPV*



(f) *NPV*

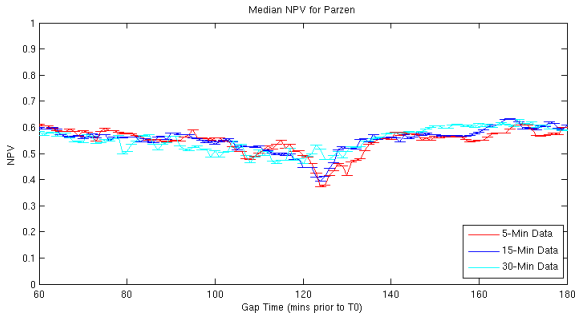
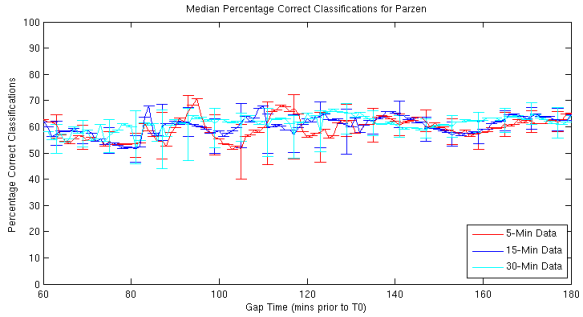
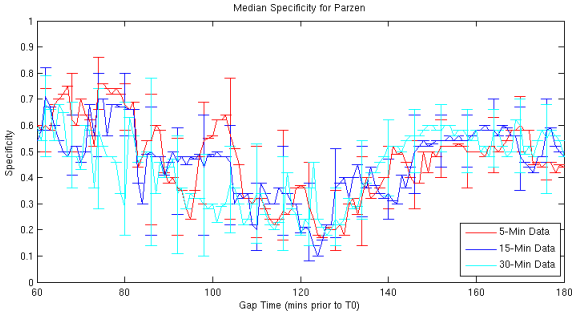


Figure 5.10: *Parzen Model Performance on Unbalanced Dataset*

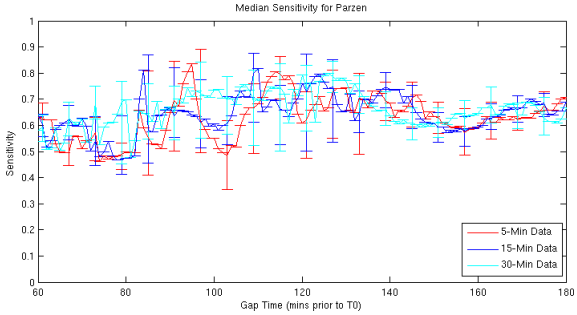
(a) *Percent Correctly Classified*



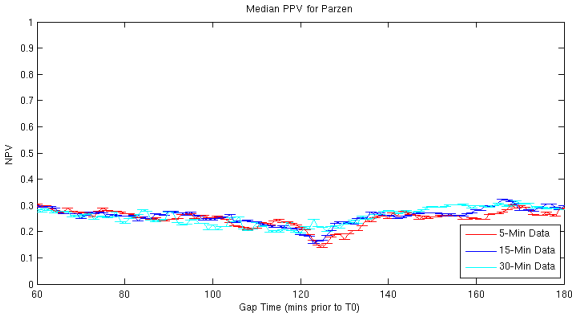
(b) *Specificity*



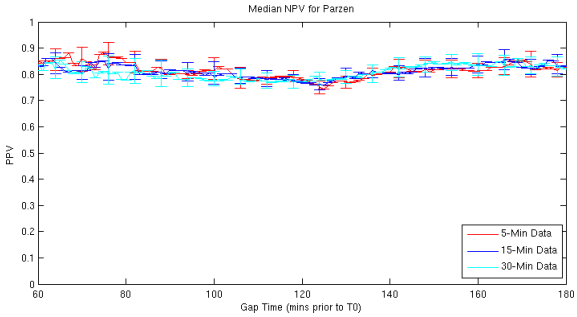
(c) *Sensitivity*



(d) *PPV*



(e) *NPV*



5.4 Discussion and Comparison

Three separate classification methods were used on both balanced and unbalanced test sets to determine their ability to classify AHEs. While the single-layer and multi-layer perceptrons apply non-linear functions to weighted combinations of features extracted from data, Parzen models build a model of the data space defined by each group (H and C), and class membership is evaluated as a probability.

Performance for the SLP and MLP models was comparable, with average classification

accuracy near 68%. Positive predictivity was never better than 0.1, reflecting the prior probabilities. Parzen model performance was worse, averaging near 60%, indicating that the data may not be well distributed for class separation. All results were obtained using only features extracted from the ABPMean signal. The use of multiple features rather than multiple values from a single signal may improve classification performance.

Chapter 6

Additional Variables

As described in Chapter 3, additional numeric and waveform parameters are available in the MIMIC II database. Possible numeric parameters to be considered for prediction of acute hypotension are cardiac output (CO), heart rate (HR), pulmonary artery pressure (PAP), central venous pressure (CVP), respiration rate (RESP), and SpO_2. Clinical data such as medications given, fluids administered, and other chart data are also available to provide additional information. In addition to the variables themselves, the variability in any given parameter may contain predictive information. Here we examine whether the use of some of these parameters may improve the predictive power of any of the indices or models so far considered.

6.1 Review of Variables

In order to further identify other features that may be relevant to the presentation of an AHE, signals with all values present in at least one patient in each sub-group (H and C) were plotted for all hypotensive and non-hypotensive data points. Once values are normalised for the sample size, the value of the variable as a predictive indicator is determined by examining offsets in the distribution of values seen.

Variables that were examined are cardiac output (CO), pulmonary artery wedge pres-

sure (PAWP), temperature (TEMP), heart rate (HR), pulmonary artery pressure (PAP), central venous pressure (CVP), respiration rate (RESP), SpO_2, and non-invasive blood pressure (NBP). Not all patients had all of these variables collected, so only the subset of patients with these variables was reviewed. A summary of the analysis can be seen in Table 6.1 below.

Table 6.1: *Variable Presence*

	C	H	Percentage of Total
HR	84	84	100%
ABPSys	84	84	100%
ABPDias	84	84	100%
ABPMean	84	84	100%
SpO_2	84	84	100%
RESP	83	83	98.8%
NBPSys	75	73	88.1%
NBPDias	75	73	88.1%
NBPMean	75	73	88.1%
CVP	43	57	59.5%
PAPSys	31	39	41.7%
PAPDias	31	39	41.7%
PAPMean	36	43	47.0%
CO	19	22	24.4%
PAWP	2	13	8.9%

Of the 16 variables, CVP, PAPSys, PAPDias, PAPMean, CO and PAWP were not present in sufficient numbers of patients to be used in multi-parameter models. Additionally, the NBPSys, NBPDias and NBPMean data was of low quality in a majority of patients as to be rendered non-usable. As a result, the remaining variables to be considered are HR, ABPSys, ABPDias, ABPMean, SpO_2, and RESP. In the appendices, a normalised histogram is shown for each of the variables across the two subgroups: H and C. Note that the histograms for the three ABP metrics (examined in previous chapters) are also plotted to confirm that they are indeed significant. For visualization purposes,

the fitted normal distribution is also plotted. (Appendix A, Figure E.1)

6.2 Shape of Distributions

All variables were tested for normality using Lilliefors test. First the variables were normalised to have zero mean and unit standard deviation. The test statistic T is defined as the largest distance between the cumulative frequencies of a truly normal variable and the normalized variable. The null hypothesis was that the random sample comes from a population with a normal distribution. The null hypothesis will be rejected at our approximate level of significance if T exceeds the value given in statistical tables.

Lilliefors test rejected the null hypothesis (H = 1) at the 5% significance level for all variables. This implies that none of the variables were statistically close to a normal distribution, and that the effects of a correlative test should therefore be genuine.

6.2.1 Sub-group Distribution Similarity

Given that none of the variables are normally distributed, the presence of a significant difference in the subgroup distributions was evaluated using the Kolmogorov-Smirnov test (KS-test). The KS-test determines whether the values in two datasets are from the same continuous distribution (the null hypothesis) or different continuous distributions. The result is 1 if the test rejects the null hypothesis at the 5% significance level, and 0 otherwise. The H distribution was compared to the C distribution in order to determine if there was a significant difference. As summarised in Table 6.2, the KS-test rejected the null hypothesis (H = 1) for the normalized distributions of ABPSys, ABPDias, ABPMean, HR, RESP, and SpO_2. This implies that these variables exhibit a significant shift in distribution between the AHE presenting (H) and non AHE-presenting (C) sub-groups.

Table 6.2: *Kolmogorov-Smirnov Test Results*

	H vs. C
ABPSys	1
ABPDias	1
ABPMean	1
HR	1
RESP	1
SpO_2	1

6.3 Variable Similarity

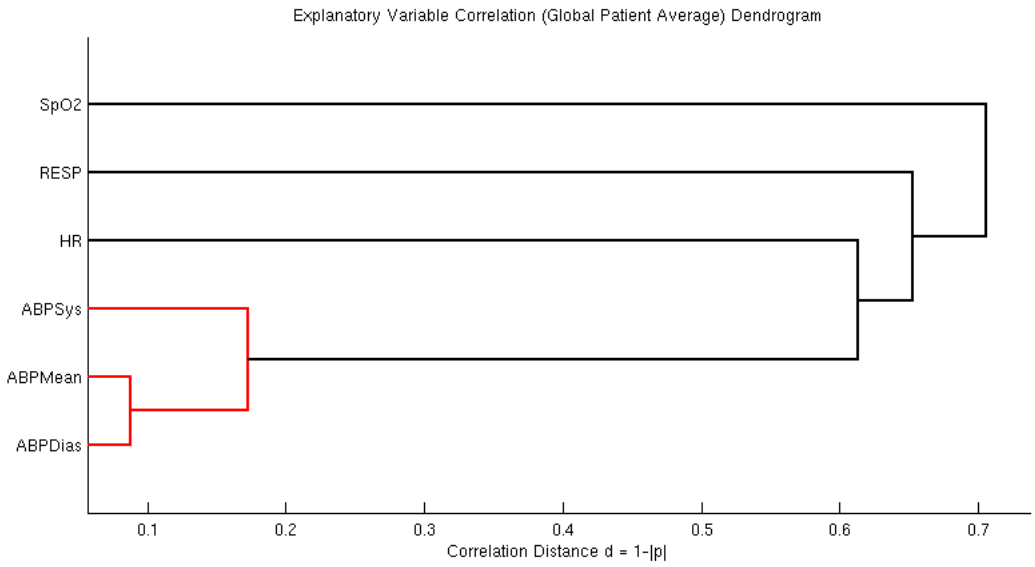
Given our list of identified predictive variables (ABPSys, ABPDias, ABPMean, HR, RESP, and SpO_2), and the similarity in the physical signs that they measure, it is useful to identify whether there are variables which only duplicate the predictive value of others. In order to identify these extraneous variables, we studied the correlation between variables. The correlation equation produces a set of correlation coefficients which can be used to determine the amount of variance in one variable that can be explained by the other. The correlation coefficients (ρ) can be calculated to determine the possibility of a linear relationship between two random variables. ρ can have any value in the range -1 to $+1$, where 0 implies no linear relationship, $+1$ implies perfect positive correlation, and -1 implies perfect negative correlation. When squared (and multiplied by 100) the coefficient gives the percentage of relation in the variable's variation in the other, and the correlation distance d between the two variables is calculated as $d = 1 - |\rho|$.

Figure 6.1 contains a dendrogram representing the similarity between the explanatory variables, where the distance on the horizontal axis displays variable similarity. This means that d = 0 represents perfect correlation and d = 1 represents complete independence. The dendrogram groups the most closely related variables into smaller branch/leaf nodes. Signal correlation was performed for each patient's data for all signals which were valid for at least 20% of the test window (6 hours).

The results were then averaged across the C sub-group, H sub-group, and all patients

to obtain a global mean correlation between all variables. The results for the three groups yielded the same dendrogram, shown in Figure 6.1, indicating that several of the signals were closely related, with tight groupings in the three ABP signals.

Figure 6.1: *Dendrogram of Variable Correlation*



Choosing the most prevalent variable from each correlated grouping narrows the list of relevant variables to ABPMean, HR, RESP, and SpO_2.

6.4 Conclusion

A total of 16 variables were examined to determine whether they could be used as relevant features in the presentation of an AHE. Four variables were found to be both consistently present in the patients, and uncorrelated with one another: ABPMean, HR, RESP, and SpO_2. These variables were extracted for all patients, and filtered as noted in Chapter 3. In the following Chapter, variables, and various aggregates, are used as model features.

Chapter 7

Long-term AHE Prediction Using Multivariate Classification Methods

No univariate index or model was able to achieve better than 75% correct classification across gap sizes. For many of the algorithms, performance was significantly worse as the gap size increased. In Chapter 6, additional relevant variables were identified for possible incorporation into a predictive model. To this end, multivariate models are explored in this chapter to investigate whether more robust classification is possible.

7.1 Multivariate Neural Networks

Features were extracted from the ABPMean, HR, RESP, and SpO_2 signals using the same methods as previously described (individual values, percentage change from one reading to the next, 5-minute mean, 5-minute variation ((max-min)/mean), 5-minute median, and 3-hour exponentially weighted mean). This led to a total of 24 variables; models were again run for every fifth gap time (every 5 minutes starting at 60 minutes prior to AHE onset). Each feature was again normalised using a zero-mean, unit-variance transformation to help avoid poor local minima, and MLP neural network models were re-run with the multivariate data as inputs.

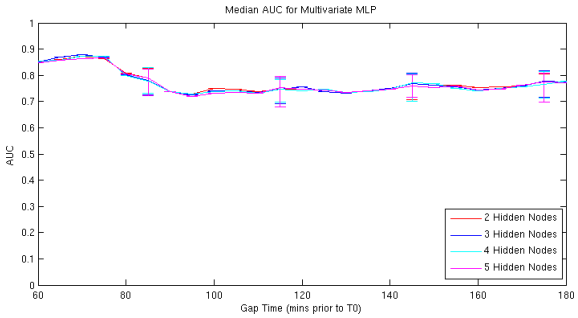
As described in Section 4.1.2, at each of the ten independent folds there are 126 patients in the training set, and the remaining 42 patients form the test set. When training the MLP, the test set was split in half to form a validation set and a test set (each with 21 patients). The validation set is used during training as an internal check as to whether the network is learning the specific training data or the underlying mapping. Training is stopped if the network classification performance on the validation set decreases during any training epoch by a factor of 10, or if performance decreases by any factor during more than 5 epochs; this was done in order to prevent loss of generalisability. Given 24 features, a training set of 126 patients is unable to maintain the desired ratio of 10 times as many training patterns to free network parameters. [57] In order to prevent the network from becoming severely under-specified, a total of 2, 3, 4 and 5 hidden nodes were used for simulation.

As seen below in Figures 7.1 and 7.2, the use of multivariate input data improved classification performance. In the balanced test set, correct classifications on the test set were approximately 82% up to the 80-minute gap time, after which classification accuracy averaged near 67%. The improved classification reflects the improved ability of the neural network to create models with a complex relationship between the input variables and the outcome. In the unbalanced case, the percentage of correct classifications reached 84% before dropping to an average value of 70% beyond the 80-minute gap time. Sensitivity and specificity were similar, and also decreased after the 80-minute gap time. While sensitivity and PPV were slightly higher than specificity and NPV in the balanced test set, the positive predictive value was still low in the unbalanced test set due to the low incidence rate (one positive case versus 21 controls in each of the folds).

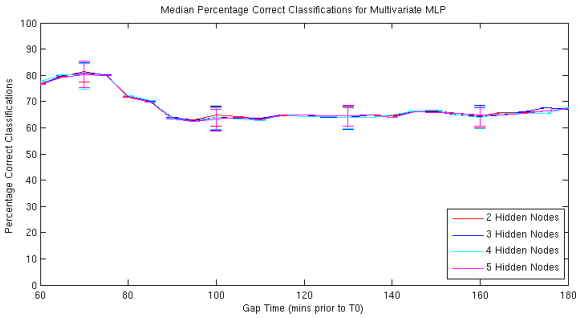
By using features generated from variables other than just the blood pressure, classification performance increased to higher values across the longer gap times. However, the low PPV indicates that (in a real ICU setting) many of the model's positive results would be false positives.

Figure 7.1: *Multivariate MLP Performance on Balanced Dataset*

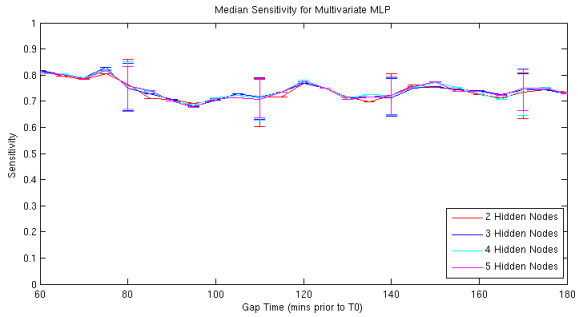
(a) *AUC*



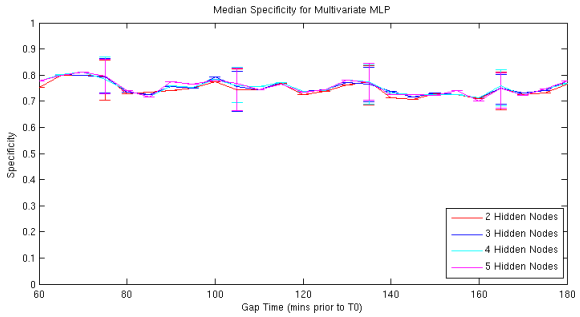
(b) *Percent Correctly Classified*



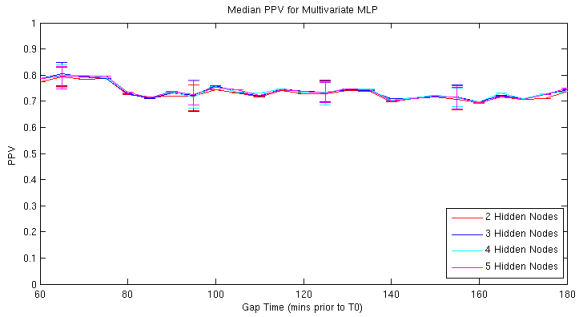
(c) *Sensitivity*



(d) *Specificity*



(e) *PPV*



(f) *NPV*

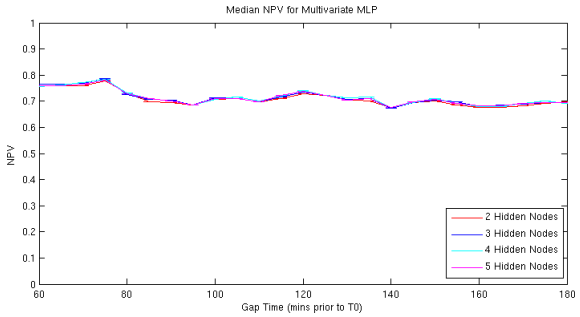
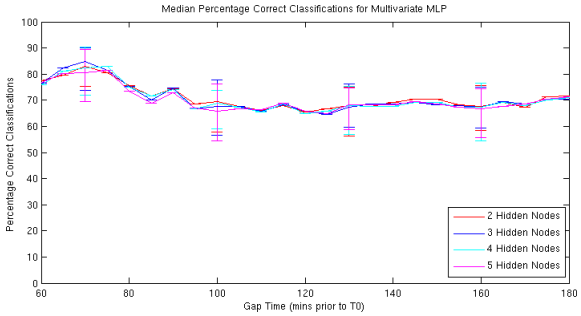
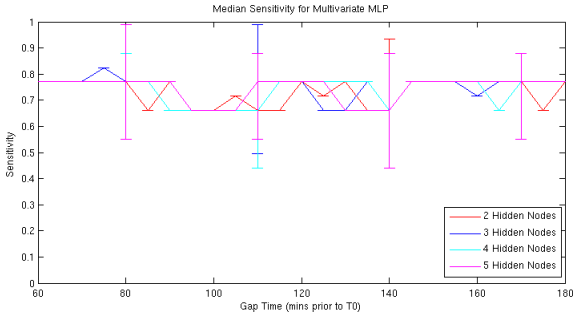


Figure 7.2: *Multivariate MLP Performance on Unbalanced Dataset*

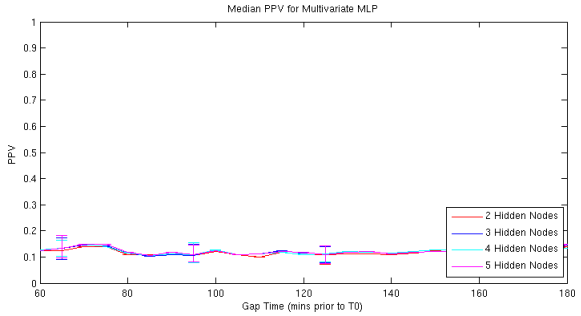
(a) *Percent Correctly Classified*



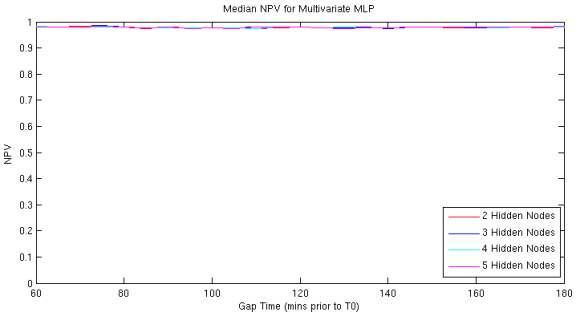
(b) *Sensitivity*



(c) *PPV*



(d) *NPV*



7.2 Multivariate Logistic Regression Models

Multivariate logistic regression models are commonly used in prediction, and have successfully been used in the past to predict the onset of septic shock, for which AHEs are a critical determinant. [58]

7.2.1 Septic Shock Early Warning System

In his 2007 Master's thesis, Shavdia focused on the development of a septic shock early warning system (EWS) which uses commonly measured clinical variables. [58] The EWS takes the form of a multivariate logistic regression model designed to predict the hallmark of septic shock: hypotension despite fluid resuscitation. Shavdia also drew his dataset from the MIMIC II database, choosing to use nurse-verified data with a minimum data threshold that corresponded to approximately one day of ICU data: two white blood cells counts and 10 hours of systolic blood pressure, heart rate, temperature, and respiratory rate. Patients with sufficient data were then pre-annotated as being in states of (i) non-sepsis, (ii) sepsis; (iii) severe sepsis without septic shock, or (iv) septic shock. Additionally, in patients that progressed to septic shock, there was an attempt to identify the transition point from sepsis or severe sepsis to septic shock. For both of these tasks, the definitions used for systemic inflammatory response syndrome (SIRS) and septic shock were those given in the 1991 ACCP / SCCM Consensus Conference. Finally, the pre-annotations were confirmed or rejected by manual review. Of the 17,082 patient records in MIMIC II, 2.7% (459) had an ICD-9 septic shock code. A subgroup of 459, 56.9% (261 of 459) had sufficient data for analysis, and of those 261 patients, 250 exhibited SIRS. It is these 250 patients that were used in the EWS developed by Shavdia.

Clinical variables extracted for each patient included systolic, diastolic and mean blood pressure (BP), pulse pressure (PP), central venous pressure, heart rate, temperature, oxygen saturation, respiratory rate, and arterial pH. Lab values for lactate, creatinine, troponin, white blood cell count (WBC), and creatine phosphokinase (CPK)

were also extracted. Any fluids or medications administered were noted, and cardiac output/peripheral resistance were calculated using the Liljestrand technique, which estimates cardiac output using heart rate, pulse pressure, and the systolic/diastolic arterial pressure as $CO = \frac{PulsePressure}{ABPSys+ABPDias} * HR$. [59].

185 of Shavdia's 250 patients exhibited sepsis or severe sepsis while in the ICU and 65 progressed to septic shock. For his purposes, patients that did not progress to septic shock were classified as negatives and those that did were classified as positives. Shavdia constructed six multivariate logistic regression models designed to differentiate between patients who progressed to septic shock and those that did not. Model performance was evaluated in a forward/causal way using a random sample of patients. The six models varied chiefly in the selection of a reference time point to gather physiological values. For septic shock patients reference times of 30, 60, 90, 120, 180, or 240 minutes prior to shock onset were used.

All models selected systolic blood pressure as the best indicator for progressing to septic shock, and eight other high-risk factors were identified: SpO_2, arterial pH, systolic BP, PP, HR, cardiac output, RESP, and white blood cell count.

The six models were run for gap times of 30 – 240 minutes prior to the onset of shock in increments of 30 minutes; note that this is compared to our considered gap times of 60 - 180 minutes prior to AHE onset in increments of 5 minutes. The overall performance of the six models was very similar. Models were tested on a random group of 500 patients with sufficient data, and a range of 1-5 consecutive model outputs were combined using weighted and unweighted sums. The summing algorithms used different methods (exponentially weighted sums, linearly weighted sums, etc.) to combine the output from 2-5 previous prediction points into a single classification output. This was done in order to reduce the impact of miscalculation at a single prediction point. A total of 21 classifiers for each of the six models were tested (single unmodified model output and four additional for each summing algorithm combining 2 to 5 consecutive outputs). This adds up to a total of 126 different classifiers for predicting the onset of an episode

of hypotension despite fluid resuscitation (HDFR).

The best model was the EWS model for prediction at 120 minutes prior to onset with no summing algorithm. The model was able to provide early warning with high specificity (upper 80% to mid 90%) and sensitivity (mid 80% to upper 80%), but displayed a substantially lower positive predictive value (PPV) of 70%. The author suggests that future work should focus on increasing the PPV of the EWS, and further stratifying early warnings into septic and non-septic aetiologies.

7.2.1.1 Logistic Regression

Logistic regression makes use of the logistic function to map any set of independent input values to a binary value. Using all relevant independent variables x, a measure z is defined representing the contribution of all variables. The variable z itself is a linear combination of regression coefficients (β) and variables (x), which can be written as $z = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$ for n variables. This is then transformed to $f(z)$ in the range of 0 to 1 using the logistic function $f(z) = \frac{1}{1+e^{-z}}$, where $f(z)$ represents the probability of a specific outcome for the variables given.

7.2.2 Logistic Regression Classifier

Using the same features as noted previously in Chapter 7.1, a logistic regression classifier was trained using a total of 24 possible features. Initially, the training data is used to construct a logistic classifier for AHEs. The training dataset is comprised of two types of patients: those who present an AHE (outcome = 1), and those who do not present an AHE (outcome = 0). The data is therefore already binomially distributed.

The classifier was trained using a greedy forward method to select the top n variables from the feature matrix as follows. Initial models were built as univariate logistic regression models, and performance results were judged based on the calculated AUC. (Chapter 4.1.1.3) Subsequent models were built by adding each of the non-selected variables to the feature set, and the best performance amongst these was again identified

using AUC. During the training process, variables were not considered if their p-value was greater than 0.10 (that is, if there was more than a 10% probability of the observed correlation values being generated by random chance, when the true correlation is zero). The training process continued until the AUC improvement was less than 1% for the addition of any further variables.

7.2.3 Optimal Variable Selection

The dataset was again split using 126 points as training data and 42 points as testing data at each iteration. Logistic regression was then run 100 times to determine the most predictive variables for each time gap (from 60 to 180 minutes prior to AHE onset in steps of 5 minutes). Shown below in Table 7.1 is the mean number of times (across the 100 runs) each variable was used over all gap sizes, e.g. if ABPMean was used 50% of the time for the 60-minute gap time models, 25% of the 65-minute gap time models, and not used at all in any other gap time, the mean percentage use would be 3% (mean of 25, 50 and 23 zeros).

The most prevalent variables used to form the optimal logistic regression models were the Three-Hour Weighted Mean ABPMean, Three-Hour Weighted Mean RESP, Three-Hour Weighted Mean SpO_2, ABPMean, and Five-Min Mean SpO_2. It is worth noting that the Three-Hour Weighted Mean ABPMean and RESP only begin to dominate model creation after the 85-minute and 120-minute gap times respectively. After these points, they are the first and second model parameters selected.

Table 7.1: *Logistic Regression Variable Use*

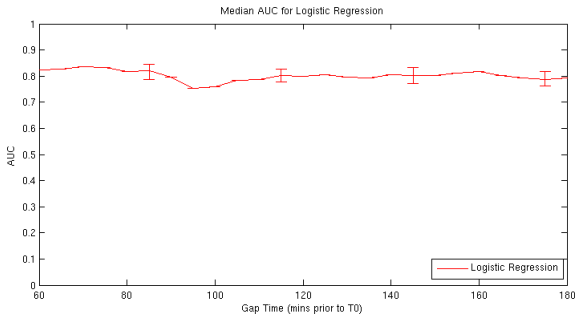
Variable	Percentage Used
ABPMean	16%
HR	0.48%
RESP	13.12%
SpO_2	12.88%
Percent Change ABPMean	1.88%
Percent Change HR	0.12%
Percent Change RESP	0.16%
Percent Change SpO_2	0.32%
Five-Min Deviation ABPMean	5.12%
Five-Min Deviation HR	3.20%
Five-Min Deviation RESP	3.16%
Five-Min Deviation SpO_2	6.20%
Five-Min Mean ABPMean	13.72%
Five-Min Mean HR	0.28%
Five-Min Mean RESP	6.48%
Five-Min Mean SpO_2	14.48%
Five-Min Median ABPMean	9.12%
Five-Min Median HR	0.36%
Five-Min Median RESP	5.20%
Five-Min Median SpO_2	8.28%
Three-Hour Weighted Mean ABPMean	83.24%
Three-Hour Weighted Mean HR	1.16%
Three-Hour Weighted Mean RESP	62.40%
Three-Hour Weighted Mean SpO_2	39.80%

7.2.4 Model Performance

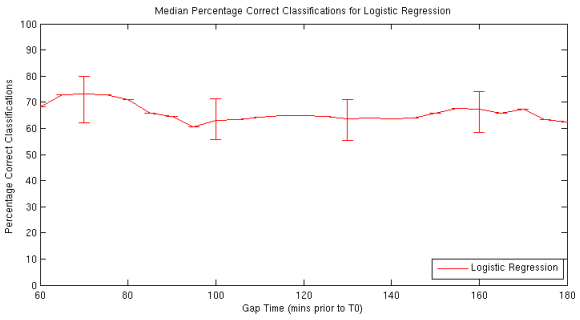
The performance of the logistic regression model over all time gaps was evaluated on the testing data using the optimal point selected by each model's ROC curve. The AUC and percentage of correct classifications are shown in Figure 7.3. As seen in the Figures, the logistic regression classifier does not outperform the MLP models in the balanced and unbalanced case. The correct classification percentage is 73% to 62% in the balanced case, and 67% to 60% in the unbalanced case.

Figure 7.3: *Logistic Regression Model Performance*

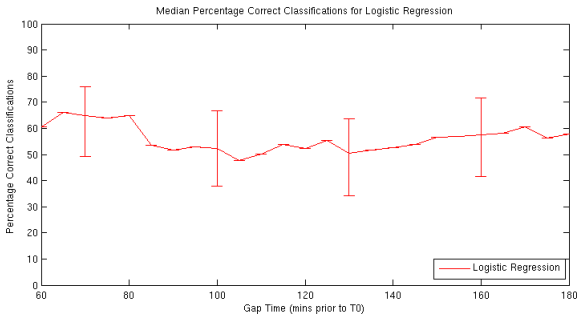
(a) *AUC Performance*



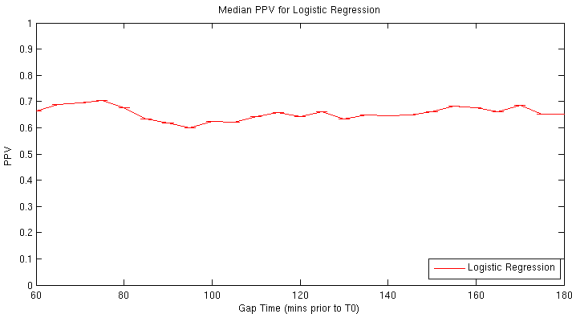
(b) *Balanced Correct Classification Performance*



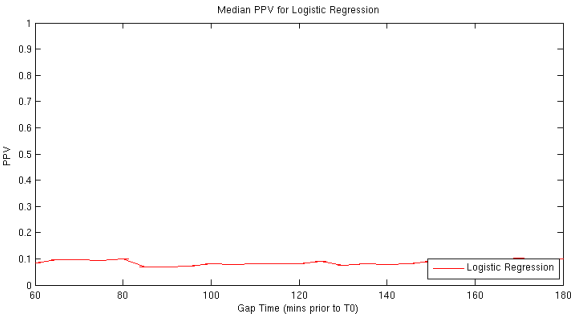
(c) *Unbalanced Correct Classification Performance*



(d) *Balanced PPV*



(e) *Unbalanced PPV*



7.3 Use of Model

As a demonstration of how the both regression and MLP models might be used in practice, the best model generated for any gap time in both cases was applied to the testing data in a forward manner to mimic an ICU alert system. This was the 65-minute model for the regression model and the 2 hidden node 70-minute MLP model.

The models produced an output, called an alert, in the range of 0 - 1 for each time point, which was then used to determine whether an AHE would occur in the future. Four different prediction schemes were evaluated, whereby an AHE was predicted if there were:

1. four or more alerts in any five minute window
2. seven or more alerts in any 10 minute window
3. 10 or more alerts in any 15 minute window
4. 13 or more alerts in any 20 minute window

Note that these limits correspond to more than 60% of a contiguous time period containing alerts. If the output of the classifier indicated that there was an alert for the preceding criteria, the patient was classified as eventually belonging to an AHE presenting state. Otherwise the patient was classified as being in a control state. An example of the use of this method is shown below in Figure 7.4.

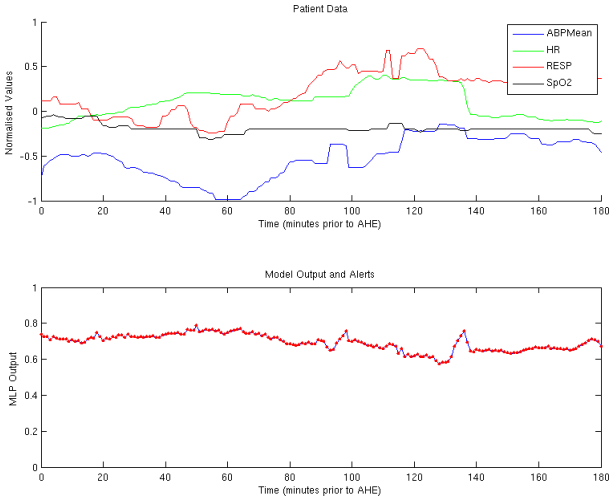
The logistic regression classifier was compared against the MLP model, with performance on a balanced test set similar to that achieved previously. The MLP model performed best when predicting an AHE after four or more alerts in any five minute window, with a correct classification performance of 78.57%. The logistic regression model performed best using seven out of any 10 or 10 out of any 15 minute window, achieving 71.43% correct classifications. (Table 7.2)

Table 7.2: *Logistic Regression Variable Use*

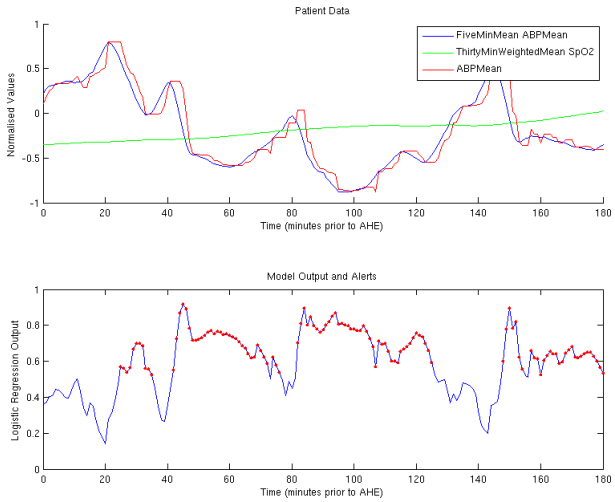
Alert Minutes/Window Minutes	Patients Correctly Classified	
	Logistic Regression	MLP
4/5	66.67%	78.57%
7/10	71.43%	76.19%
10/15	71.43%	76.19%
13/20	69.05%	76.19%

Figure 7.4: *Prediction of Acute Hypotensive Episode*

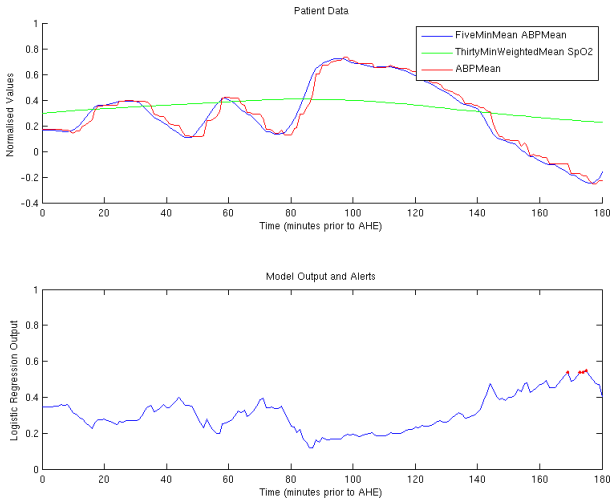
(a) *Model Output For Patient 48*



(b) *Model Output For Patient 22*



(c) *Model Output For Patient 7*



7.4 Discussion

Two different classification methods were used on both balanced and unbalanced test sets to determine their ability to classify AHEs. The performance of MLP models was improved by the addition of physiological variables other than blood pressure, with a top AUC of 0.88 and 82% correct classifications. The logistic regression classifier performed worse than the MLP model. This is likely due to its lack of a hidden layer, which allows MLPs to capture non-linear data relationships. Logistic regression is essentially a single layer perceptron with a logistic output function.

Shavdia's EWS performed considerably better than the logistic regression model trained on our data. There are several possible reasons for this difference. First, Shavdia's logistic regression classifier focused on determining whether a patient would experience septic shock rather than sepsis through identification of hypotension despite fluid resuscitation. Thus, his training group was tightly focused around those patients with sepsis, and varied only on whether they had progressed to shock. Second, Shavdia's data source is nurse-verified information, which has different sources of error than physiological data acquired from patient monitors. [60] While less frequent, data of this nature is also more plentiful in the MIMIC II database, yielding a larger patient base. Finally, our classification problem is more general than Shavdia's, and thus a harder task. Our training group is diverse, and draws upon hypotensive and control patients with many diagnoses, care units, etc. rather than focusing on a specific subset of patients with a given condition (sepsis).

A summary table of the perfomance of the best algorithm are shown in Table 7.3. The first two reported results (Chen and Ghassemi) are for the balanced MIMIC II dataset extracted for this thesis. The metrics reported are the best performance achieved after the one hour gap time as measured by the percentage of correct classifications. Lee's noted performance is based upon his own published results. Note that Lee's feature entry includes the principal component analysis (PCA) step as a layer; this is because

PCA can be performed by an MLP using a linear activation function.

Table 7.3: *Comparision of Performance*

Author	Model	Features	Correct Classification	Sens.	Spec.
Chen et al.	Five-min mean of numeric ABPMean	1	75%	0.87	0.71
Ghassemi	MLP: 24 in, 3 hidden, 1 out	24-3-1	82%	0.83	0.82
Lee et al.	MLP 45 in, 15 principal components, 20 hidden, 1 out	45-15-20-1	87%	0.83	0.88

Chapter 8

Conclusion

The goal of the work described in this thesis was to derive methods for predicting the occurrence of AHE with a lead time which would be of clinical usefulness within the ICU. To that end, we have demonstrated acute hypotensive episode prediction performance with a minimum 1 hour gap time.

Models were based on data acquired from a large-scale ICU database (MIMIC II) indicating that its results are likely to be credible if extended to real ICUs. In our particular dataset, the number of patients with pressor medications was higher in the hypotensive group (82.1% vs. 51.2%), but hypotensive ICD9 codes were not found to correlate well with the data groups (13.1% in H vs. 4.8% in C), as they were not present at a level which might be expected for a group of patients with identified AHEs. This is likely due to the broad view that ICD9 codes represent: they tend to correlate with key clinical syndromes and diagnoses, but not with detailed physiologic data.

Parzen windows were used to construct models of normality (control patients), but were not effective at detection. This may be due to our patient population, as control patients are drawn from the ICU (thus likely to exhibit abnormal physiological values) and hypotensive patient records are only used prior to the onset of their first AHE (thus guaranteed to exhibit no ABPMean value less than 60 mmHg). However, one interesting result of this exploration was the different types of abnormality exhibited by control

versus hypotensive patients. The most abnormal hypotensive patients had a consistently low ABPMean, elevated HR, elevated RESP and lowered SpO_2 level, which were distinct from those seen in the most abnormal control patients.

Neural network models were built using statistical indices which performed well on the 2009 Computers in Cardiology datasets as input features, first exclusively with blood pressure data. Models were not able to distinguish between control and hypotensive patients using only the ABPMean data, so three additional vital signs (HR, RESP SpO_2) were added as potential feature sources. Ultimately, MLP models were found to be the best predictors of AHEs, as evaluated with our dataset. Data consisting of a patient's ABP, HR, SpO_2 and RESP (along with first-order statistical transformations of these values) were used as the model parameters, yielding a best AUC of 0.84 at a 60-minute gap time. This corresponded to a correct classification percentage of 82% in the balanced case and 84% in the unbalanced case.

The evaluated models used simple features generated only from first-order statistical functions of prior data. The fact that these models were able to perform classification of 84% accuracy at a 60-minute gap time indicates that there are early patterns underlying general hypotension detectable in patient physiology. Our results also clearly show that overall performance degrades as gap size increases; this is as expected, as predicting further in the future is more difficult. However, there are further issues for exploration and improvement.

One critical issue is the low PPV of models in the unbalanced test set. All of the top algorithms considered (see Table 7.3) achieved a PPV of less than 0.16, which means that more than 84% of all alarms generated would be false. While the low PPV achieved by all models in the unbalanced test set is attributable to the large imbalance between hypotensive and control patients, similar prediction performance could be expected if such a system were to be used in the ICU as an early predictor of hypotension. In this case, the prediction algorithms could be used to identify patients at an increased risk only, e.g. alert nursing staff to patients who needed more attention. A gap-time of more than

an hour prior to onset would support this usage model: by noting an increased risk an hour before an AHE's likely occurrence, clinical staff could be more attentive to a given patient's progress, allowing for an earlier intervention. This might be combined with a triggering system based on whether a patient had needed an intervention previously (pressors, fluid resuscitation, etc.), or led by the physician. By setting patient monitoring to be clinician initiated, staff would esentially be using the alerts as a prompt for more regular patient evaluation. Clinician-initiation would also provide additional incentive for caring staff to respond promptly to the alerts. By having the alert system run only on those patients who are more likely to be hypotensive, we are essentially re-balancing the prior distribution of patients, thus leading to a better PPV score (as shown by the balanced datasets).

As noted by Lee [3], a limitation of any binary classification system is the inherent need to assign a single threshold to distinguish between control and hypotensive patients. It is of course possible that blood pressure values above 60 mmHg could constitute hypotensive episodes for some individuals. Additionally, in such a diverse patient population as studied in this thesis, hypotension may be due to any number of root causes. While the high ratio of CCU and CSRU patients indicates that most of the patients are not experiencing hypotension due to hypovolaemia, the root causes underlying an event as broadly defined as "experiencing an AHE" are diverse. This makes the target of our prediction difficult: many different individual states and subsequent paths through an ICU lead to an eventual AHE. Attempting to learn these complex mappings would require a much larger example set with many examples of the possible states and paths. One key reason the model may have a low PPV in the unbalanced case is a lack of training examples which cover the physiology for a number of the root causes of hypotension. A possible way of determining the hypotensive cause could be to examine the text-based discharge summaries for each patient. These documents often list the discharge diagnosis, and the relevant parts of the patient's physiology towards that diagnosis. Another possibility is to attempt to identify patient clusters using a clustering algorihtm such as KNN. Once

rough groups of hypotensive patients are identified, their general characteristics could be explored in order to determine what the possible underlying causes might be for the hypotensive event.

The quality of the data itself is always a factor in modeling applications. A recent study of the MIMIC II database's signal quality noted that nurse-verified data was critically undersampled and introduced human error. [60] Previous studies have shown that the signal quality of MIMIC II's data varies within a given patient record, and the use of robust signal quality metrics could be a critical way of reducing the false positive rate. [61–63]

Finally, as these patients are in the ICU, intervention, especially pressor administration, is a key modifier of their physiology. While pressor admission was not included as one of the standard physiological signals for modelling, it may be valuable to consider it as a model parameter in the future. Accurate information on the type, amount, and time of pressor administration is available in the MIMIC II database, and may help to provide a more accurate prediction of hypotension during the patient's stay in the ICU.

Bibliography

[1] A. Kumar, D. Roberts, K. E. Wood, B. Light, J. E. Parrillo, S. Sharma, R. Suppes, D. Feinstein, S. Zanotti, and M. Taiberg, LeoDavid Gurka, MD; Aseem Kumar, PhD; Mary Cheang, "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock," *Critical Care Medicine*, vol. 34, no. 6, pp. 1589–1596, 2006.

[2] E. Rivers, B. Nguyen, S. Havstad, J. Ressler, a. Muzzin, B. Knoblich, E. Peterson, and M. Tomlanovich, "Early goal-directed therapy in the treatment of severe sepsis and septic shock.," *The New England journal of medicine*, vol. 345, pp. 1368–77, November 2001.

[3] J. Lee and R. Mark, "An investigation of patterns in hemodynamic data indicative of impending hypotension in intensive care," *Computers in Cardiology*, vol. 37, no. 1, pp. 81–84, 2010.

[4] X. Chen, D. Xu, G. Zhang, and R. Mukkamala, "Forecasting Acute Hypotensive Episodes in Intensive Care Patients Based on a Peripheral Arterial Blood Pressure Waveform," *Computers in Cardiology 2009*, vol. 36, pp. 545–548, 2009.

[5] J. H. Henriques and T. R. Rocha, "Prediction of Acute Hypotensive Episodes Using Neural Network Multi-models," *Computers in Cardiology 2009*, vol. 36, pp. 549–552, 2009.

[6] M. Mneimneh and R. Povinelli, "A Rule-Based Approach for the Prediction of Acute Hypotensive Episodes," *Computers in Cardiology 2009*, vol. 36, pp. 557–560, 2009.

[7] O. Alexis, "Providing best practice in manual blood pressure measurement," *British Journal of Nursing*, vol. 18, pp. 410–415, April 2009.

[8] ADAM Medical Encyclopedia. [Internet], "Hypotension; [updated 2009 feb 22; cited 2010 jun 11]," Atlanta (GA): A.D.A.M., Inc. ©2010.

[9] A. Chobanian, G. Bakris, H. Black, W. Cushman, L. Green, J. Izzo, D. Jones, B. Materson, S. Oparil, J. Wright, and E. Roccella, "Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure," *Hypertension*, vol. 42, no. 6, pp. 1206–52, 2003.

[10] A. Jones, V. Tayal, and D. e. a. Sullivan, "Randomized controlled trial of immediate versus delayed goal-directed ultrasound to identify the cause of nontraumatic hypotension in emergency department patients," *Crit Care Med*, vol. 32, no. 1, pp. 1703–1708, 2004.

[11] NHS Medical Encyclopedia. [Internet], "Hypotension; [updated 2009 oct 2; cited 2010 jun 11]," British National Health Services ©2010.

[12] S. Goldhaber, L. Visani, and M. De Rosa, "Acute pulmonary embolism: clinical outcomes in the International Cooperative Pulmonary Embolism Registry (ICOPER)," *Lancet*, vol. 353, no. 9162, pp. 1386–1389, 1999.

[13] K. Lee, L. Woodlief, E. Topol, W. Weaver, A. Betriu, J. Col, M. Simoons, P. Aylward, F. Van de Werf, and C. RM, "Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction: results from an international trial of 41,021 patients: GUSTO-I Investigators," *Circulation*, vol. 91, no. 6, pp. 1659–1668, 1995.

[14] J. H. Kilgannon, B. W. Roberts, L. R. Reihl, M. E. Chansky, A. E. Jones, R. P. Dellinger, J. E. Parrillo, and S. Trzeciak, "Early arterial hypotension is common in

the post-cardiac arrest syndrome and associated with increased in-hospital mortality," *Resuscitation*, vol. 79, no. 3, pp. 410 – 416, 2008.

[15] G. Martin, D. Mannino, S. Eaton, and M. Moss, "The Epidemiology of Sepsis in the United States from 1979 through 2000," *The New England Journal of Medicine*, vol. 348, no. 16, pp. 1546–1554, 2003.

[16] ADAM Medical Encyclopedia. [Internet], "Sepsis; [updated 2009 sep 28; cited 2010 jun 11]," Atlanta (GA): A.D.A.M., Inc. ©2010.

[17] A. G. Tsiotou, G. H. Sakorafas, G. Anagnostopoulos, and J. Bramis, "Septic shock; current pathogenetic concepts from a clinical perspective.," *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, vol. 11, pp. RA76–85, March 2005.

[18] S. McGee, W. B. A. III, and D. L. Simel, "Is this patient hypovolemic?," *Journal of the American Medical Assoc.*, vol. 281, no. 11, pp. 1022–1029, 1999.

[19] Mayo Clinic. [Internet], "Bradycardia; [updated 2009 may 28; cited 2010 oct 24]," Mayo Foundation for Medical Education and Research (MFMER): Mayo Clinic ©1998-2010.

[20] K. Hassanpour, S. HotzBoendermaker, P. Dokladal, E. M. S. for Human Spinal Cord Injury Study group, and A. Curt, "Low depressive symptoms in acute spinal cord injury compared to other neurological disorders," *Journal of Neurology*, vol. Epub ahead of print: (18 November 2011), pp. 1–9, 2011.

[21] J. E. Peacock, D. A. Herrington, J. C. Wade, H. M. Lazarus, M. D. Reed, J. W. Sinclair, D. C. Haverstock, S. F. Kowalsky, D. D. Hurd, D. A. Cushing, C. P. Harman, and G. R. Donowitz, "Ciprofloxacin plus piperacillin compared with tobramycin plus piperacillin as empirical therapy in febrile neutropenic patients: A randomized, double-blind trial," *Annals of Internal Medicine*, vol. 137, no. 2, pp. 77–87, 2002.

[22] R. Pumphrey, "Lessons for management of anaphylaxis from a study of fatal reactions," *Clinical and Experimental Allergy*, vol. 30, pp. 1144–1150, 2000.

[23] J. Bassale, "Hypotension Prediction Arterial Blood Pressure Variability," *Portland State University. BSP Technical Report 01-002*, June 2001.

[24] F. Prescott, "Clinical evaluation of the pressor activity of methedrine, neosynephrine, paredrine, and pholedrine.," *Br Heart J*, vol. 6, pp. 214–220, Oct 1944.

[25] L. Tarassenko, A. Hann, A. Patterson, E. Braithwaite, K. Davidson, V. Barber, and D. Young, "Biosign: multi-parameter monitoring for early warning of patient deterioration," in *3rd IEE International Seminar on Medical Applications of Signal Processing*, pp. 71–76, 2005.

[26] C. Crespo, J. Mcnames, M. Aboy, J. Bassale, M. Ellenby, S. Lai, and B. Goldstein, "Precursors In The Arterial Blood Pressure Signal To Episodes Of Acute Hypotension In Sepsis," *Proceedings of the 16th International EURASIP Conference Biosignal*, vol. 16, pp. 206–208, 2002.

[27] L. Lehman, M. Saeed, G. Moody, and R. Mark, "Similarity-based searching in multiparameter time series databases," *Computers in Cardiology 2008*, vol. 35, pp. 653–656, September 2008.

[28] M. Saeed, M. Villarroel, A. T. Reisner, G. D. Clifford, L. Lehman, G. B. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, "Multiparameter Intelligent Monitoring in Intensive Care II: A public-access intensive care unit database," *Critical Care Medicine*, vol. 39, pp. 952–960, May 2011.

[29] A. Stell, R. Sinnott, J. Jiang, R. Donald, I. Chambers, G. Citerio, P. Enblad, B. Gregson, T. Howells, K. Kiening, P. Nilsson, A. Ragauskas, J. Sahuquillo, and I. Piper, "Federating distributed clinical data for the prediction of adverse hypotensive events," *Philosophical Transactions of the Royal Society A*, vol. 367, no. 1898, pp. 2679–2690, 2009.

[30] A. Stell, R. Sinnott, and J. Jiang, "A clinical grid infrastructure supporting adverse hypotensive event prediction," in *Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, pp. 508–513, 2009.

[31] P. Langley, S. King, D. Zheng, E. Bowers, K. Wang, J. Allen, and A. Murray, "Predicting Acute Hypotensive Episodes from Mean Arterial Pressure," *Computers in Cardiology 2009*, vol. 36, pp. 553–556, 2009.

[32] P. Fournier and J. Roy, "Acute Hypotension Episode Prediction Using Information Divergence for Feature Selection, and Non-Parametric Methods for Classification," *Computers in Cardiology 2009*, vol. 36, pp. 625–628, 2009.

[33] D. Hayn, B. Jammerbund, A. Kollmann, and G. Schreier, "A Biosignal Analysis System Applied for Developing an Algorithm Predicting Critical Situations of High Risk Cardiac Patients by Hemodynamic Monitoring," *Computers in Cardiology*, vol. 36, pp. 629–632, 2009.

[34] S. W. Roberts, "Control chart tests based on geometric moving averages," *Technometrics*, vol. 1, no. 3, pp. 239–250, 1959.

[35] K. Jin and N. Stockbridge, "Smoothing and Discriminating MAP Data," *Computers in Cardiology 2009*, vol. 36, pp. 633–636, 2009.

[36] T. Ho and X. Chen, "Utilizing Histogram to Identify Patients Using Pressors for Acute Hypotension," *Computers in Cardiology 2009*, vol. 36, pp. 797–800, 2009.

[37] D. Specht, "A general regression neural network.," *IEEE Transactions on Neural Networks*, vol. 2, no. 6, pp. 568–576, 1991.

[38] M. J. D. Powell, *Radial basis functions for multivariable interpolation: a review*. New York, NY, USA: Clarendon Press, 1987.

[39] S. Ghosh and D. L. Reilly, "Credit card fraud detection with a neural-network," in *Proceedings Twenty-Seventh Hawaii International System Sciences Vol. III: Infor-*

mation Systems: Decision Support and Knowledge-Based Systems Conference, vol. 3, pp. 621–630, 1994.

[40] R. Plamondon and S. N. Srihari, "Online and off-line handwriting recognition: a comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63–84, 2000.

[41] H. B. Burke, D. B. Rosen, and P. H. Goodman, "Comparing artificial neural networks to other statistical methods for medical outcome prediction," in *Proceedings of IEEE International Neural Networks IEEE World Congress Computational Intelligence Conference*, vol. 4, pp. 2213–2216, 1994.

[42] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, pp. 1065–1076, 1962.

[43] D. F. Specht, "Generation of polynomial discriminant functions for pattern recognition," *IEEE Transactions on Electronic Computing*, vol. EC-16, pp. 308–319, 1967.

[44] M. Saeed, *Temporal Pattern Recognition in Multiparameter ICU Data*. PhD thesis, Massachusetts Institute of Technology, 2007.

[45] J. Shao and D. Tu, *The Jackknife and Bootstrap*. Springer-Verlag, Inc., 1995.

[46] J. Xie and Z. Qiu, "Bootstrap technique for ROC analysis: A stable evaluation of fisher classifier performance," *Journal of Electronics (China)*, vol. 24, no. 4, pp. 523–527, 2007.

[47] L. Tarassenko, A. Hann, and D. Young, "Integrated monitoring and analysis for early warning of patient deterioration," *British Journal of Anaesthesia*, vol. 97, no. 1, pp. 64–68, 2006.

[48] L. Tarassenko, A. Hann, A. Patterson, E. Braithwaite, V. B. K. Davidson, and D. Young, "Multi-parameter monitoring for early warning of patient deterioration,"

Proceedings of the 3rd IEE International Seminar on Medical Applications of Signal Processing, London, pp. 71–76, 2005.

[49] A. Hann, *Multi-parameter Monitoring for Early Warning of Patient Deterioration*. PhD thesis, University of Oxford, 2008.

[50] M. D. Buist, E. Jarmolowski, and e. a. P. R. Burton, "Recognising clinical instability in hospital patients before cardiac-arrest or unplanned admission to intensive care. a pilot study in a tertiary-care hospital," *Medical Journal of Australia*, vol. 171, pp. 22–25, 1999.

[51] J. Kause, G. Smith, D. Prytherch, and et al., "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*, vol. 62, pp. 275–282, 2004.

[52] F. L. Sax and M. Charlson, "Medical patients at high risk for catastrophic deterioration," *Critical Care Medicine*, vol. 15, pp. 510–515, 1987.

[53] K. Hillman, P. Bristow, T. Chey, K. Daffurn, T. Jacques, S. Norman, G. Bishop, and G. Simmons, "Antecedents to hospital deaths," *Internal Medicine Journal*, vol. 31, pp. 343–348, 2001.

[54] C. M. Bishop, "Novelty detection and neural network validation," *Vision, Image and Signal Processing, IEE Proceedings*, vol. 141, pp. 217–222, 1994.

[55] F. Rosenblatt, *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Spartan Books, 1962.

[56] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning Internal Representations by Error Propagation*, vol. 1:Foundations. Cambridge, MA: Bradford Books/MIT Press, 1986.

[57] A. Jain and B. Chandrasekaran, *Handbook of Statistics: Dimensionality and sample size considerations in pattern recognition practice.*, vol. 2. North-Holland Publishing Company, 1982.

[58] D. Shavdia, *Septic Shock: Providing Early Warnings Through Multivariate Logistic Regression Models.* PhD thesis, Massachusetts Institute of Technology, 2007.

[59] G. Liljestrand and E. Zander, "Vergleichende bestimmung des minutenvolumens des herzens beim menschen mittels der stickoxydulmethode und durch blutdruckmessung," *Z. Ges. Experimental Medicine*, vol. 59, pp. 105–122, 1928.

[60] C. W. Hug, G. D. Clifford, and A. T. Reisner, "Clinician blood pressure documentation of stable intensive care patients: An intelligent archiving agent has a higher association with future hypotension," *Critical Care Medicine*, vol. 39, no. 5, pp. 1006–1014, 2011.

[61] W. Zong, G. B. Moody, and R. G. Mark, "Reduction of false arterial blood pressure alarms using signal quality assessment and relationships between the electrocardiogram and arterial blood pressure," *Medical and Biological Engineering and Computing*, vol. 42, pp. 698–706, 2004.

[62] C. W. Hug and G. D. Clifford, "An analysis of the errors in recorded heart rate and blood pressure in the ICU using a complex set of signal quality metrics," *Computers in Cardiology*, vol. 34, pp. 641–645, 2007.

[63] T. Chen, G. D. Clifford, and R. G. Mark, "The effect of signal quality on six cardiac output estimators," *Computers in Cardiology*, vol. 36, pp. 197–200, 2009.

Appendix A

Glossary of Terms

ABP - Arterial Blood Pressure

AHE - Acute Hypotensive Episode

AUC - Area under the ROC curve, a value of 0.5 corresponds to chance

Arrhythmia - Abnormal cardiac electrical activity

Bradycardia - Low resting heart rate

Cardiac Arrest - Inability of the heart to contract

Cardiogenic Shock - Failure of the heart to pump

CO - Cardiac output

CVP - Central venous pressure

ECG - Electrocardiogram

Haemorrhage - Excessive bleeding

HR - Heart rate

ICU - Intensive care unit

MLP - Multi-layer Perceptron

Myocardial Infarction - Heart attack

PAP - Pulmonary artery pressure

Pressor - Medication used to regulate blood pressure

Pulmonary Embolism - Blockage of the lung or one of its branches

RESP - Respiration rate

ROC - Receiver Operating Characteristic; an ROC curve plots sensitivity versus 1-specificity as a threshold varies

SLP - Single-layer Perceptron

Vasodilation - Decreased vascular resistance caused through the relaxation of the smooth muscle surrounding blood vessels

Appendix B

Pressor Description

Dopamine (Intropin) At low doses, dopamine acts predominately on dopamine-1 receptors in the renal mesenteric, cerebral, and coronary beds, resulting in selective vasodilation. At moderate doses, dopamine also stimulates Beta-1 receptors and increases CO, predominately by increasing stroke volume with variable effects on HR. Can result in dose-limiting dysrhythmias. At higher doses, dopamine stimulates alpha receptors and produces vasoconstriction with an increased systemic vascular resistance. MIMIC II refers to this as Dopamine, and has two categories for the medication (drip and non-drip).

Epinephrine (Adrenalin) Potent beta-1 receptor activity and moderate beta-2 and alpha-1 receptor effects. MIMIC II refers to this as Epinephrine, and has three categories for the medication (drip, non-drip, and -k).

Isoproterenol Stimulates both beta1 and beta2 adrenergic receptors (no alpha receptor capabilities). MIMIC II does not contain a reference to this medication.

Norepinephrine (Levophed) Potent vasoconstriction as well as a less pronounced increase in cardiac output. MIMIC II refers to this as Levophed, and has two categories for the medication (non-drip, and -k).

Phenylephrine (Neosynephrine) Vasoconstriction with minimal cardiac inotropy or

chronotropy. MIMIC II refers to this as Neosynephrine, and has two categories for the medication (non-drip, and -k).

Vasopressin Usually used in the setting of DI or esophageal variceal bleeding.

Appendix C

Hypotensive ICD9 Codes

458.0: Orthostatic hypotension

A fall in blood pressure associated with dizziness, syncope and blurred vision occurring upon standing or when standing motionless in a fixed position.

458.1: Chronic hypotension

458.2: Iatrogenic hypotension

458.21: Hypotension of hemodialysis

458.29: Other iatrogenic hypotension

458.8: Other specified hypotension

458.9: Hypotension unspecified

Any blood pressure that is below the normal expected for an individual in a given environment.

Appendix D

Error Backpropagation Algorithm

Given a network with $L+1$ layers, N inputs and M output nodes, the error backpropagation algorithm first performs an inital forward pass where the input vector $x_1 \dots x_N$ is transformed into the output vector $y_L = y_1, \dots y_i, \dots y_M$ by iterating through each layer l as follows:

$$y_i^l = f(v_i^l) = f(\sum_{j=1}^{n_{l-1}} x_j^{l-1} * w_{i,j}^l + b_i^l) \tag{D.1}$$

Note that the final layer value x_i^L will be the output values y_L.

The total error between the output y_i and the targets t_i is calculated using:

$$\delta_i^L = f'(v_i^L) * (t_i - y_i) \tag{D.2}$$

The output error is propagated backwards (from $l = L \dots 1$) through the entire MLP, by evaluating:

$$\delta_j^{l-1} = f'(v_i^{l-1}) * \sum_{i=1}^{n_l} \delta_i^l * w_{i,j}^l \tag{D.3}$$

Once the proportional error is calculated for each node in all previous layers, the weights and biases are updated:

$$w_{i,j}^l = w_{i,j}^l + \alpha * \delta_i^l * x_j^{l-1} \tag{D.4}$$

$$b_i^l = b_i^l + \alpha * \delta_i^l \tag{D.5}$$

The squared error between the target and output is defined as:

$$E(w, b) = \frac{1}{2} \sum_{i=1}^{M} (t_i - y_i)^2 \tag{D.6}$$

The error backpropagation algorithm is a gradient descent algorithm which minimises this cost function. The weight update equations are:

$$w_{i,j}^l = w_{i,j}^l - \alpha * \frac{\partial E}{\partial * w_{i,j}^l} \tag{D.7}$$

$$b_i^l = b_i^l - \alpha * \frac{\partial E}{\partial * b_i^l} \tag{D.8}$$

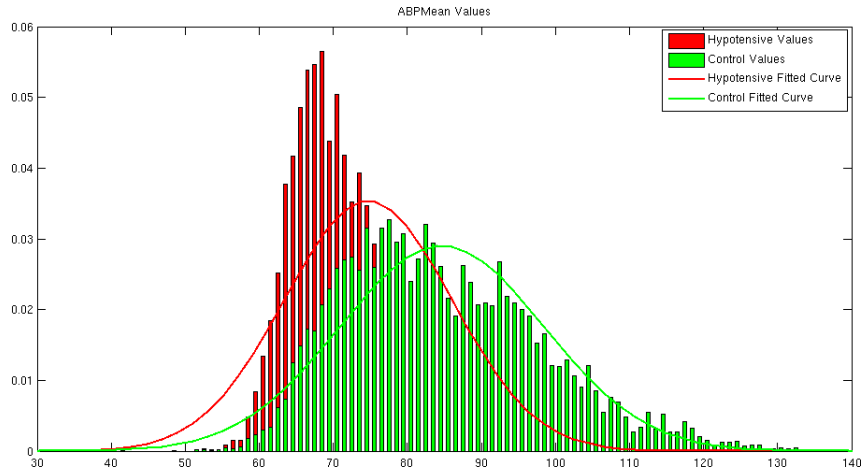
Backpropagation training can continue until the magnitude of the gradient is below a certain threshold or an error rate has been reached.

Appendix E

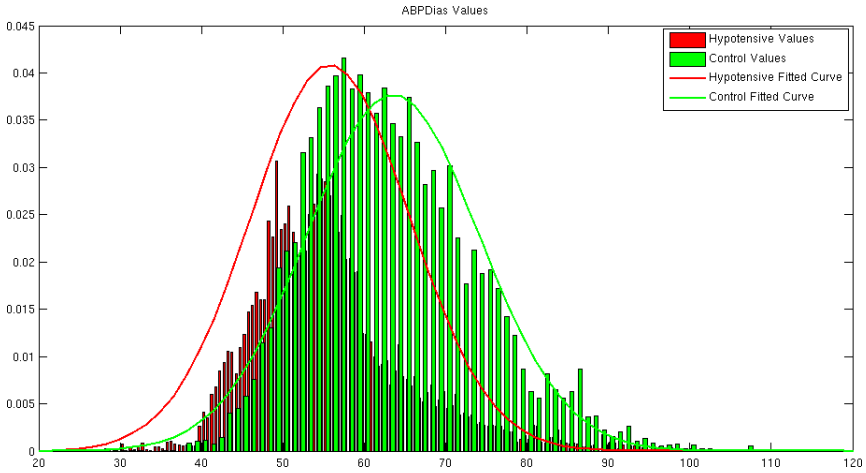
Variable Distributions

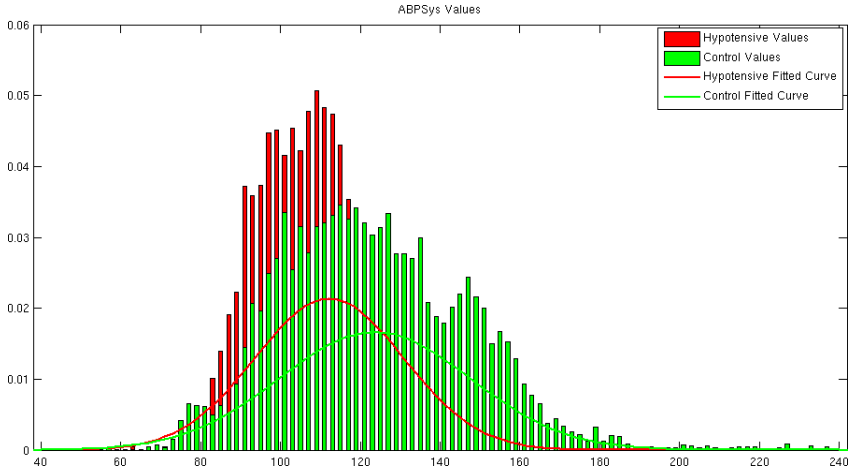
Figure E.1: *Histograms of Variable Distributions*

(a) *ABPMean Distribution*

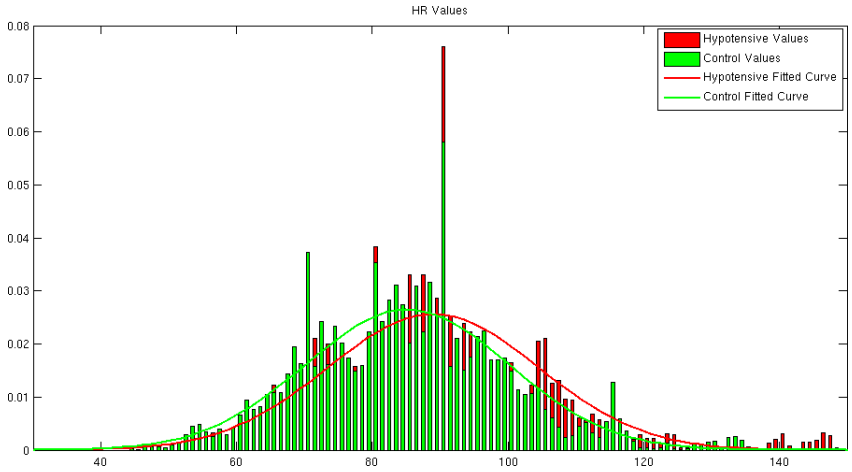


(b) *ABPDias Distribution*

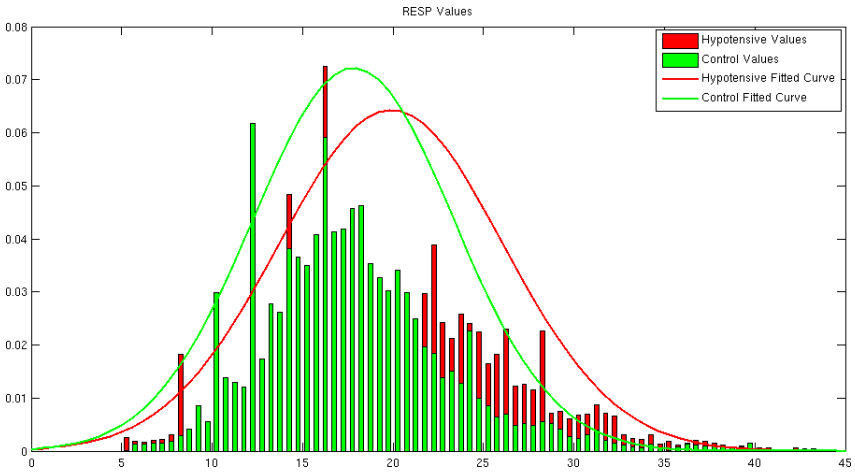




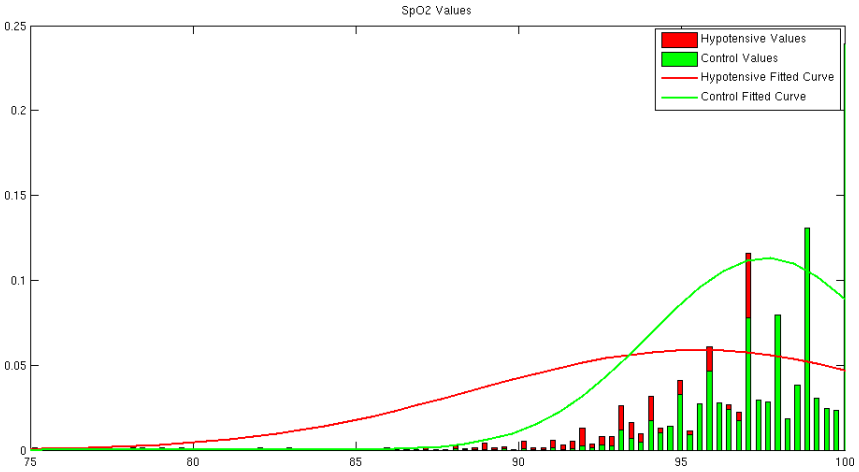
(c) *ABPSys Distribution*



(d) *HR Distribution*



(e) *RESP Distribution*



(f) *SpO_2 Distribution*