College of Science and Engineering

Development of Machine Learning Schemes for use in Non-Invasive and Continuous Patient Health Monitoring

Thesis submitted by

Stephanie Baker, B.Eng (Hons)

in January 2021

for the degree of

Doctor of Philosophy

Acknowledgments

I am deeply grateful to the following people, and many others, for their support. This thesis would not have been possible without you.

Thank you to my supervisor Prof. Wei Xiang for your guidance through this project. Working with you has improved my thesis and skills as a researcher.

To my secondary supervisor, Prof. Ian Atkinson - thank you for your infectious enthusiasm and support. Thank you for constantly challenging me to take new opportunities and to grow as a researcher.

I am grateful to my wonderful colleagues, including but not limited to Dr. Kenny Leong, Prof. Mohan Jacob, Dr. Maria Papparlardo, and Melissa Norton. Your support has been invaluable throughout my PhD.

Thank you to Dr. Joe Moxon and A/Prof. Anthony Leicht, for the insight that you gave me into medical fields and concepts over many cups of coffee.

Thank you to my friends for always being there for me - even when I disappeared for months on end into this research.

Special thanks to my partner Laurance Papale, for your unending love and support throughout all of my studies. I couldn't have done this without you.

To my family, including the family that I have gained through Laurance - I am eternally grateful for your enthusiasm, encouragement, and belief in me. I would particularly like to thank my father Rodney Baker. I wouldn't be the person I am today without your guidance and wisdom.

Finally, to my late mother, Sandra Baker. You were my first teacher, my biggest fan, and an incredible mum. This is for you.

Statement of the Contribution of Others

I acknowledge the following contributions to this work with gratitude.

Financial assistance was provided by the Australian Postgraduate Award, which included living assistance and tuition fee support. Additional financial support was offered through the annual funding and additional grants from the College of Science and Engineering and Graduate Research School, which facilitated the publication of the content of this thesis.

Editorial assistance for this thesis was provided by my supervisors, Prof. Wei Xiang and Prof. Ian Atkinson. Chapters 2-6 are each based on papers published or submitted for publication, and editorial assistance was again provided by my supervisors, who were the only co-authors.

Contributions to co-authored publications were in the form of editorial assistance and technical feedback. I am first author of all publications related to this thesis, wherein I conceived all experiments, performed simulations, analysed results, developed figures and graphs, and wrote all text.

Data used for publications related to Chapters 3-6 were acquired from the open-access Medical Information Mart for Intensive Care III (MIMIC-III), an invaluable resource for researchers working in the intersection of data science and healthcare.

Abstract

Healthcare is an essential part of life, but modern healthcare systems are faced with many challenges. Costs and resource demand in healthcare facilities have been steadily rising, largely due to the ever-increasing global population. Our population is also older than ever, which in turn leads to increased prevalence of chronic health conditions and higher pressure on healthcare systems.

Many chronically ill patients need to spend large amounts of time in hospital, with a variety of manual or invasive monitoring techniques used to observe their health. This places high strain on healthcare workers such as nurses and doctors, and also places significant emotional and financial strain on patients and their families.

This thesis explores machine learning (ML) solutions that stand to greatly improve the standard of healthcare worldwide, beginning with methods of non-invasive vital sign measurement that could be used at home or in healthcare environments. Focus then turns towards better utilisation hospital resources in the intensive care unit, by quantifying severity of illness through mortality risk prediction in various windows and thus empowering healthcare workers to effectively triage and make treatment decisions..

Non-invasive and continuous methods for measuring systolic and diastolic BP using hybridized neural network from easily-obtained HR signals are explored, with the aim of eliminating the need for invasive intra-arterial measuring or the use of uncomfortable sphygmomanometers. Using raw HR signals obtained via photoplethysmogram and electrocardiogram sensor as inputs, a hybrid convolutional and long short-term memory (CNN-LSTM) neural network (NN) is shown to perform strongly for the task, and meets the high industry standards for blood pressure monitoring devices.

A solution with fewer input features is also explored, with the aim of improving computational efficiency for low-powered devices. This scheme uses features of the ECG and PPG waveforms rather than manually calculated features and a smaller CNN-LSTM network. It is shown to perform comparably to the scheme utilising raw waveforms as inputs despite the significantly reduced number of features, and again meets industry standards.

The RR is commonly measured through manual counting of breaths, a tedious and time-consuming process. As such, the automatic and continuous measurement of this parameter is investigated. Multiple respiratory variations are extracted from HR signals, with each used to determine an estimate for RR. A respiratory quality index (RQI) is also developed to determine the quality of the respiratory variation signals. RR estimates and corresponding RQIs are used as inputs to a bidirectional long short-term memory (BiLSTM) network. This process was repeated for three different segment lengths - 20, 30, and 60 seconds. In all three cases, the BiLSTM network performed significantly better when the novel RQIs were included as features.

The second section of the thesis explores mortality risk prediction, beginning with adult patients. In this work, it is shown that adult mortality risk in intensive care units within short- and long-term windows can be reliably assessed using hybrid CNN-LSTM networks and vital sign data. The easy-to-acquire features also ensures that mortality risk can be continuously updated without manual intervention, allowing healthcare staff to observe trends in how the patient is responding to treatment.

Lastly, this work investigates the prediction of neonatal mortality risk. Premature babies are a high-risk group of patients, and quantifying their risk levels can assist in treatment decisions. In this work, we show high prediction performance for neonatal mortality using only gestational age, birth weight, RR and HR data as input features to a CNN-LSTM network.

List of Publications

The following publications were produced during the period of candidature:

[1] **S. Baker**, W. Xiang, and I. Atkinson, "Internet of Things for Smart Healthcare: Technologies, Challenges, and Opportunities", *IEEE Access*, vol. 5, pp. 26521-26544, November 2017.

[2] **S. Baker**, W. Xiang, and I. Atkinson, "Continuous and Automatic Mortality Risk Prediction using Vital Signs in the Intensive Care Unit: A Hybrid Neural Network Approach," *Scientific Reports*, vol. 10, pp. 21282, December 2020.

[3] **S. Baker**, W. Xiang, and I. Atkinson, "Determining respiratory rate from photoplethysmogram and electrocardiogram signals using respiratory quality indices and neural networks," *PLoS One*, vol. 16, no. 4, p. e0249843, February 2021.

[4] **S. Baker**, W. Xiang, and I. Atkinson, "A Hybrid Neural Network for Continuous and Non-Invasive Estimation of Blood Pressure from Raw Electrocardiogram and Photoplethysmogram Waveforms," *Computer Methods and Programs Biomedicine*, p. 106191, 2021.

[5] **S. Baker**, W. Xiang, and I. Atkinson, "Non-invasive and Continuous Neonatal Mortality Risk Assessment using Respiratory Rate and Heart Rate," accepted for publication in *Computers in Biology and Medicine*

Contents

List of Tables

List of Figures

List of Abbreviations

AMI - Association for the Advancement of Medical Instrumentation

ABP - Arterial Blood Pressure

AUPRC - Area Under the Precision-Recall Curve

AUROC - Area Under the Receiver-Operator Curve

BHS - British Hypertension Society

BiLSTM - Bidirectional Long Short-Term Memory

BP - Blood Pressure

bpm - Beats Per Minute

BrPM - Breaths Per Minute

CNN - Convolutional Neural Network

DBP - Diastolic Blood Pressure

DNN - Deep Neural Network

ECG - Echocardiogram

FFNN - Feed-Forward Neural Network

HR - Heart Rate

ICU - Intensive Care Unit

LOA - Limits of Agreement

LSTM - Long Short-Term Memory

MAE - Mean Absolute Error

MAP - Mean Arterial Pressure

MD - Mean Difference

MIMIC - Medical Information Mart for Intensive Care

PPG - Photoplethysmogram

PRC - Precision-Recall Curve

RMSE - Root Mean Squared Error

RNN - Recurrent Neural Network

ROC - Receiver-Operator Curve

RR - Respiratory Rate

SBP - Systolic Blood Pressure

SD - Standard Deviation

SQI - Signal Quality Index

TNR - True Negative Rate

TPR - True Positive Rate

UniLSTM - Unidirectional Long Short-Term Memory

Chapter 1

Introduction

1.1 Background & Motivation

Healthcare is an essential part of life. Worryingly, modern healthcare systems are under immense strain due to a growing and ageing population and a related rise in chronic illness [6]. Resources from medical professionals to hospital beds are in high demand across the healthcare sector [7].

Critical care is one of many healthcare services experiencing strain as a result of an ageing population, with over 60% of admissions to Victorian Intensive Care Units (ICUs) in 2010-11 occurring in patients over 60, despite this group representing less than 20% of the population during this period [8].

ICUs treat the most critically ill of patients, and as such have the highest mortality rate of all hospital units [9], ranging between 13.0%-14.4% in tertiary hospitals in Victoria from 2001-2011 [8]. Patient outcomes are improved with a high staff-to-patient ratio in ICUs [10], however providing this level of care comes at a high cost. In 2013/14, intensive care in Australia cost $4,375 per patient bed-day, resulting in an annual expenditure of $2,119 million dollars [11].

On the other end of the age range, Special Care Nurseries (SCNs) and Neonatal Intensive Care Units (NICUs) provide high levels of care to critically ill newborns. The rate of infant admission to SCNs and NICUs has been steadily increasing, with 18.3% of all babies born in Australia admitted to one of these critical care environments in 2018 [12]. As of 2018, the mortality rate for neonatal patients was 2.2% [12].

While critical care for all age groups has undergone significant innovation and

improvement in recent years, many issues remain. Some monitoring techniques remain largely manual, with manual counting of breaths over a one-minute period for respiratory rate measurement remaining the accepted method [13]. Such manual measurement has been shown to be limited by factors such as patient awareness, time constraints, external interruptions, and patient agitation [14–16]. The time cost of patient monitoring is high, with nurses spending 7.2% of their time performing patient assessment [17]. Other monitoring techniques in critical care units are often invasive, such as gold-standard intra-arterial blood pressure monitoring [18], which in turn leads to increased infection risk.

Treatments can also be highly invasive, with mechanical ventilation received by 42% of ICU patients [8]. These invasive methods are only used where absolutely necessary, however it can be challenging to make decisions about when to initiate, alter, or withdraw such treatments. Recent studies have shown that mortality risk assessment can aid in making difficult treatment decisions [19].

Decision making becomes increasingly important and no less challenging in times of crisis. The ongoing COVID-19 pandemic has highlighted how rapidly modern healthcare systems can be overwhelmed by widespread disease. Impacts of the pandemic on ICUs and the wider healthcare system have been reported across the world, in countries including Italy [20], France [21], Brazil [22], the United Kingdom [23], the United States [24], and China [25]. Such impacts have included overcrowding of hospitals [20], lack of resources [20, 23, 25], and increased risk of infection among essential healthcare workers [23]. In Australia, ICUs have identified their ability to add more beds, however the effectiveness of this is restricted by a low capability to increase mechanical ventilation units and staffing levels [26]. The widespread and significant impacts have lead to challenging decision making, with several reports examining ethical and fair rationing of increasingly limited healthcare resources [27–29].

Mortality risk prediction could ease the burden of such decisions, allowing healthcare workers to assess which patients require the most urgent access to limited resources such as invasive ventilation. Unfortunately, existing methods for quantifying mortality risk are limited in that they depend heavily upon laboratory test results, are calculated once at admission and not updated throughout

the stay, and decline in performance over time [30, 31].

The burden of COVID-19 on healthcare systems has also had significant impacts on out-of-hospital care. Following the 2020 outbreak of COVID-19 in Lombardy, emergency services took longer on average to arrive at medical incidents than in 2019. Additionally, occurrences of out-of-hospital death where resuscitation was attempted increased by 14.9 percentage points [32]. Out-of-hospital monitoring with non-invasive devices would improve at-home care and allow for rapid detection of a medical emergency, and thus improve response times. Such at-home monitoring would have ongoing benefits in routine healthcare beyond the pandemic, particularly for elderly, chronically ill, or otherwise at-risk patients.

Improvements in monitoring would offer significant advantages to the healthcare system. The development of automatic, continuous, and non-invasive methods for measuring vital signs such as respiratory rate and blood pressure would greatly enhance monitoring both at-home and in the hospital. Moving to non-invasive methods of measurement would also reduce the high infection risks in intensive care units, particularly for the at-risk populations of neonatal and elderly patients. For patients admitted to the hospital, and more seriously the ICU, mortality risk monitoring is an essential tool for improving patient outcomes by enabling informed decision making around treatments.

In recent years, machine learning has emerged as a technique for improving healthcare in areas including vital sign monitoring [33–45], detection of clinical events or deterioration [46–54], and mortality prediction [36, 55–63]. However, much of the research to date is not yet suitable for clinical implementation.

Motivated by the challenges in healthcare monitoring both in and out of ICU environments, there are two main objectives for this thesis. Firstly, this thesis develops machine learning methods for continuous and non-invasive measurement of respiratory rate and blood pressure, using data from sensors that are readily available and highly wearable. Secondly, this thesis aims to improve patient outcomes in the ICU and NICU by developing mortality risk assessment tools that utilize machine learning to predict mortality outcomes from readily obtained vital sign measurements and basic demographics. Overall, this thesis presents

solutions for improved measurement and monitoring of patient health in a range of settings.

1.2 Research Problems

This thesis focuses on improving healthcare through the use of machine learning techniques in several areas. The research problems investigated include the development of novel methods for monitoring blood pressure and respiratory rate continuously and non-invasively, as well as the development of tools for quantifying severity of illness using vital sign information in intensive care units. These research problems are elaborated upon as follows.

1.2.1 Blood Pressure Monitoring

Blood pressure is an important health parameter that can provide much information about a patient's cardiovascular health. The current gold-standard for blood pressure monitoring is invasive intra-arterial monitoring with a pressure transducer inserted into a suitable artery. While this method is capable of continuous monitoring, it is invasive and can only be utilised in clinical settings. Other methods, such as the use of cuff-based sphygmomanometers, cannot provide continuous measurements of the parameter and remain uncomfortable.

Continuous and non-invasive blood pressure monitoring would be an ideal solution to this problem, and much research has looked to achieve this goal. However, the research problem remains relatively new and existing schemes have not yet succeeded in meeting the standards required for clinical implementation. To address this problem, this thesis proposes two schemes - one focused on maximum performance and the other on finding a balance between performance and computational efficiency - that utilise hybridized neural networks to determine blood pressure from electrocardiogram and photoplethysmogram waveforms.

1.2.2 Respiratory Rate Monitoring

Respiratory rate is a key vital sign, however it has been historically under-recorded. The primary reason for this is that the accepted method for mea-

surement is manual counting of the breath over a one-minute period, which is time-consuming for healthcare professionals. Other methods are obstructive and uncomfortable for patients, requiring placement of devices over the mouth or nose to accurately measure the parameter.

The development of a method for measurement of respiratory rate in a non-invasive and continuous manner has long evaded researchers, with devices from microphones to stretch sensors trialled for the purpose. Unfortunately, many sensors are still obstructive and not suitable for long-term wear. Additionally, sufficient performance has not been achieved in the literature to date. This thesis addresses the research problem of accurately monitoring respiratory rate through non-invasive techniques through the development of a scheme that extracts respiratory modulations of the electrocardiogram and photoplethysmogram waveforms and assesses their quality, using quality indices and candidate respiratory rate values as inputs to a neural network for final prediction of respiratory rate.

1.2.3 Mortality Risk Assessment

Decision making regarding resource allocation in intensive care units is a challenging and yet essential task. This has become increasingly apparent throughout the COVID-19 crisis, which has seen ICUs filled beyond their capacity and thus left healthcare professionals to decide which patients should be allocated the limited numbers of life-saving equipment such as mechanical ventilators. This is an emotionally challenging task and places high strain on healthcare workers.

Making decisions regarding treatment paths is also challenging in critical care environments, and it is often difficult to determine whether a patient is improving or deteriorating following certain treatment path. As such, it can be often be hard to determine when to start or withdraw treatments.

Mortality risk assessment can ease decision making in both areas, however existing schemes are limited by their dependencies on knowledge of medical histories and extensive laboratory results. Additionally, the performance of existing schemes has been shown to deteriorate with time. Finally, the complex nature of existing schemes means that mortality risk is typically only calculated once at the commencement of the stay, and not updated throughout.

Mortality risk assessment has attracted much interest in the literature, particularly from researchers aiming to use machine learning to improve predictions. While machine learning improves the calibration problem (as machine learning models can continuously learn as they work), schemes in the literature remain dependent on complex variables and are not continuously updated throughout the stay.

This thesis aims to address these problems by developing a hybrid neural network scheme for the measurement of mortality risk within 3, 7, and 14 days. One model is proposed for adult patients, and another for neonatal patients. In both models, only vital signs, age characteristics, and sex are used as features. The simplicity of these variables ensures that the scheme can be continuously and automatically updated throughout a patient's stay to allow ongoing assessment of their condition.

1.2.4 Summary

To summarise, this thesis addresses several key research problems: the continuous and non-invasive measurement of both blood pressure and respiratory rate, and the development of mortality risk assessment schemes for both adult and infant patient groups. The development of techniques for improved vital sign monitoring not only improves at-home healthcare monitoring, but also greatly enhances the ability of mortality risk prediction schemes to provide accurate results. Addressing these major research problems has lead to the original contributions outlined in the following section.

1.3 Original Contributions

To address the research problems outlined in Section 1.2., the following original contributions to the literature are presented in this thesis:

1. In chapter 3, a robust and powerful hybrid neural network for non-invasive and continuous blood pressure measurement that uses raw waveforms from wearable electrocardiogram and photoplethysmogram sensors as inputs is

developed, and is shown to meet industry standards for blood pressure devices.

2. In chapter 3, a low-power alternative neural network scheme for blood pressure measurement using 12 features that describe the shape of the electrocardiogram and photoplethysmogram waveforms is also presented, showing little compromise on predictive performance compared to the raw waveform scheme and a significant decrease in computational time.

3. In chapter 4, a machine learning scheme for continuous and automatic respiratory rate measurement is designed, based on deriving respiratory signals from modulations to photoplethysmogram and electrocardiogram signals caused by respiration and conducting respiratory signal quality assessment.

4. In chapter 5, a hybrid neural network is constructed to create an accurate and continuously-updating scheme for mortality prediction in adult intensive care units, utilizing only vital signs and basic demographics as features.

5. In chapter 6, the work on adult mortality risk assessment is extended upon to develop a mortality risk assessment scheme for neonatal patients admitted to neonatal intensive care units, utilizing respiratory and heart activity variations along with basic demographics as features.

1.4 Significance

This thesis fills significant gaps in the healthcare monitoring literature. The techniques presented for measuring blood pressure (BP) and respiratory rate (RR) outperform previous literature, offering a strong solution for non-invasive and continuous measurement of these parameters in environments ranging from the home to the hospital. The algorithms for both blood pressure and respiratory rate are based on photoplethysmogram (PPG) and electrocardiogram (ECG) signals, which are widely used in devices from fitness trackers to medical-grade

hospital equipment. The sensors that record these signals are also non-invasive, unobtrusive, and can be continuously recorded - unlike many existing alternatives for measuring BP and RR. As such, the proposed algorithms for measuring BP and RR could be utilized by health and fitness device manufacturers to develop tools from fitness watches to non-invasive monitors for critical care environments.

Additionally, the mortality risk prediction schemes for adult and neonatal patients are the first to show that mortality risk can be continuously and accurately predicted using only vital signs and basic demographics, enabling ease-of-use and reducing the burden of care on healthcare providers in intensive care environments. As the algorithm only depends upon vital signs, it could be utilized by medical device manufacturers to develop wearable devices for automatic and continuous severity of illness assessment, both in hospital and in telehealth applications. This would minimize and potentially eliminate the time cost associated with currently used mortality risk assessment schemes, and would enable at-home monitoring for high-risk patients. Additionally, it would be possible to create an entirely non-invasive device for this purpose if the proposed schemes for measuring BP and RR were implemented alongside established techniques for non-invasively measuring the remaining vital signs, which would greatly improve patient experience.

Overall, the algorithms presented in this thesis offer solutions for a wide range of significant problems in the field of healthcare, with applications ranging from fitness to telehealth to critical care.

1.5 Document Organization

The structure of this thesis is as follows. Chapter 2 presents a thorough review of the literature, with focus placed on previous works in the areas of blood pressure monitoring, respiratory rate measurement, and mortality risk assessment. Relevant machine learning techniques are also discussed, with reference to previous works in the healthcare field.

Chapters 3-6 are the research chapters, focused on addressing the research problems outlined in Section 1.2 through achieving the objectives outlined in

Section 1.3. The following figure provides an overview of the structure of the research chapters, and the relationships between them.

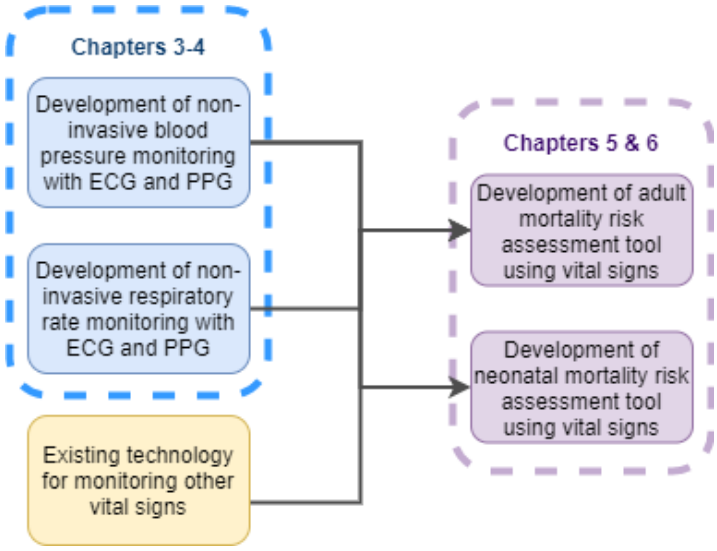


Figure 1.1: Conceptual framework illustrating the relationship between research problems

As Fig. 1.1 illustrates, Chapter 3 presents two machine learning methods for estimation of blood pressure from photoplethysmogram and echocardiogram waveform. The first scheme uses raw PPG and ECG waveforms as inputs to a hybrid neural network, while the second scheme improves computational efficiency by selecting features that describe the shape of the waveform and using these as inputs to a hybrid neural network. Respiratory rate measurement is considered in Chapter 4, where a scheme for extracting respiratory signals from heart activity information is proposed. Each extracted respiratory signal is assessed using a respiratory signal quality index, and machine learning is then utilized to predict the respiratory rate based on multiple signals extracted from heart activity. The development of these two schemes is strongly related, as they are the only two vital signs that do not have an existing method for continuous and non-invasive measurement. The same input signals are used for both works, allowing for the two schemes to be integrated into a single device readily.

After developing schemes for measuring vital signs, this thesis then turns to using vital sign information for enhanced prognostics tools in clinical settings. As shown in Fig. 1.1, Chapter 5 proposes a scheme for predicting mortality using basic demographics and statistics extracted from temporal vital signs. This

scheme is able to be readily updated throughout the stay, allowing for ongoing assessment of a patient's health and response to treatments. Building upon this, neonatal mortality risk assessment is then considered in Chapter 6, using only heart rate and respiratory rate information along with basic demographics to assess mortality risk in infants.

Following the research chapters, Chapter 7 concludes this thesis with several comments on future research directions.

Chapter 2

Background

This chapter presents a literature review of novel vital sign monitoring techniques and machine learning in healthcare, with emphasis on four key aspects: blood pressure measurement, respiratory rate measurement, and mortality risk prediction for both adults and neonates in the intensive care unit.

An earlier version of this literature review has been published in the following journal article:

[1] **S. Baker**, W. Xiang, and I. Atkinson, "Internet of Things for Smart Healthcare: Technologies, Challenges, and Opportunities", *IEEE Access*, vol. 5, pp. 26521-26544, November 2017

2.1 Vital Sign Monitoring

Vital signs quantify the status of several fundamental life-sustaining functions of the body. The four fundamental vital signs are heart rate (HR), body temperature, respiratory rate (RR), and blood pressure (BP) [64]. The monitoring of pulse rate is largely a solved problem, with wrist-based photoplethysmogram (PPG) sensors considered most comfortable for a long-term wearable system [65]. The operation of these sensors is illustrated below, and involves an LED shining light into the artery. Some light is absorbed by the blood, while the remainder reflects back to a photodiode. The amount of absorbed light varies while the heart beats, and as such heart activity waveforms can be extracted and used to assess parameters such as HR and blood oxygen saturation. Such sensors have already been widely implemented commercially, with many devices by brands

including Fitbit, Apple, and Garmin validated in the literature [66–69] and used for a variety of research projects [70].

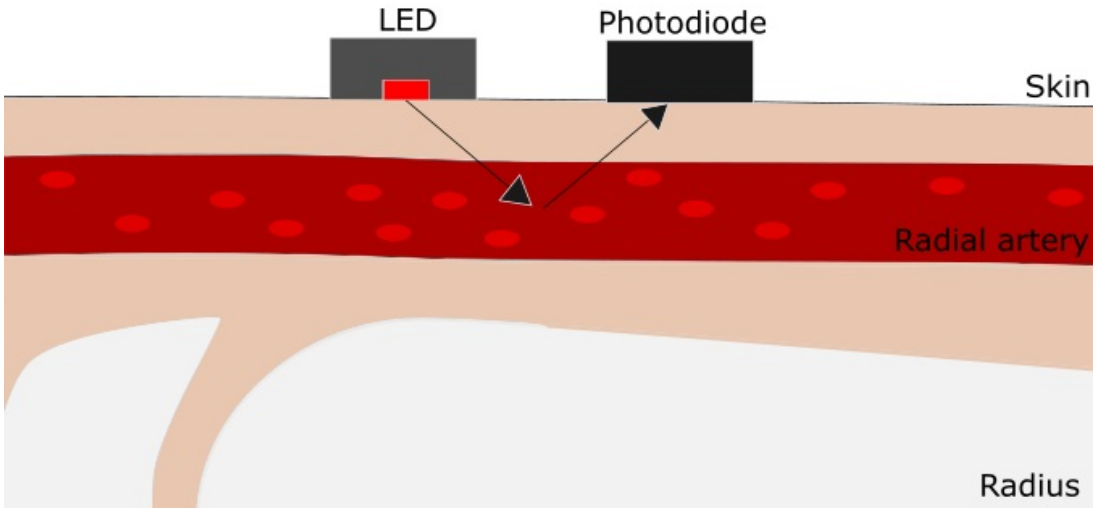


Figure 2.1: Photoplethysmographic pulse sensor

Aside from PPG sensors, echocardiogram (ECG) sensors can also be used to acquire information about heart health, including the vital sign of heart rate. ECG measurements are typically performed in healthcare settings, with multiple electrodes used to monitor the electrical activity of a patient's heart. However, recent research has shown that ECGs can be reduced to as few as one electrode, therefore making them more suitable for implementation into wearables including wrist bands and chest straps [71–73].

Another of the vital signs is body temperature, which can be used to detect hypothermia, heat stroke, fevers, and more. Measurement of body temperature is also a largely solved problems, with recent works surrounding the topic primarily focusing on thermistor-type sensors. In [74, 75], the common negative-temperature-coefficient (NTC) type temperature sensors were used, while positive-temperature-coefficient (PTC) sensors were considered in [76, 77]. In all studies, the thermistors were shown to measure a suitable range of temperatures for monitoring the human body, with acceptable levels of error. These small sensors could readily be included in a wrist-wearable device alongside PPG and ECG sensing capabilities.

The two vital signs that are currently missing from the wearable wrist-watch

picture are blood pressure and respiratory rate. The measurement of these two parameters has become the focus of much literature in recent years. In this section, existing methods in the literature for measuring blood pressure and respiratory are critically analysed.

2.1.1 Blood Pressure Measurement

Blood pressure (BP) is a key parameter for assessing patient health. Hypertension, a condition where BP is elevated above 140 mmHg for systolic blood pressure (SBP) or 90 mmHg for diastolic blood pressure (DBP) [78], is a leading risk factor for developing cardiovascular disease (CVD). It is also one of the most common chronic illnesses, affecting 32% of adult Australians. Of those affected, 68% had uncontrolled or unmanaged hypertension [79]. Complications resulting from hypertension lead to 9.4 million deaths per year [80]. Treatment to reduce blood pressure reduces the patient's risk of developing CVD [81], however treatment cannot commence until the condition is diagnosed.

Significant diagnostic benefit is also seen with accurate and continuous measurement of mean arterial pressure (MAP), which is quantified by the relationship between SBP and DBP:

$$MAP = \frac{SBP + (2 \times DBP)}{3} \tag{2.1}$$

MAP is a useful parameter for determining overall blood flow and thus the level of organ perfusion. Lower values for MAP can indicate high mortality risk conditions such as sceptic shock, as well as less critical but still serious conditions including syncope [82]. Conversely, elevated MAP causes cardiovascular strain and can lead to serious and potentially fatal CVDs including stroke [83].

BP is an extremely important parameter, however there are no commercially available devices capable of measuring BP continuously and non-invasively. The current gold-standard method for measuring BP continuously is intra-arterial monitoring, a highly invasive procedure that involves inserting an arterial line into a patient's blood stream [18]. This procedure must be performed in a sterile clinical environment and increases the risk of infection for the patient. Intra-arterial monitoring is therefore unsuitable for long-term BP monitoring.

Sphygmomanometers are a much more common method for measuring BP. These non-invasive devices are based on an inflatable cuff that is manually or automatically inflated beyond the expected SBP, causing blood flow to be cut off. The cuff is then slowly deflated, and the pressure at which audible blood flow sounds begin is the SBP. Pressure continues to be reduced, and then the pressure at which the blood sounds cease is the DBP [84]. While non-invasive, sphygmomanometers still cause significant discomfort and are even used to assess pain sensitivity in research applications [85, 86]. There are also several limitations of sphygmomanometers. Firstly, they cannot perform continuous BP monitoring [87]. Regular measurements are possible using an ambulatory sphygmomanometer, which can automatically inflate and deflate at regular intervals to take multiple measurements within a 24 hour period. However this is not truly continuous, and is disruptive to daily activities and sleep, and thus is not a sustainable long-term solution. Another issue is that vibrations in the arterial wall, which can be caused by conditions such as arrhythmias, can compromise the ability of a sphygmomanometer to accurately measure BP [84]. They also cannot be used on people with several pre-existing conditions, such as lymphedema [88]. Factors including cuff size and arterial wall stiffness (which varies with age) can also greatly affect the performance of sphygmomanometers [84].

With the limitations of existing devices, there is a strong need for new methods of measurement of blood pressure. Continuous, non-invasive, and comfortable tools for measuring blood pressure would greatly improve diagnostic ability and patient quality of life. The majority of the literature focuses on using manually extracted features of the electrocardiogram (ECG) and photoplethysmogram (PPG) signals to estimate BP using various mathematical approaches.

Several early studies [89–92] attempted to derive simple equations that could be used to accurately measure BP, based on the well-known inverse relationship between systolic BP and pulse transit time (PTT). PTT is the time that it takes for the pulse to travel from the heart to another point, usually the radial artery at the wrist. The early works focusing on this [89–92] each used a combination of ECG and PPG sensors, then derived expressions that utilised PTT and a variety of other parameters to estimate BP. Each of these works were limited

by a need to frequently recalibrate the included parameters to obtain accurate results. These problems arose even with the small patient samples, such as the 6 patients used to develop the model in [90] and 9 patients in [92].

As a result of the limitations around manually derived algorithms for BP estimation, several recent works have turned their attention to machine learning (ML) techniques for calculating BP. Some have considered ML to predict BP using basic demographics and health parameters including age and body mass index (BMI) [93] or using sphygmomanometers [94–96]. However, these methods are non-continuous, and as such the majority of the literature instead focuses on using ML to estimate BP from ECG and PPG signals [33, 34, 36–38, 97–100].

Several of these works [33, 38, 97] use the large Medical Information Mart for Intensive Care (MIMIC-III) database, which includes over 40,000 patient records from multiple critical care units in the period of 2001-2012. The aforementioned works extracting features including pulse transit time (PTT) from the ECG and PPG waveforms. These features were then used as the inputs to ML algorithms, including AdaBoost in [33], multi-regression in [97], and multivariate adaptive regression spline (MARS) analysis in [38]. While the MIMIC-III database is a valuable resource, it suffers from intra-waveform alignment issues as the PPG and ECG signals are not time-synchronised. This prevents accurate calculation of PTT and other time difference features between the ECG and PPG waveforms [101]. As such, the schemes using MIMIC-III to obtain PTT and other time-dependent parameters could not reliably be implemented in healthcare environments without further testing.

Two other recent works [37, 98] also used features including PTT extracted from ECG and PPG signals, measured with the same equipment across all participants. This likely improved consistency between measurements, however synchronisation between devices measuring each waveform would remain challenging. The databases used in these works were small, with 85 patients and 110 patients considered by Miao *et al.* [98] and Song *et al.*, respectively. Mean absolute errors (MAEs) of 6.13 mmHg and 4.54 mmHg for SBP and DBP, respectively, were achieved by the regression algorithm presented in [98]. Meanwhile, a MAE of 4.8 mmHg was reported for both SBP and DBP by [37], achieved with a deep

fully connected NN. These results are reasonable, however further validation of the algorithms on larger patient databases would be required to confirm these results.

The use of raw waveforms as input features is considered by [34] and [36]. In [34], regression algorithms including decision tree, support vector, adaptive boosting, and random forest are trained to predict SBP, DBP, and MAP using raw PPG waveforms. The adaptive boosting regressor performed the strongest, achieving MAE values of 3.97 mmHg, 2.43 mmHg, and 2.61 mmHg, respectively. However, the model shows signs of overfitting. The standard deviation (SD) is high at 8.901, and 16% of all measurements had errors greater than 15 mmHg.

Raw ECG waveforms are used in [36] to train a regressor neural network (NN) comprised of residual network and long short-term memory (LSTM) layers for SBP, DBP, and MAP prediction. The MIMIC-III database was used along with a second independent database, however synchronicity was not an issue given that a single waveform was used with no timing-dependent features derived. Results varied dramatically between databases, with the results on the MIMIC-III database showing a MAE of 7.10 mmHg and SD of 9.99 mmHg for SBP, along with a MAE of 4.61 mmHg and SD of 6.29 mmHg for DBP.

Several other works have investigated NNs for the prediction of BP, including long short-term memory (LSTM) networks. However, the works that investigated these used limited databases of 26 patients [100] and 96 patients [99], respectively. Each work used manually extracted features of the ECG and PPG waveforms as features. Minimal statistics were cited by each paper, with [100] citing root mean square errors (RMSEs) of 2.571 mmHg and 1.604 mmHg for SBP and DBP, respectively, while [99] reported RMSE values of 3.90 mmHg and 2.66 mmHg for SBP and DBP, respectively. These values show promise for LSTM networks, but significant further validation on larger databases and with more rigorous statistical analysis is clearly required.

Aside from densely-connected and LSTM NNs, another candidate for BP estimation is found in convolutional neural networks (CNNs). These networks are well-known for their high performance in computer vision tasks, and raw PPG and ECG waveforms could be effectively considered as images. They have

previously been used to analyse ECG signals to detect conditions including atrial fibrillation [51–53], however to the best of the author's knowledge they have not been used for BP detection from such waveforms.

The results of both [34] and [36] indicate that raw waveforms can be used to predict BP with some success, however are not yet sufficient for clinical use. Meanwhile, limited studies on small databases have suggested that NNs may be suitable for analysing ECG and PPG signals, however there have been no attempts to use NNs to analyse both waveforms simultaneously for the estimation of SBP, DBP, and MAP. These are significant gaps in the literature.

In terms of assessing the performance of BP measuring algorithms and devices, the most common technique in the literature is comparison to the Association for the Advancement of Medical Instrumentation (AAMI) standard and British Hypertension Society (BHS) protocol for assessment of BP devices, each of which is used for the validation of devices used in clinical applications. The AAMI standard [102] is a pass-or-fail test that states that a device must be tested on ≥ 85 people, achieve a mean absolute error (MAE) of ≤ 5 mmHg and standard deviation (SD) of ≤ 8 mmHg compared to a gold-standard method to pass, otherwise the device fails. Meanwhile, the BHS protocol [103] assigns a grade between A-D depending on how many measurements have errors less than several thresholds when compared to a gold-standard method. Only devices that achieve 'A' or 'B' grades are recommended for clinical use. The criteria for reaching different grades under the BHS protocol is outlined in Table 2.1 below.

Table 2.1: Grading criteria defined by the BHS protocol.

	Absolute Difference (mmHg)		
Grade	**≤5**	**≤10**	**≤15**
A	60%	85%	95%
B	50%	75%	90%
C	40%	65%	80%
D	Worse than C		

In terms of these two standards, none of the aforementioned works have achieved 'pass' grades for the AAMI standard and 'A' grades for the BHS standard across SBP, DBP, and MAP. As such, it is unlikely that any of these schemes

are suitable for clinical use. Thus, developing an algorithm that meets the requirements set by the AAMI and the BHS remains a significant gap in the literature.

2.1.2 Respiratory Rate Measurement

Another of the vital signs is respiratory rate (RR), or the number of breaths a patient takes per minute. Respiratory rate abnormality is one of the earliest indicators of critical illness. Elevated RR has been linked to clinical deterioration after emergency department discharge [104], cardiac arrest [105], pneumonia in children [13, 14] and general mortality risk [106]. Meanwhile, fluctuations in RR have been found to be strongly linked with patient stability [107].

Despite the importance of RR, it has been historically under-recorded compared to other vital signs [15, 104, 108, 109]. One study has found that nurses do not measure RR in 50% of cases due to time constraints and lack of equipment [15].

In terms of equipment, the most common tools used for automatic RR measurement are oronasal systems comprised of capnography, temperature, or moisture sensors, however these have not been widely adopted [108].

Manual measurement through counting the number of breaths a patient takes over a one-minute period remains the accepted method for determining RR [15]. This method has the obvious limitation of being non-continuous, however manual measurement can also be negatively impacted by patient awareness, time constraints, interruption, and patient agitation [14–16, 110]. The need for manual measurement of respiratory rate is also time consuming, with one study finding that nurses spent up to 7.2% of their time on manual patient assessment [17].

Due to the diagnostic importance of RR and the major limitations in existing methods of measuring this vital sign, many previous works have investigated a variety of sensors and devices for measuring respiratory rate, including thermistors [111], microphones [112], fibre optic vibration sensor [113], pressure sensors, stretch sensors [74, 114, 115]. Each of these sensors has shown promise, however they are not highly wearable. Thermistor-based devices involve sensor placement in the nose, while the other sensor types are typically chest-worn.

An emerging candidate for wearable RR measurement is the extraction of respiration signals from PPG and ECG signals. There is significant movement associated with breathing, along with changes in intrathoracic pressure. This results in respiration modulating heart activity signals in three main ways - baseline wander (BW), amplitude modulation (AM), and respiratory sinus arrhythmia (RSA) modulation [116]. The extraction of these modulations and subsequent use in estimating RR has been considered in several recent works [39, 40, 42, 43, 117], with mixed success. Machine learning techniques including linear regression and support vector regression have been considered [43], but this has not been explored broadly on large datasets.

Overall, there are still substantial improvements to be made in the field of measuring RR automatically and continuously. The most promising method is the extraction of respiration modulations from the ECG and PPG signal. Machine learning has shown some promise on smaller datasets, however its potential has not been widely explored for this purpose.

2.1.3 Summary

Blood pressure and respiratory rate remain the two most challenging vital signs to measure, particularly in a continuous and non-invasive manner. Recent literature suggests that the heart activity signals of ECG and PPG contain information that can be used to calculate both BP and RR, however existing literature has not yet met the standards required for clinical use. Machine learning techniques have been explored to improve BP and RR estimation from ECG and PPG signals, but this has not yet led to clinically-suitable devices.

2.2 Mortality Risk Assessment

The automatic and continuous measurement of vital signs can support a wide variety of healthcare applications, including the assessment of mortality risk in critical care environments. Intensive care units (ICUs) treat the most critically ill patients and have the highest mortality rate of all hospital units [9], ranging between 11.3-12.6% [118]. Mortality risk assessment can aid healthcare pro-

fessionals in making treatment decisions and determining the effectiveness of treatments [19].

Assessment of mortality risk has been considered in the literature for two main age ranges: adults and neonates. In this section, the literature on mortality risk assessment for adult and neonatal patients is thoroughly examined.

2.2.1 Adult Mortality Risk Assessment

Several points-based schemes are currently used in adult ICUs to quantify mortality risk. The most prevalent of these are the Acute Physiology and Chronic Health Evaluation (APACHE) score [119], Simplified Acute Physiology Score (SAPS) [120] and the Sequential Organ Failure Assessment (SOFA) score [121]. APACHE and SAPS have each undergone several updates, with the most recent versions being APACHE-IV and SAPS-III, respectively. Despite this, APACHE-II and SAPS-II remain the most commonly used versions of these schemes worldwide [122].

These scores rely on health parameters that are often time-consuming and difficult to obtain, as well as manual data entry. In addition to these disadvantages, it has been found that the performance of these schemes decrease fairly rapidly over time, with SAPS-II found to be out of calibration within 12 years of its development [30]. Several recent studies have also found calibration issues with APACHE, SAPS, and SOFA [123–125]. Changing patient population and medical treatments account for much of the calibration loss, as each of these schemes is trained on a singular dataset at a particular point in time [30]. It has also been observed that the schemes perform poorly on cohorts from different regions than those they were trained on, including Europe and Singapore [123, 125]. This indicates that insufficient consideration of diverse populations affects the performance of the schemes.

The limitations of traditional scoring systems has lead to a rise in researchers investigating machine learning for mortality prediction [35, 55–61] and related applications including detecting sepsis risk [46–48, 126] and general clinical deterioration [49, 50]. Techniques used included random forest [35], logistic regression [55], gradient boosting [55, 59, 60], and neural networks [56–58, 61]. These recent

works have focused on using machine learning techniques for binary classification; that is, determining whether a patient is a mortality risk or not. Neural networks were the most commonly used, and showed strong performance on smaller numbers of variables than was achieved with other techniques. Neural networks also offer the benefit of being reasonably easy to configure for continuous learning, ensuring that calibration is always up-to-date. This in turn leads to enhanced ability to generalise to current populations, even as populations, treatments, and outcomes change with time.

Most recent works [35, 56–60] have developed schemes that are heavily dependent on laboratory results including those obtained from extensive blood, urine, breath, and other clinical analyses. These parameters are often complex and time-consuming to obtain, and then additionally require entry into the patient's medical records for the scores to be calculated.

Performance of mortality risk assessment schemes is typically quantified using the area under the receiver-operator curve (AUROC), which compares the false-positive rate (percentage of incorrect predictions of mortality) with the true-positive rate (percentage of correct predictions of mortality) for a varying cut-off threshold for mortality prediction. AUROC can range from 0-1. In terms of this parameter, the highest performing of the aforementioned schemes was that presented in with an AUROC of 0.94, however the scheme depended on other scoring schemes including the All Patients Refined Diagnosis Related Groups (APR-DRG) and Medicare Diagnosis Risk Groups (MS-DRG). The grouping of patients under these schemes is dependent on doctor diagnosis, introducing a significant human bias. APR-DRG and MS-DRG are also not used in all hospitals, limiting the broader usage.

Another strong work was that presented in [59], achieving an AUROC of 0.927. However, this scheme depended on 148 features comprised predominantly of complex laboratory results. The high number of parameters required would place significant burden on healthcare workers in terms of measurement and data entry, limiting the practical usefulness of the scheme.

Other schemes have attempted to use fewer variables, such as the scheme presented in [55] which investigated the use of only vital signs as parameters.

However, an AUROC of only 0.65 was achieved. Through expanding their vital sign feature vector to include Glasgow Coma Score (GCS) and SAPS-II score, they were able to raise the AUROC to 0.84 - however this again introduces dependency on complex parameters. Nonetheless, this work showed promise for the use of vital signs in predicting mortality risk. Another work to identify the importance of vital signs in critical care was that presented in [126], which found that 10 of the 20 most important parameters for detecting onset of septic shock were derived from vital signs. Further exploration of the use of vital signs in predicting mortality should be considered given that they measure the most critical functions of the human body [127].

Another trend in the literature is the assessment of mortality risk for the entire stay using data acquired at admission. However, one recent work [61] has identified that this is inflexible given that a patient's condition can dramatically change during the stay, and investigated the use of a shifting window to identify mortality risk at any time. However, this scheme depended on 48 hours of extensive laboratory values that would be time-consuming and difficult to continuously update. Further exploration of repeatable mortality risk prediction would undoubtedly be valuable for assessing patient response to treatments and detecting any deterioration.

While machine learning has been broadly explored in the literature, relatively little research has investigated the use of neural networks (NNs) for mortality prediction. Early works investigating NNs focused on simple feed-forward networks and achieved comparable results to traditional scoring schemes [128–130]. More recent works have identified long short-term memory (LSTM) networks as candidates for mortality prediction [56, 61, 131], as have hybrid networks comprised of convolutional neural network (CNN) and LSTM layers [57].

Overall, the major limitations in current mortality risk assessment schemes for adult patients are the use of extensive and complex features, and the focus on predicting mortality for the entire stay using data taken at and immediately after the time of admission. Further investigation into developing a mortality scheme that is able to be continuously recalculated throughout the stay with simple features is required, and the use of NNs is a promising avenue to achieve

this goal.

2.2.2 Neonatal Mortality Risk Assessment

Complications resulting from preterm birth are the leading cause of death in children under 5 [132], causing over 1.1 million deaths per year globally [133]. Preterm infants are often admitted to Neonatal Intensive Care Units (NICUs). As many as 84.41% of infants with birthweight between 500g-1499 g are admitted to NICU in the United States, while 48.17% of infants weighing 1500-2499 g are admitted.

Several scoring schemes comparable to APACHE and SAPS are used in the NICU for mortality risk assessment. One common scheme is the Clinical Risk Index for Babies (CRIB-II) [134], a simpified version of the earlier CRIB score [135]. The Score for Neonatal Acute Physiology (SNAP) [136] and SNAP Perinatal Expansion (SNAPPE) [137] are also commonly used, as are their successors SNAP-II and SNAPPE-II. Other scores such as the Berlin score [138] and Neonatal Mortality Prognostic Index (NMPI) [139] are also in use to a lesser extent.

The limitations of these scores are comparable to the limitations of the APACHE, SOFA, and SAPS scores for adult mortality prediction. Each of the scores for predicting mortality risk in neonates was developed over fifteen years ago, with the publication of CRIB-II in 2003 marking it as the most recent scheme. Recent works have identified a significant decrease in performance for scores in the SNAP/SNAPPE family of scores [140] and CRIB-II [19]. A recent extensive review of multiple scoring schemes [31] concluded that updated and enhanced scoring systems are required to account for the significant advancements in neonatal care.

In addition to calibration loss, several scoring schemes feature complexity or inflexibility in the variables used. CRIB-II is the simplest scoring scheme, relying on five variables, however each of these are not updated after admission, limiting the usefulness in determining patient response to treatment. Meanwhile SNAP-II, SNAPPE-II, Berlin score, and NMPI each use complex variables including PO_2/FiO_2, serum pH, presence of seizures, urine output, base excess, and more.

Due to the limitations of existing scores, several recent works have investi-

gated alternative methods for predicting neonatal mortality risk. Studies have investigated logistic regression [141, 142], densely-connected neural networks [63], random forest [143], and fusion of multiple algorithms into a so-called "super-learner" [62] for binary classification of mortality risk in neonates. Of these schemes, the highest performing were those based on neural network techniques [62, 63] and random forest [143].

While the use of machine learning can aid in overcoming calibration issues, the works in the literature remain limited by the selection of variables that are challenging to measure regularly. Parameters considered by the schemes in the literature include laboratory results [62], maternal characteristics [63, 141], existing conditions [62, 63], and more. Other parameters are challenging to quantify definitively, such as the condition of the baby through visual inspection used in [141]. The use of such parameters increases burden on healthcare workers and thus limits the usefulness of the scheme.

On the other hand, several works in the literature are limited by their selection of variables that do not change, including pre-birth and start-of-labour characteristics [141], blood oxygen at admission [142], and respiratory support within the first 24 hours after birth [142]. Fixed variables prevent recalculation of risk during the stay, thus preventing assessment of response to treatment and other changes to the infant's condition throughout their stay.

Given the similarities between adult and neonatal mortality risk prediction, neural networks including CNNs and LSTMs are also strong candidates for assessment of mortality risk prediction in infants. An ideal scheme would include variables that are easy to calculate regularly, utilising a machine learning strategy that would ensure calibration remains strong throughout the future.

2.2.3 Summary

Mortality risk prediction is an essential component of ICU decision making. Accurate prediction of this parameter enables treatment decisions, resource allocation, and assessment of patient condition. However, existing schemes are limited by the use of complex or fixed variables to quantify mortality risk. They are further limited by their assessment at the start of the stay, without further up-

dates during the stay to enable constant reassessment of the patient's condition. There are indications in the literature that machine learning techniques may be suitable for mortality risk prediction, however schemes that have attempted the use of machine learning have remained dependent on a high number of complex input variables. There remains a significant gap in the literature regarding the use of machine learning on non-complex variables that can be readily assessed throughout the stay to enable continuously updating mortality risk assessment.

2.3 Neural Networks

Machine learning has been identified as a candidate solution for blood pressure measurement [33, 34, 36–38, 93–100], respiratory rate measurement [43, 144], and mortality risk assessment in both neonatal [62, 63, 141–143] and adult [35, 55–61, 131] patient cohorts. Recently, several works in these healthcare applications have focused on neural networks specifically [34, 36, 56, 57, 61–63, 99, 100, 131] due to their enhanced ability to learn complex patterns compared to traditional machine learning techniques such as logistic regression and random forest. As such, this thesis has placed focus on neural network techniques for healthcare applications.

Every neural network shares some core components. Firstly, neural networks are comprised of three layer types: input, hidden, and output layers. The input layer receives the features, the hidden layers are where the inputs are processed, and the output layer returns the result of the operation. Within each layer there are a number of units (also called cells or neurons) that perform a mathematical operation. They each have their own weights and biases that are updated through the process of training using an optimization algorithm. It is this process that allows the network to learn from the data it is seeing. Units within the hidden layers are typically called hidden units. In training a neural network, a loss function is used to tell the network which parameter it is aiming to optimize (for example, error or accuracy).

While neural networks have these characteristics in common, there are also many differences across the various types of neural networks. In this section,

several fundamental neural network structures are introduced and described in detail.

2.3.1 Feed-Forward Neural Networks

Feed-forward neural networks (FFNNs) are also known as densely-connected, fully-connected, or dense NNs. FFNNs are the most basic form of NN, with each cell in a hidden layer connected to every cell from the previous layer. The major advantage of FFNNs is that they are mathematically simple and therefore highly efficient for small feature vectors. In many cases, if a FFNN were to perform well, there would be little need for more complex NNs to be utilized. However, works utilizing FFNNs for applications such as mortality prediction [128–130] saw little improvement compared to manually-derived scoring systems. A single layer of a FFNN can be mathematically described as follows:

$$y_{\mathrm{i}} = g(x_{\mathrm{i}} \bullet w_{\mathrm{y}} + b_{\mathrm{y}}) \tag{2.2}$$

where y_{i} and x_{i} represent the output and input of layer i respectively, while w_{y} and b_{y} are the weights and biases, learned using the powerful Adam optimizer. The symbol '$\bullet$' denotes element-wise matrix multiplication, and $g()$ represents the activation function chosen. There are several common activation functions, including ReLU ($g(z) = max(0, z)$), sigmoid ($g(z) = \frac{1}{1+e^{-z}}$), and tanh ($g(z) = tanh(z)$).

An example of a FFNN with an input layer, three hidden layers, and output layer is shown in Figure 2.2 below.

Aside from multi-layer FFNNs, it is common for a singular densely-connected layer to be used as the output layer of other neural network types. This will be illustrated in the following subsections.

2.3.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are broadly used in image and video recognition tasks due to their ability to identify patterns. They have found use in healthcare applications for tasks related to blood pressure and respiratory

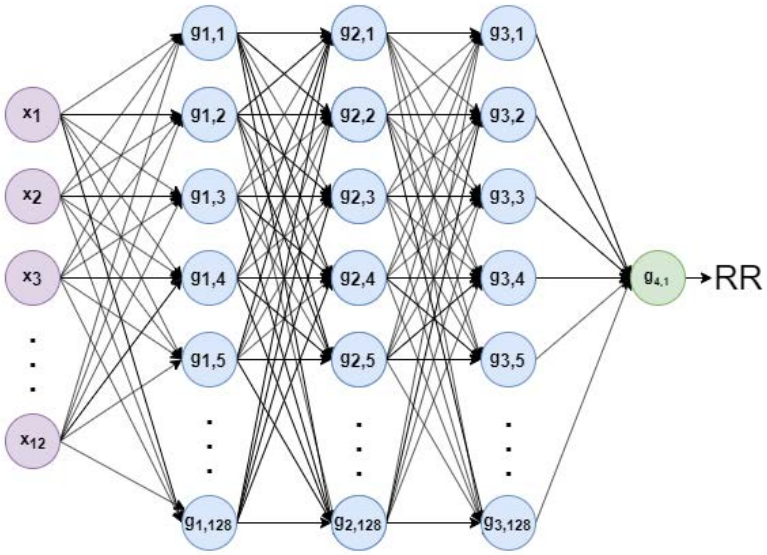


Figure 2.2: Example of a FFNN with three hidden layers.

rate measurement, such as recognizing variation in Korotkoff sounds [54] and detecting heart abnormalities in ECG signals [51–53]. They have also been considered in combination with other network types for mortality risk assessment [57]. Typically, they are most useful in applications where there are a large number of input variables and patterns are not already known. A single CNN layer is described mathematically as follows:

$$y_{\mathrm{j}}{}^{\mathrm{i}} = g(\sum_{n=1}^{N} w_{\mathrm{jn}}{}^{\mathrm{i}} * x_{\mathrm{m}}{}^{(\mathrm{i\text{-}1})} + b_{\mathrm{j}}{}^{\mathrm{i}}) \tag{2.3}$$

where y_j^i is the output jth feature map of the ith layer, after convolution has been performed and passed through the activation function. The symbol $*$ represents the convolution operation. The parameter $w_{\mathrm{jn}}{}^{\mathrm{i}}$ is the nth weight of the jth feature map from the previous layer, where $n = 1, ..., N$. The term $x_{\mathrm{m}}{}^{(\mathrm{i\text{-}1})}$ represents the outputs of the previous $(i-1)$th layer, and lastly b_j is the jth bias term of the lth layer.

CNNs can be implemented for one-dimensional, two-dimensional, or three-dimensional data. Time-series data, such as ECG and PPG signals, are examples of one-dimensional inputs. An example of a CNN with three hidden units and a densely-connected output layer is illustrated in Fig. 2.3.

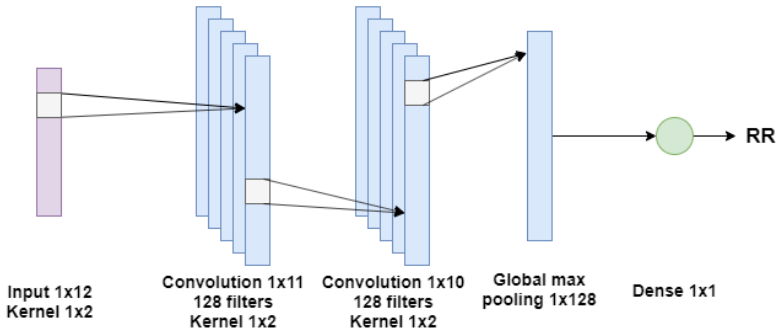


Figure 2.3: Example of a CNN with three hidden layers.

2.3.3 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are largely used for sequential data based on their enhanced ability to 'remember' what they have learned in the past. This makes the suitable for tasks such as handwriting recognition and language processing. A traditional RNN is mathematically expressed through two equations. Firstly, the equation that defines the activation value from time step t, with the activation function denoted by g, is as follows:

$$a_t = g(w_{\mathrm{aa}} a_{(t-1)} + w_{\mathrm{ax}} x_t + b_{\mathrm{a}}) \tag{2.4}$$

where a_{t} is the activation value from time-step t that will is inputted to the following time-step $t+1$ and $a_{(\mathrm{t\text{-}1})}$ is the activation value from time-step $(t-1)$, providing information about the past. The parameter x_{t} is the input for time-step t, and b_{a} are the relevant biases for calculating the activation value. w_{aa} and w_{ax} are the weights for the activation value $a_{(\mathrm{t\text{-}1})}$ and input x_{t} respectively.

The second relevant equation for RNNs gives the predicted output, and can be defined as follows:

$$y_{\mathrm{t}} = g(w_{\mathrm{ya}} a_{\mathrm{t}} + b_{\mathrm{y}}) \tag{2.5}$$

where y_{t} is the output of layer t, a_{t} is the activation value of layer t, and w_{ya} and b_{y} are the weights and biases used for calculating the output prediction, respectively. The activation function is again denoted by $g()$

A graphical example of an RNN with three hidden layers followed by a densely-connected output layer is illustrated in Figure 2.4.

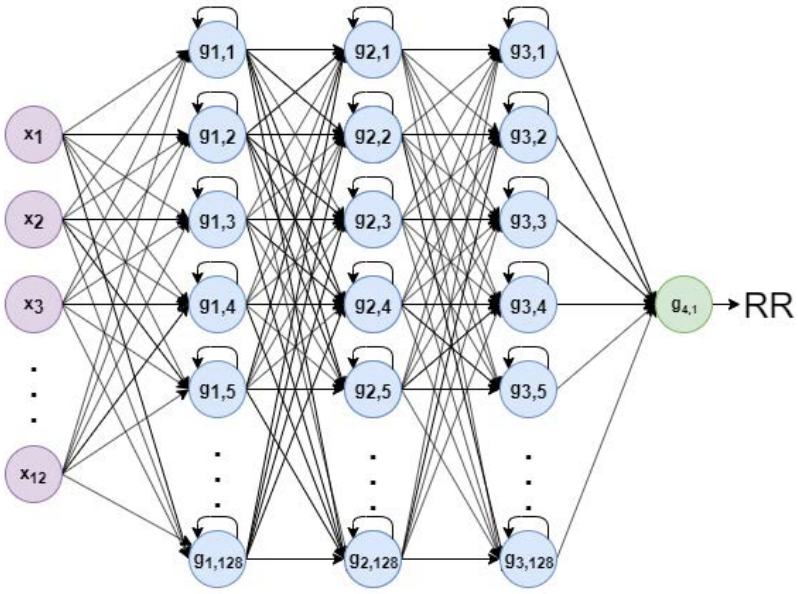


Figure 2.4: Example of an RNN with three hidden layers.

2.3.4 Long Short-Term Memory Networks

Long Short-Term Memory (LSTM) networks are an advanced type of recurrent neural network, offering several additional benefits over simple RNN. LSTM introduces several gates that make decisions regarding what to remember and what to forget. LSTMs have been widely used in healthcare applications for tasks such as predicting sceptic shock [145], seizure detection [146], cancer prediction [147], heart anomaly detection [148] and blood pressure estimation [99, 100].

The following series of equations represent the process for updating the cell state c_{t} in a single layer t of an LSTM network. Activation functions are again represented by $g()$, however activation functions used often vary for the gate equations and final output equation.

$$\tilde{c}_{\mathrm{t}} = \tanh(w_{\mathrm{c}}[\, a_{(\mathrm{t\text{-}1})}, x_{\mathrm{t}}] \, + b_{\mathrm{c}}) \tag{2.6}$$

$$f_{\mathrm{t}} = g(w_{\mathrm{f}}[\, a_{(\mathrm{t\text{-}1})}, x_{\mathrm{t}}] \, + b_{\mathrm{f}}) \tag{2.7}$$

$$u_{\mathrm{t}} = g(w_{\mathrm{u}}[\, a_{(\mathrm{t\text{-}1})}, x_{\mathrm{t}}] \, + b_{\mathrm{u}}) \tag{2.8}$$

$$o_{\mathrm{t}} = g(w_{\mathrm{o}}[\, a_{(\mathrm{t\text{-}1})}, x_{\mathrm{t}}] \, + b_{\mathrm{o}}) \tag{2.9}$$

$$c_{\mathrm{t}} = u_{\mathrm{t}} \bullet \tilde{c}_{\mathrm{t}} + f_{\mathrm{t}} \bullet c_{(\mathrm{t\text{-}1})} \tag{2.10}$$

$$a_{\mathrm{t}} = o_{\mathrm{t}} \bullet g(c_{\mathrm{t}}) \tag{2.11}$$

Eqn. (2.6) represents the calculation of candidate values $\tilde{c}_t$ that may be used to update the cell state c_t. Then, (2.7)-(2.9) show the calculation of the forget gate f_t, update gate u_t and output gate o_t respectively. Finally, the cell state c_t is updated in (2.10), while the layer output is determined in (2.11). The '•' symbol in (2.10) and (2.11) represents element-wise matrix multiplication.

In (2.6)-(2.9), w_c, w_f, w_u and w_o refer to the learned weights for their respective operations, while b_c, b_f, b_u and b_o are the learned biases. Additionally, the parameter $a_{(t-1)}$ refers to the output of the previous layer, while x_t is the input for time-step t. Eqn. (2.10) utilizes the results of (2.6)-(2.8) as well as the cell state of the previous time step, $c_{(t-1)}$ to update the cell state, and (2.11) uses the resultant c_c as well as the output gate results.

An example of a LSTM NN with three hidden layers and a densely-connected output layer is illustrated in Fig 2.5.

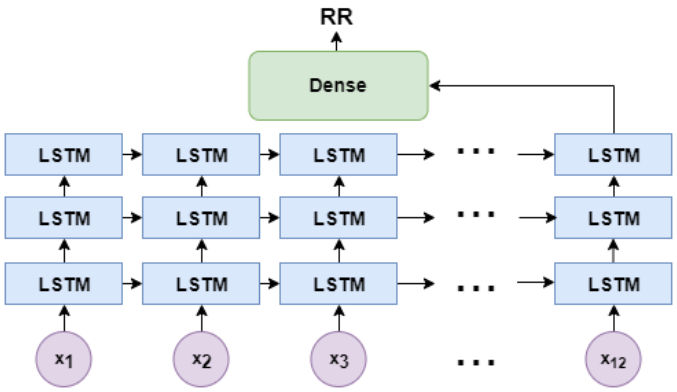


Figure 2.5: Example of an LSTM NN with three hidden layers.

A variation on LSTM is Bidirectional LSTM (BiLSTM). The LSTM network described above feeds data through the network from beginning to end of the sequence. BiLSTM also does this, but additionally passes the data through the network in reversed order. Results of both forward and reversed passes are then concatenated after each layer before being passed to the next layer. This enables the network to learn from both past and future data. BiLSTMs follow the same mathematical structure as unidirectional LSTM, applying this structure to both passes of the data.

An example of a BiLSTM network is illustrated in Fig. 2.6. A single bidirec-

tional layer is comprised of both forward and backward LSTM pass, as well as the concatenation operation that combines them.

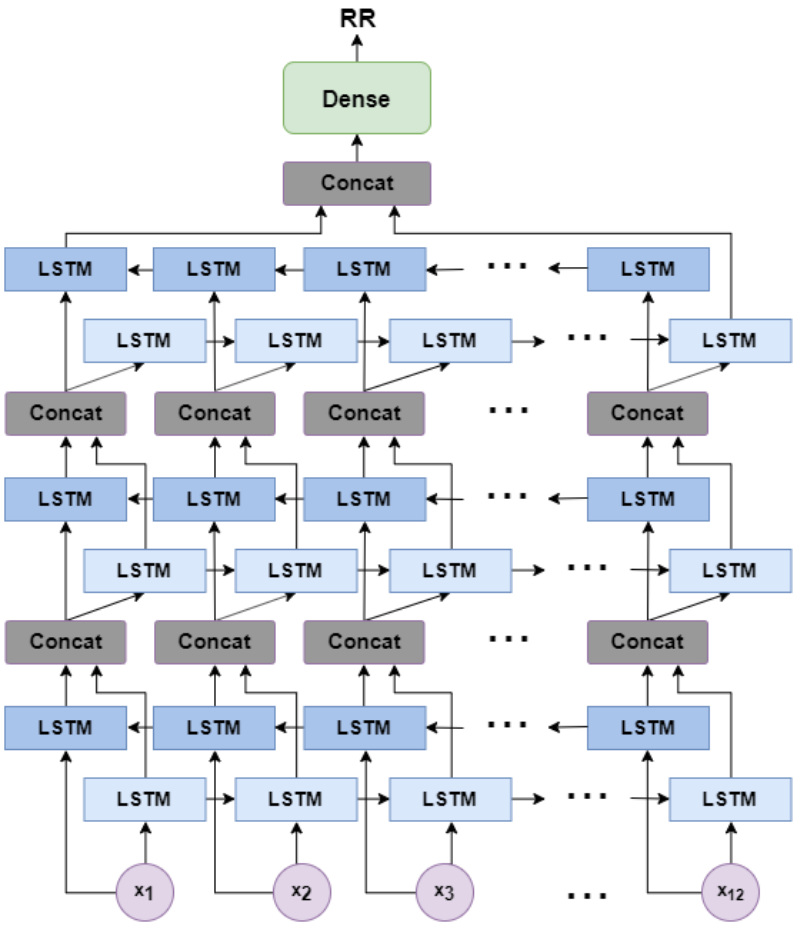


Figure 2.6: An example of a BiLSTM NN with three hidden layers.

2.3.5 Summary

In this section, several NN structures that have been previously used in the literature for healthcare applications have been presented. FFNNs are the most mathematically simple of these, however have found limited success in healthcare applications previously. Despite this, a fully-connected output layer is a standard feature of most NN structures. CNNs are strong contenders for large feature vectors due to their ability to recognize previously unknown patterns, while LSTM NNs have been found to perform strongly on sequential data. The use and hybridization of these networks could provide substantial value to the applications of blood pressure measurement, respiratory rate measurement, and mortality risk prediction

2.4 Conclusion

This literature review has conducted a comprehensive analysis of existing works and technologies for monitoring vital signs and overall severity of illness. Through this analysis, several key research gaps have been identified.

Firstly, there is no known scheme for non-invasive and continuous blood pressure measurement that meets the key industry standards outlined by the British Hypertension Society and the Association for the Advancement of Medical Instrumentation. This indicates that existing schemes would not be suitable for clinical implementation, leaving a substantial gap in the literature.

A similar issue is seen with respiratory rate measurement. Existing methods are either manual or obstructive and uncomfortable. Several pilot studies have considered non-invasive methods for continuous respiratory rate measurement, however low error has not been achieved on large datasets. A significant research gap remains in developing a robust technique for automatic, continuous and non-invasive respiratory rate measurement that is highly accurate.

Currently, blood pressure and respiratory rate are the only vital signs that cannot be measured continuously and non-invasively in clinical settings. The development of schemes that fill these gaps would strongly support the development of enhanced diagnostics and prognostics tools, including those for mortality risk assessment.

In terms of assessing severity of illness or mortality risk in critical care patients, existing schemes suffer from a wide range of issues including time cost, complexity of variables, and poor accuracy. The development of an easy-to-interpret scheme based on parameters that can be automatically recorded would fill a substantial gap in the literature.

This thesis addresses these significant research gaps, providing a broad range of techniques designed for implementation in environments ranging from health trackers to critical care units.

Chapter 3

Deep Learning for Blood Pressure Estimation using Electrocardiogram and Photoplethysmogram Data

This chapter contains materials published in the following article, which has been accepted for publication with *Computer Methods and Programs in Biomedicine*:

[4] **S. Baker**, W. Xiang, and I. Atkinson, "A Hybrid Neural Network for Continuous and Non-Invasive Estimation of Blood Pressure from Raw Electrocardiogram and Photoplethysmogram Waveforms," accepted by *Computer Methods and Programs in Biomedicine* in May 2021.

This chapter also contains materials from the following manuscript:

[149] **S. Baker**, W. Xiang, and I. Atkinson, "A Computationally Efficient CNN-LSTM Network for Estimation of Blood Pressure," manuscript prepared, publication to be pursued following the acceptance of [4].

3.1 Introduction

Blood pressure (BP) is a key diagnostic tool for a variety of life-threatening conditions. Elevated BP, or hypertension, is a major risk factor for cardiovascular disease (CVD), contributing to the deaths of 9.4 million people every year [80]. Additionally, poor organ perfusion can be identified through the measurement of BP-derived parameters, particularly mean arterial pressure (MAP). MAP is useful in determining overall blood flow and thus the level of nutrient delivery

to organs, and therefore is routinely measured when dealing with high-mortality conditions like septic shock [82]. MAP that is too low can lead to shock, syncope, and poor perfusion to organs, while elevated MAP places strain on the cardiovascular system and can eventually to various CVDs including stroke [83].

Despite the importance of monitoring BP, there are currently no commercially available devices capable of continuous and non-invasive BP measurement that have been approved for medical use. Currently, the gold-standard method for continuous and accurate BP monitoring is intra-arterial monitoring, which involves the invasive insertion of a catheter equipped with a pressure transducer into a patient's artery [18]. This is clearly not suitable for long-term monitoring as it must be performed in a clinical environment and it increases infection risk for the patient. Typically, BP is measured using less invasive sphygmomanometers, cuff-based devices which is manually or automatically inflated to determine BP. However, sphygmomanometers are incapable of continuous monitoring and cause significant discomfort to many patients [87]. They also cannot be used on people with several pre-existing conditions, such as lymphedema [88].

There is a clear need for improved methods of clinical and at-home BP monitoring, especially for high-risk patients. As such, many recent works have investigated methods for non-invasive measurement of BP. One promising area of research lies in machine learning (ML). ML techniques have been used to estimate BP from various health factors such as age and gender [93], as well as for improving sphygmomanometer measurements [95, 96].

More recently, many researchers have investigated the calculation of BP from electrocardiogram (ECG) and photoplethysmogram (PPG) signals [33, 34, 36–38, 97–100]. In [33, 38, 97], the Medical Information Mart for Intensive Care III (MIMIC-III) database was used to obtain features such as pulse transit time (PTT) and other manually extracted features of the ECG and PPG waveforms. These were then used with algorithms including AdaBoost in [33], multi-regression in [97] and multivariate adaptive regression spline (MARS) analysis in [38]. Unfortunately, the MIMIC-III database used by these works suffers from intra-waveform alignment issues that make calculation of PTT and other time-dependent features between ECG and PPG signals unreliable [101]. As such,

these schemes could not be reliably applied in healthcare applications.

In [34], raw PPG waveforms are used to train an AdaBoostR algorithm to estimate SBP, DBP and MAP. This model was trained on a small subset of the MIMIC-II data of 1,323 records and not validated on a distinct testing set. In results presented from the training set, the model was clearly suffering from a large number of high-range errors and large standard deviation (SD), particularly in SBP estimation. This indicates that the model had overfit to the training data, and therefore would be unlikely to perform strongly on new data.

Meanwhile, in the recent paper [36], raw ECG waveforms are used to train a neural network for SBP, DBP, and MAP prediction. Testing was performed on MIMIC-III data, as well as a second independent database. Results across these two databases varied significantly, with the scheme shown to not perform as strongly on the large MIMIC-III database. It is likely that utilizing both PPG and ECG data would significantly improve performance, and PPG signals are comparatively easy to obtain from wearable devices compared to ECG signals.

Two recent works [37, 98] obtained features from ECG and PPG signals before using ML techniques to predict BP. Features considered included PTT, which was measured using the same equipment across all participants. This likely improved synchronization between devices when compared to the waveform synchronity issues in [33, 97, 150], however clock drift could still impair PTT calculation over longer periods. Each of these works built small databases using measurements from healthy volunteers, with [98] obtaining readings from 85 patients and [37] using 20-second segments from 110 subjects. Good results were presented in both works, however larger databases would be needed to verify that these schemes would perform well on a wide range of patients.

In this chapter, two hybrid neural network (NN) schemes are proposed. Each scheme incorporate temporal convolutional neural network (CNN) and long short-term memory (LSTM) layers for the estimation of BP. In each hybrid NN, he CNN layers act to identify the most important features, while the LSTM layers have a strong ability to remember information and thus identify relationships between features.

The first scheme proposed in this chapter uses 5-second windows of both raw

ECG and PPG waveforms as inputs. By using raw waveforms as inputs, the hybrid CNN-LSTM network is able to learn from all of the available information, rather than from manually identified features. Avoiding manual feature selection also has the advantage of removing human bias from training and testing. Additionally, the use of a short, 5-second window of data ensures that BP can be calculated rapidly and continuously, and is less prone to suffering from interference than longer windows.

In the second scheme proposed by this chapter, focus was placed on improving computational efficiency. Twelve features describing the shape of 5-second segments of ECG and PPG are derived and used as input features. This strategy minimises the risk of human bias as it focuses on describing what can be seen in the waveform, rather than deriving complex features such as PTT.

The remainder of this chapter is structured as follows. Section 3.2. presents the methodology, including schemes for preprocessing, signal quality assessment and developing the hybrid NNs for BP estimation. Section 3.3. discusses the results of testing conducted on the NN algorithms to assess their performance. Finally, Section 3.4. briefly concludes this work and summarises its significance.

3.2 Methodology

3.2.1 Data Acquisition

Deep learning is most successful when large quantities of data are used for training, validating, and testing the models. The Medical Information Mart for Intensive Care (MIMIC) [151] database features many de-identified patient records from critical care environments and has been used in several significant and high-impact studies focused on developing biomedical algorithms, including [33, 97]. To train neural networks to estimate both SBP and DBP, ECG and PPG signals are required. Additionally, reference "true" values for SBP and DBP are needed, which can easily be derived from arterial blood pressure (ABP) waveforms available in the MIMIC database. As such, all records that contained ECG, PPG and ABP waveforms were obtained, resulting in a database comprised of 6,972 unique patients.

3.2.2 Data Preprocessing

Following acquisition, each record was split into 5-second segments. This segment length allows for extremely rapid BP estimation, while also providing a wide enough window to accurately calculate BPs even where the heart rate is extremely low. Segments were taken sequentially, with no overlap between segments. This ensured that each segment contained completely unique data. During the segmentation process, any segments with missing or flatlining signals were immediately discarded.

3.2.3 Data Selection

Signal quality indices (SQIs) have been developed in several previous works to assess the quality of ECG signals, using techniques including spectral analysis [152], fuzzy support vector machines [153], [150], and simple sanity checks [154].

While these works offer significant SQI tools for ECG signal assessment, they do not consider PPG signals. As such, for this work a straightforward SQI strategy is implemented, comprised of sanity checks for PPG and ECG waveforms. Heart rate (HR), beat-to-beat (BTB) intervals and waveform heights are calculated for ECG and PPG in each record. HR values derived from each signal must be equal and fall within 40-180 bpm for a record to be considered "good" by the SQI tool. This is the range which is physiologically probable for HR [154] and thus is a reasonable indicator of signal quality. The consistency of the signal is also considered by finding the maximum-to-minimum ratio for both beat-to-beat intervals and peak heights and ensuring that the maximum is no more than 50% larger than the minimum.

Records were also excluded if pulse pressure (the difference between SBP and DBP) was not between 20-60 mmHg. Pulse pressure is considered high when it is over 60 mmHg [155, 156], and is usually indicative of an immediate health problem. Meanwhile, pulse pressure is considered low beneath 40mmHg and indicates poor heart function [155, 156], so an intentionally conservative lower limit of 20 mmHg was chosen for this application given that data is acquired from critical care units.

Algorithm 1 Signal Selection Algorithm

Input: hr_ppg, hr_ecg, ppg_peak_ratio, ecg_peak_ratio, ppg_btb_ratio, ecg_btb_ratio, true_sbp, true_dbp, pulse_pressure

Output: use_record

1: **if** (hr_ppg == hr_ecg) & (hr_ppg > 40) & (hr_ppg < 180) & (ppg_peak_ratio < 1.5) &
(ecg_peak_ratio < 1.5) & (ppg_btb_ratio < 1.5) &
(ecg_btb_ratio < 1.5) & (pulse_pressure > 20
& (pulse_pressure < 60)] **then**
2: record_quality = 1
3: **else**
4: record_quality = 0
5: **end if**

In Algorithm 1, *hr_ppg* and *hr_ecg* are the HRs calculated from the PPG and ECG signal respectively. Additionally, *ppg_peak_ratio* and *ecg_peak_ratio* are the ratios of the maximum peak height to the minimum peak height for each signal, while *ppg_btb_ratio* and *ecg_btb_ratio* are the ratios of the widest to smallest BTB intervals for each signal. Each of these metrics offers a measure of signal consistency, which in turn is indicative of signal quality. Lastly, the parameter *pulse_pressure* is the pulse pressure values calculated from the ABP signal.

After assessing the suitability of all signals using Algorithm 1, the resulting data was inspected and outlier BP values were excluded. The final database contained over 200,000 records for use in training and testing of the proposed NNs.

3.2.4 Feature Extraction

This chapter describes two schemes for prediction of blood pressure. The first scheme uses the raw ECG and PPG waveforms, while the second scheme uses a small vector of features which describe the shape of the waveforms. The feature selection process is described for each scheme as follows.

Raw Waveform Scheme

In the first scheme, the input features used are the amplitudes of raw ECG and PPG waveforms over a 5-second period. As ECG and PPG waveforms were both sampled at 125Hz, the feature vector included 625 amplitude data features from both the ECG and PPG waveforms, for a total feature vector size of 1,250 amplitude data points.

Feature-Based Scheme

In the second scheme, the aim was to improve computational efficiency by greatly reducing the number of features used. In selecting which features to extract from the signals, human bias was minimised by predominantly choosing features *of* the signal, rather than features calculated *from* the signals. In other words, the majority of the chosen features aimed to describe the shape of the waveform.

For ECG, the waveforms were described by extracting all R, P, T, Q, and S-wave amplitudes from the signal, as well as the beat-to-beat (BTB) interval - that is, the time between R-waves. These features are illustrated in Fig. 3.1 below. To quantify the typical R, P, T, Q, and S-waves within the 5-second signal segment, median was used rather than mean to minimise the impacts of outliers. Mean was however used to quantify the typical BTB inerval between R-waves, as the time scale was less likely to be affected by noise than the various wave heights.

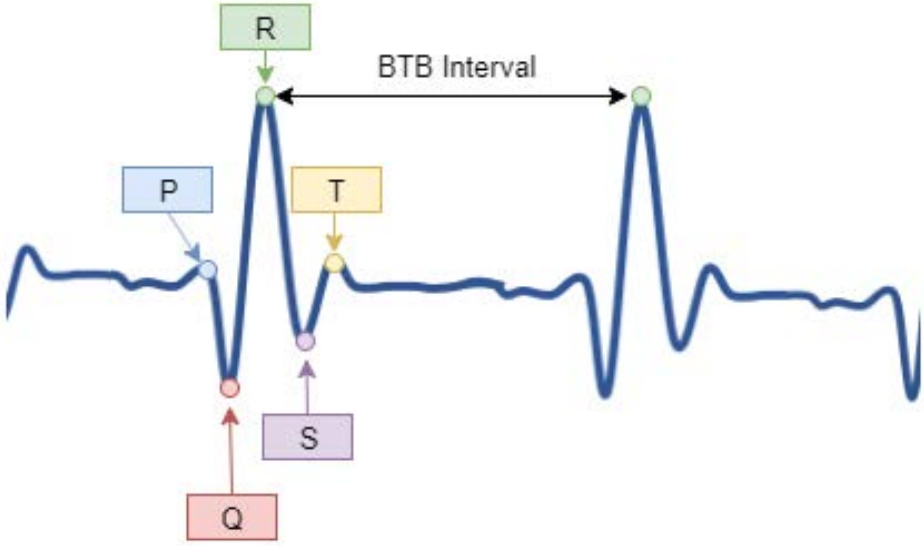


Figure 3.1: A typical ECG waveform.

The median peak and trough heights of the PPG signal were then calculated after extracting all peaks and troughs from the signal. The BTB intervals between two peaks were also extracted, with the mean then taken to quantify the typical BTB interval. These features are illustrated in the sample PPG signal shown in Fig. 3.2.

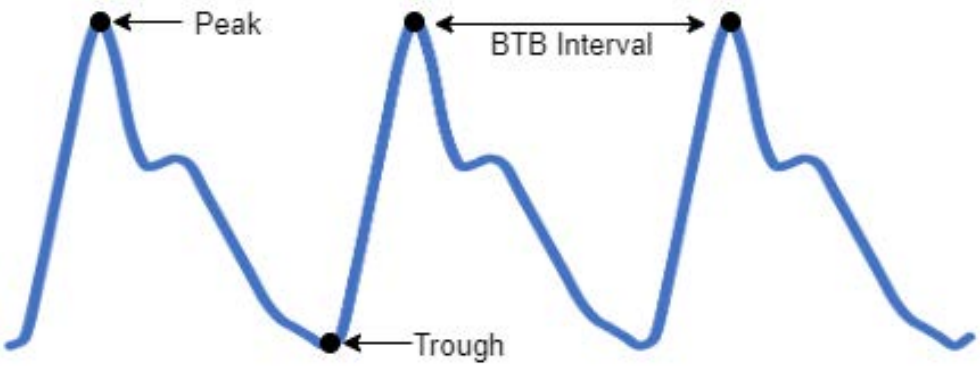


Figure 3.2: A typical PPG waveform.

The few features that required slightly more calculation were PPG wave height, heart rate, and mean up-time. As illustrated in Fig. 3.3, the PPG wave height was calculated as the difference between the first peak and trough heights, while the median of the HRs calculated from the PPG and ECG signals was used as HR in the feature vector. Lastly, up-time was calculated as the time

taken for the PPG signal to go from trough to peak. This was calculated for all trough-peak pairs and averaged to get a mean up-time, which was then used in the feature vector.

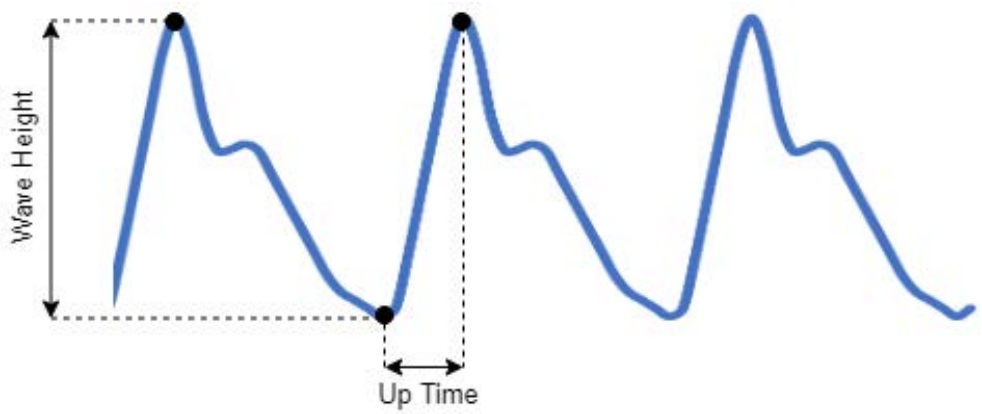


Figure 3.3: Calculation of additional PPG features.

Overall, the feature vector contained 12 important features of the ECG and PPG signals. These features required minimal calculation and thus minimize the risk of human bias impacting upon the ability of a NN to learn from the data. The twelve features used were the R-wave median, S-wave median, Q-wave median, P-wave median, T-wave median, mean BTB interval of the ECG, mean BTB interval of the PPG, PPG wave height, heart rate, median PPG peak height, median PPG trough depth, and mean up time.

3.2.5 Proposed Neural Network

Raw Waveform Scheme

For the task of blood pressure estimation from raw waveforms, a hybridised deep neural network (DNN) is proposed, combining temporal convolutional layers with long short-term memory (LSTM) layers, as shown in Fig. 3.4. CNNs are typically used to identify important features and patterns within a signal, regardless of their location, and have previously been used in the related problems in ECG anomaly detection [51–53]. Meanwhile LSTM networks perform exceptionally well on sequential data due to their ability to 'remember' what they have previously seen. This enables them to draw links between multiple features readily, and has lead to them being trialled in ECG and BP related problems

[99, 100, 148]. Combining the two network structures draws on the benefits of both to create a powerful hybrid NN with strong predictive abilities for sequential waveform data. The proposed hybrid CNN-LSTM outperformed separate CNN and LSTM networks with respect to MAE, SD, and error distribution in preliminary testing.

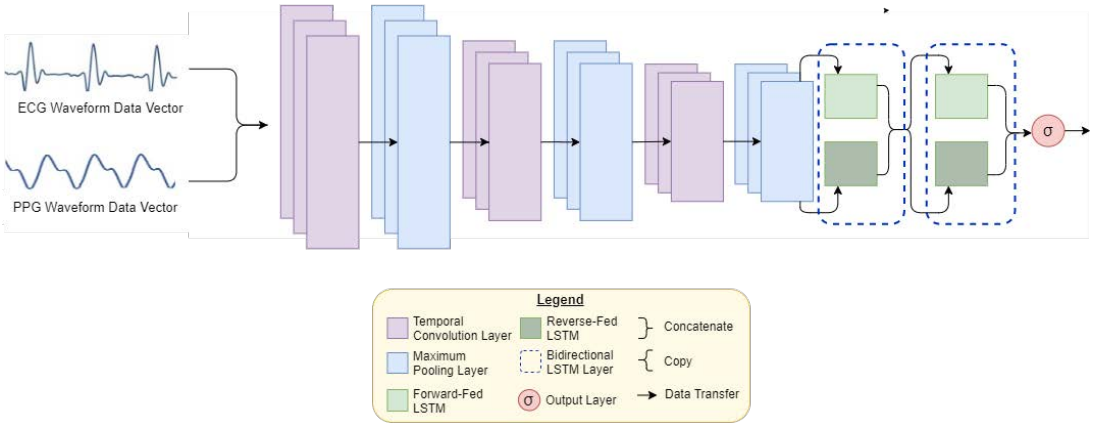


Figure 3.4: System model of the proposed NN for BP estimation using raw waveforms as inputs.

The proposed network then utilizes three temporal CNN layers, each with 128 hidden units and utilizing ReLU activation. The first two CNN layers use a kernel size of 10, while the third uses a kernel size of 4. All CNN layers use a stride of 1. The CNN layers are mathematically described by Eqn. (3.1).

$$y_{\mathrm{j}}^{\mathrm{i}} = relu(\sum_{n=1}^{N} w_{\mathrm{jn}}^{\mathrm{i}} * x_{\mathrm{m}}^{(\mathrm{i\text{-}1})} + b_{\mathrm{j}}^{\mathrm{i}}) \tag{3.1}$$

where y_j^i is the jth feature map of the ith layer. Convolution is denoted with the $*$ symbol. Weights $w_{\mathrm{jn}}^{\mathrm{i}}$ describe the nth weight of the jth feature map from the $(i-1)$th layer, where $n = 1, ..., N$. The outputs of the $(i-1)$th layer are denoted as $x_{\mathrm{m}}^{(\mathrm{i\text{-}1})}$, while bias is denoted as b_j for the jth bias term of the ith layer. Biases are initialised to zero and updated using the Adam optimizer algorithm [157] with a learning rate of 0.01.

Maximum pooling is applied following each convolutional layer, as shown in Fig. 3.4. Pool1 and Pool2 both use pool and stride sizes of 10, while Pool3 uses pool and stride sizes of 4. Applying maximum pooling after CNN downsamples the outputs, which aids in the prevention of overfitting.

As shown in Figure 3.4, the unravelled ECG and PPG data is passed to the regression network as a feature vector, before passing through an efficient network comprised of multiple interleaved convolutional, dimensionality reduction, and LSTM layers. The output of the final hidden layer is passed to a densely connected layer. This output layer utilises ReLU activation to predict blood pressure as a decimal value. The same network structure was used for training separate SBP and DBP prediction networks.

Following the convolutional section of the network, there two bidirectional LSTM network layers with 128 hidden units. Bidirectional LSTMs (BiLSTMs) consider data in both original and reversed order, allowing them to learn from values both in the past and future within the sequence. Results from both forward and reversed sequences are concatenated to provide the overall output, however the mathematical structure for both passes remains the same as standard LSTM. This mathematical process is described Eqns. (3.2-3.7) below.

$$\tilde{c}_{\mathrm{t}} = \tanh(w_{\mathrm{c}}[\, a_{(\mathrm{t\text{-}1})}, x_{\mathrm{t}}] \, + b_{\mathrm{c}}) \tag{3.2}$$

$$f_{\mathrm{t}} = \sigma(w_{\mathrm{f}}[\, a_{(\mathrm{t\text{-}1})}, x_{\mathrm{t}}] \, + b_{\mathrm{f}}) \tag{3.3}$$

$$u_{\mathrm{t}} = \sigma(w_{\mathrm{u}}[\, a_{(\mathrm{t\text{-}1})}, x_{\mathrm{t}}] \, + b_{\mathrm{u}}) \tag{3.4}$$

$$o_{\mathrm{t}} = \sigma(w_{\mathrm{o}}[\, a_{(\mathrm{t\text{-}1})}, x_{\mathrm{t}}] \, + b_{\mathrm{o}}) \tag{3.5}$$

$$c_{\mathrm{t}} = u_{\mathrm{t}} \bullet \tilde{c}_{\mathrm{t}} + f_{\mathrm{t}} \bullet c_{(\mathrm{t\text{-}1})} \tag{3.6}$$

$$a_{\mathrm{t}} = o_{\mathrm{t}} \bullet \tanh(c_{\mathrm{t}}) \tag{3.7}$$

where the weights are $\mathrm{w_c}$, $\mathrm{w_f}$, $\mathrm{w_u}$ and $\mathrm{w_o}$, while biases are $\mathrm{b_c}$, $\mathrm{b_f}$, $\mathrm{b_u}$ and $\mathrm{b_o}$. Biases and weights are learnt using the Adam optimization algorithm [157] with a learning rate of 0.01. The previous layer output is denoted as $a_{(\mathrm{t\text{-}1})}$, while x_{t} is the input to timestep t. Lastly, Equations (3.6) and (3.7) are the updated cell state and layer output respectively.

The final layer of the network is a simple densely-connected node that provides the final output of the network, which is the prediction for blood pressure. This network structure was used for training and testing of networks for SBP and DBP estimation separately.

Feature-Based Scheme

For the feature-based scheme, a shallower hybrid CNN-LSTM network was used. As shown in Figure 3.5, the 12-feature vector is used as the input to the network comprised of multiple convolutional, pooling, and LSTM hidden layers. Each layer had 128 hidden units, with the CNN layers featuring a kernel size of 2 with stride of 1 and the maximum pooling layers featuring pool and stride sizes of 2. The output of the final hidden layer is passed to a densely connected layer utilizing sigmoid activation, which then determines the most likely blood pressure.

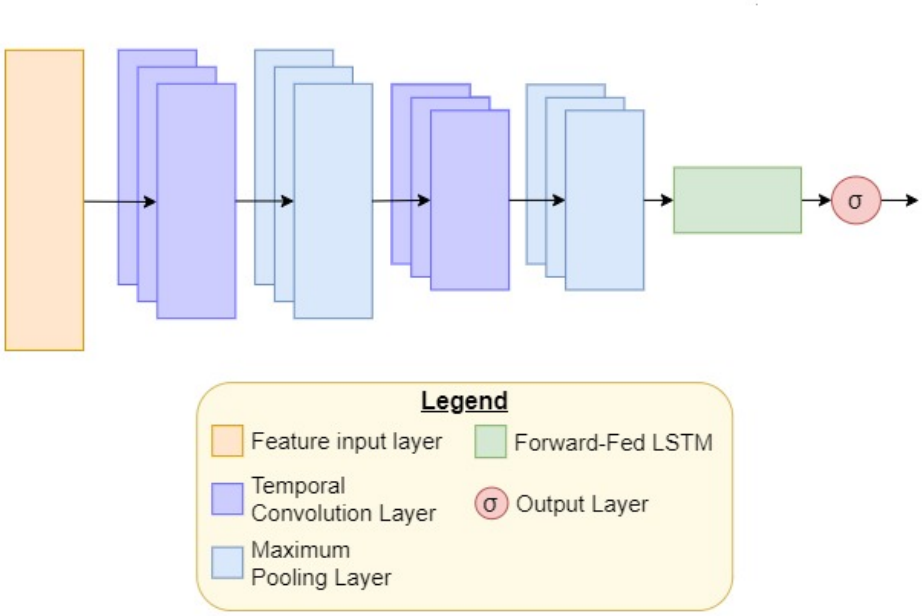


Figure 3.5: System model of the proposed NN for BP estimation using twelve features of the waveforms as inputs.

This network was able to be shallower than that of the raw waveform scheme, as fewer input features lead to convergence on a solution without as many layers. The layer depth illustrated in Fig. 3.5 was found to provide the optimal MAE, with increased depth showing no improvement with respect to this key parameter. Additionally, unidirectional LSTM was used rather than bidirectional LSTM in this case, as the latter showed no improvement in predictive performance. Therefore it was preferable to use unidirectional LSTM, which offers lower computational cost. As with the raw waveform scheme, biases are

initialised to zero and updated using Adam optimization [157] with a learning rate of 0.01.

3.2.6 Training & Testing of the NNs

The networks for both the raw waveform and feature-based scheme were trained using 80% of the data. A further 10% of the data was used for validation of SBP and DBP, allowing for fine-tuning of the hyperparameters. The final 10% of data remained unseen to the networks for use in testing.

For the raw waveform network, training and validation was performed over 750 epochs with mean absolute error (MAE) used as the loss function. Meanwhile, the feature-based network converged rapidly after 50 epochs, with any further iterations found to cause overfitting to the training set. For both schemes, network performance was checked at the end of each iteration; if the network was achieving a lower MAE than all previous iterations, then the network weights were saved. If not, then training moved on to the next iteration. This ensured that the best weight combination encountered during training and validation was used for the final network.

Testing was then conducted using the highest-performing networks for SBP and DBP estimation in each scheme. The predictions made by the SBP and DBP networks are also combined to produce a prediction for mean arterial pressure (MAP), which represents the average pressure in a person's arteries during a single cardiac cycle [82] and is mathematically defined as follows:

$$MAP = \frac{SBP + (2 \times DBP)}{3}$$

The results achieved by the networks were recorded and analysed, and are presented with discussion in the following section.

3.3 Results & Discussions

In evaluating the performance of the proposed CNN-LSTM model for the estimation of SBP, DBP, and MAP, two widely accepted standards for the approval of blood pressure devices for use in clinical environments are considered - the British

Hypertension Society (BHS) protocol and the Association for the Advancement of Medical Instrumentation (AAMI) standard.

Furthermore, this section evaluates the level of agreement between the calculations made by the CNN-LSTM networks and the expected SBP, DBP, and MAP values as determined from ABP waveforms within the MIMIC-III database, which were obtained using gold-standard intra-arterial blood pressure measurement. The schemes proposed in this chapter are also compared to previous related works, highlighting the improvement that the proposed schemes make to accurate blood pressure estimation.

3.3.1 Comparison to the BHS Protocol

The BHS protocol [103] assigns grades of A-D to blood pressure measurement devices, based on the percentages of measurements that achieve absolute differences of less than 5mmHg, 10mmHg, and 15mmHg respectively, when compared to gold-standard measurement techniques such as intra-arterial monitoring. The grading criteria established by the BHS protocol are illustrated in Table 2.1. Devices that achieve grades of A or B in accordance with the BHS grading criteria are considered suitable for clinical use, while those that achieve lower grades are not recommended for clinical use.

Raw Waveform Scheme

As shown in Table 3.1, the proposed raw waveform scheme satisfies the requirements for an A grade device in the estimation of SBP, DBP, and MAP, and thus would be recommended for use in clinical settings.

Table 3.1: Assessment of raw waveform scheme based on BHS protocol.

	Absolute Difference (mmHg)			
	≤5	**≤10**	**≤15**	**Grade**
SBP	67.66%	89.82%	96.82%	A
DBP	82.79%	96.12%	99.09%	A
MAP	84.21%	97.38%	99.58%	A

Feature-Based Scheme

As shown in Table 3.2, the feature-based scheme also comfortably satisfies the requirements for an A grade device in the estimation of SBP, DBP, and MAP. It can be seen that fewer predictions fell within each error category than with the raw waveform scheme, however the maximum grade for the BHS standard is still comfortably achieved.

Table 3.2: Assessment of feature-based scheme based on BHS protocol.

	Absolute Difference (mmHg)			
	≤5	**≤10**	**≤15**	**Grade**
SBP	64.96%	89.90%	97.76%	A
DBP	82.08%	96.18%	98.86%	A
MAP	81.04%	96.91%	99.50%	A

3.3.2 Comparison to the AAMI Protocol

Blood pressure devices are often evaluated with respect to both the AAMI and BHS protocols, as they have different mechanisms for determining device suitability. The AAMI standard [102] states that a device must have a mean difference of ≤5 mmHg and a standard deviation (SD) of ≤8 mmHg from gold standard measurements. Devices are assigned a grade of "Pass" if the aforementioned criteria are met, otherwise the device is given a grade of "Fail". Each of the proposed schemes were compared to the AAMI standard as follows.

Raw Waveform Scheme

As illustrated in Table 3.3, the raw waveform scheme comfortably achieve "Pass" grades with respect to the AAMI criteria. The proposed algorithms achieve impressively low MAEs and SD in estimation of all BP parameters, and would be suitable for implementation in healthcare.

Feature-Based Scheme

After testing the proposed CNN-LSTM model on the test set, it was found that the algorithm achieved acceptably low MAE and SD in estimating SBP, DBP, and

Table 3.3: Assessment of raw waveform scheme based on AAMI standard.

	MAE (mmHg)	**SD (mmHg)**	**Grade**
SBP	4.4097	6.1075	Pass
DBP	2.9105	4.2347	Pass
MAP	2.7663	3.8832	Pass

MAP, as is illustrated in 3.4. This model therefore achieves a comfortable "pass" grade in all areas of BP estimation according to the AAMI standard. While MAE is higher than was achieved by the raw waveform scheme, this scheme would still be suitable for healthcare applications.

Table 3.4: Assessment of feature-based scheme based on AAMI standard.

	MAE (mmHg)	**SD (mmHg)**	**Grade**
SBP	4.5010	5.9678	Pass
DBP	3.0167	4.2987	Pass
MAP	3.0517	4.1357	Pass

3.3.3 Analysis of Error Distribution

Accurate measurement of BP is of vital importance in healthcare applications. To further analyse the performance of the proposed schemes, error histograms were generated for SBP, DBP and MAP prediction to inspect the spread of errors.

Raw Waveform Scheme

Figs. 3.6-3.8 present the error distributions for the SBP, DBP, and MAP predictions generated by the raw waveform scheme. Each of these histograms clearly shows that '0' is the most common error, and that most other errors are also extremely low.

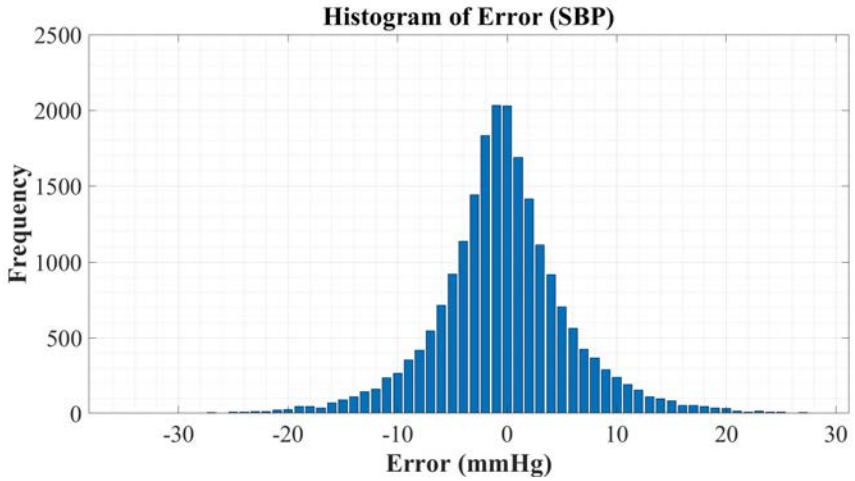


Figure 3.6: Error histogram for SBP (raw waveform scheme).

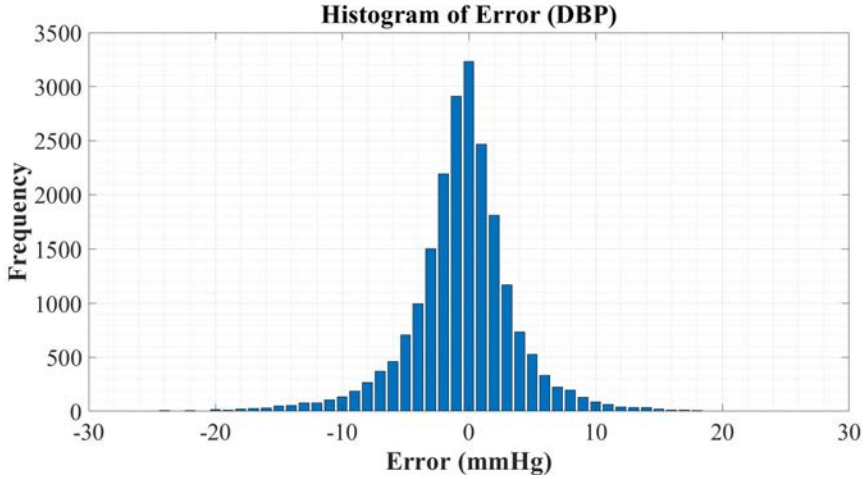


Figure 3.7: Error histogram for DBP (raw waveform scheme).

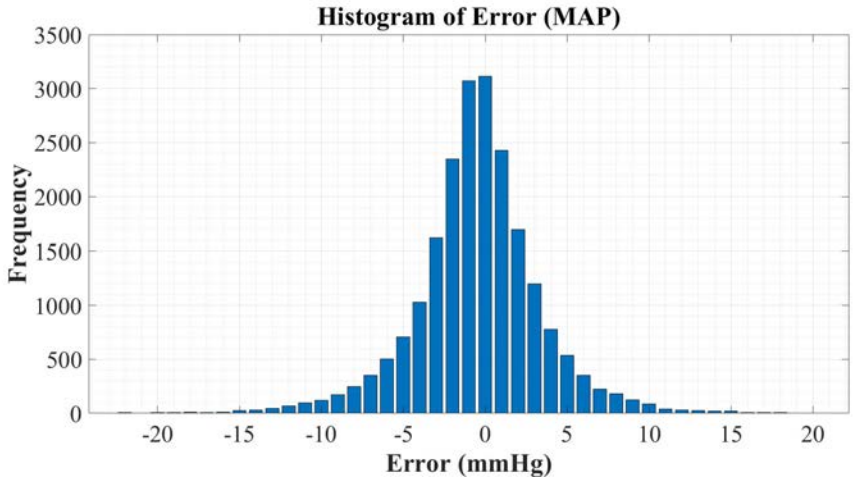


Figure 3.8: Error histogram for MAP (raw waveform scheme).

Feature-Based Scheme

The error distributions for SBP, DBP, and MAP predictions made by the feature-based scheme are illustrated in Figs. 3.9-3.11. Each of these histograms clearly shows that '0' is the most common error, and that most other errors are also extremely low. The error histograms for the feature-based scheme show a similar distribution of error as was achieved by the raw waveform scheme.

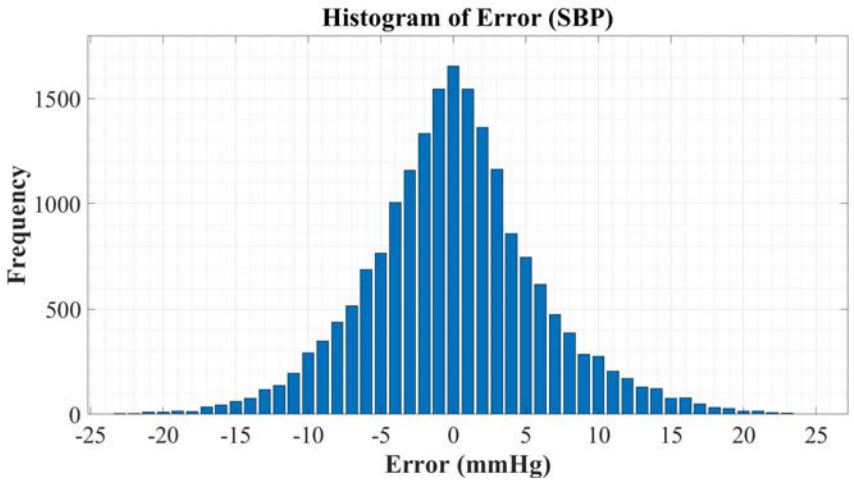


Figure 3.9: Error histogram for SBP (feature-based scheme).

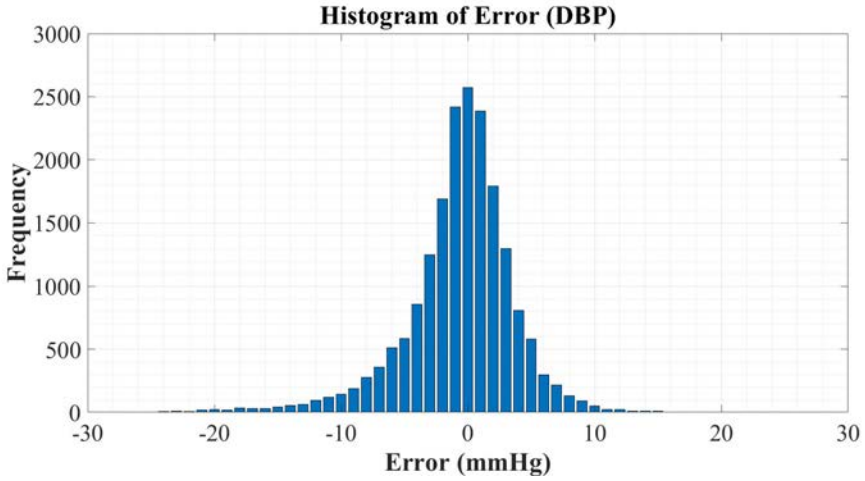


Figure 3.10: Error histogram for DBP (feature-based scheme).

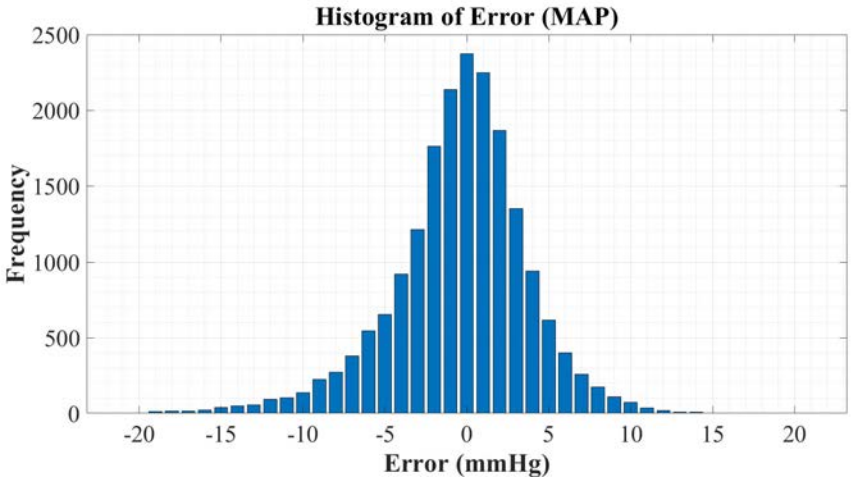


Figure 3.11: Error histogram for MAP (feature-based scheme).

3.3.4 Level of Agreement between Intra-arterial Monitoring and CNN-LSTM Networks

Bland Altman plots are a key method for assessing the level of agreement between two methods of measurement, particularly in medical applications. These plots illustrate the difference between two measurements compared to the mean of the two measurements, and as such a high density of data near the central 'mean difference' line and narrow 'limits of agreement' (LOAs) indicate a strong level of agreement between measurements.

Raw Waveform Scheme

In Figs. 3.12-3.14, the SBP, DBP, and MAP predictions made by the raw waveform scheme are compared with the values obtained using the current gold-standard of BP monitoring, intra-arterial measurement. Each figure clearly shows a high density of points near the mean difference line with narrow LOAs for SBP, DBP, and MAP graph. As such, it is clear that there is a high level of agreement between the proposed CNN-LSTM models the respective measurements made via intra-arterial monitoring.

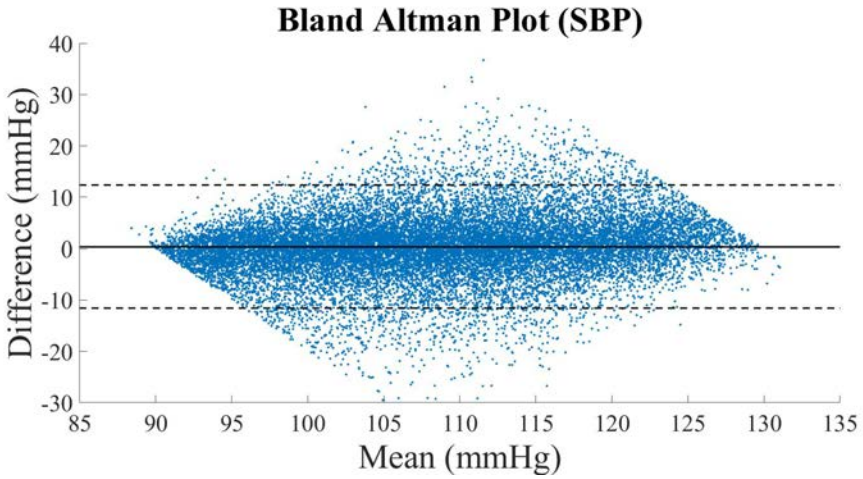


Figure 3.12: Bland Altman plot for SBP (raw waveform scheme).

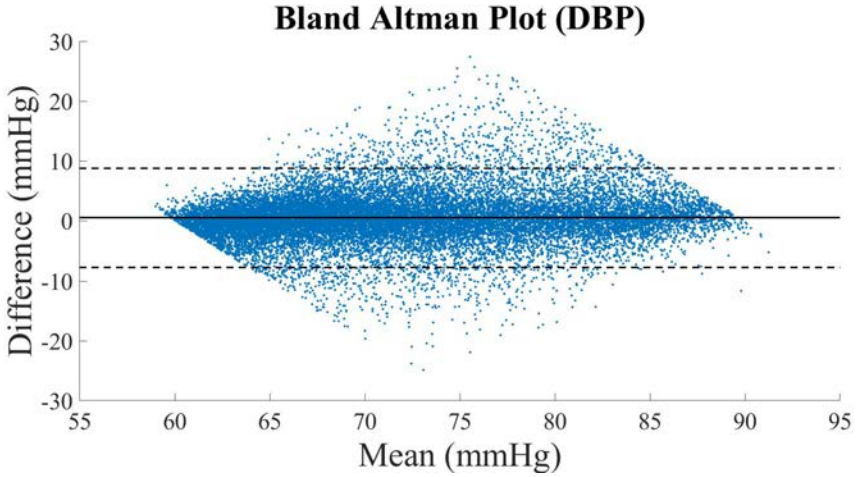


Figure 3.13: Bland Altman plot for DBP (raw waveform scheme).

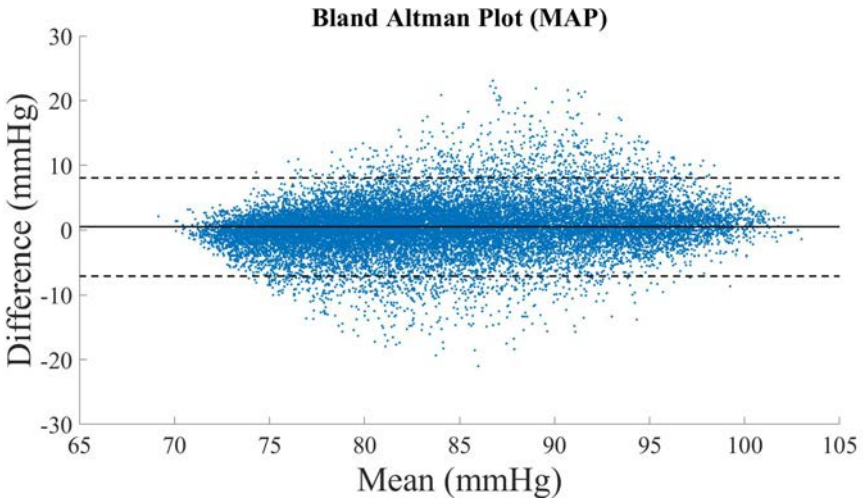


Figure 3.14: Bland Altman plot for MAP (raw waveform scheme).

To further evaluate the level of agreement between the proposed scheme and intra-arterial BP measurement, regression plots were generated and the coefficients of correlation were calculated to quantify the strength of the relationship between measurements. In all regression plots, the dashed black line shows the theoretical "perfect" correlation, while the solid black line represents the actual correlation. The regression plots for SBP, DBP, and MAP are shown in Figs. 3.15, 3.16 and 3.17 respectively.

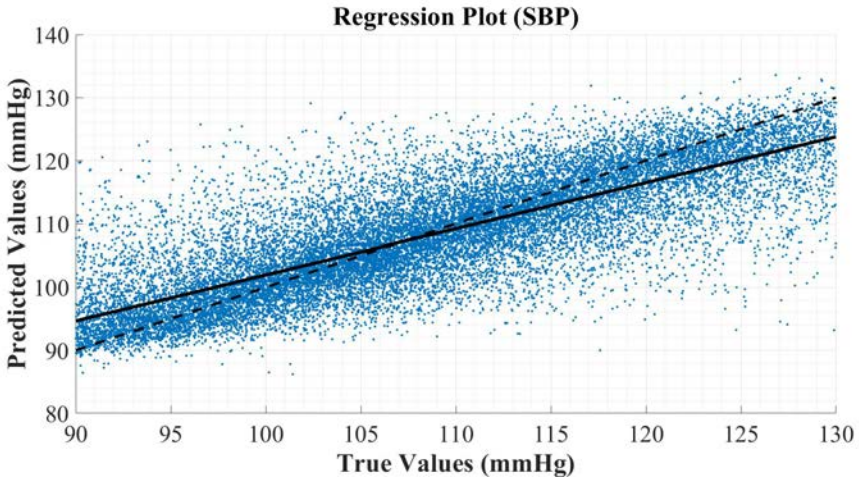


Figure 3.15: Regression plot for SBP (raw waveform scheme).

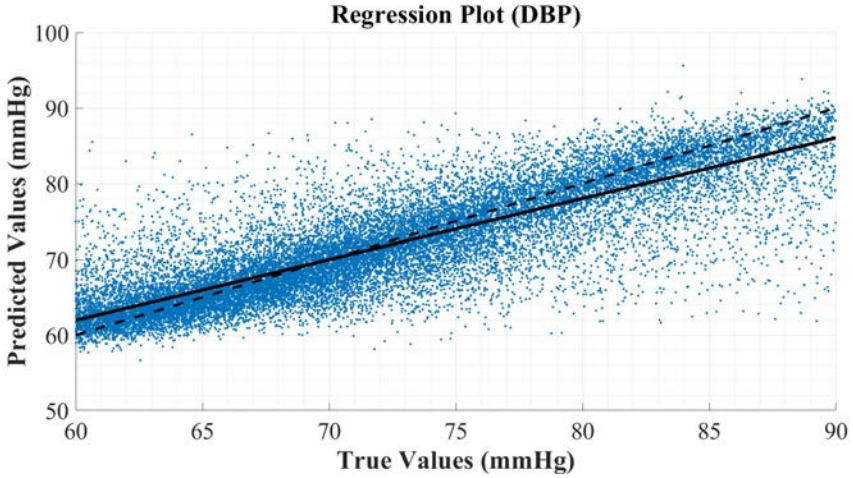


Figure 3.16: Regression plot for DBP (raw waveform scheme).

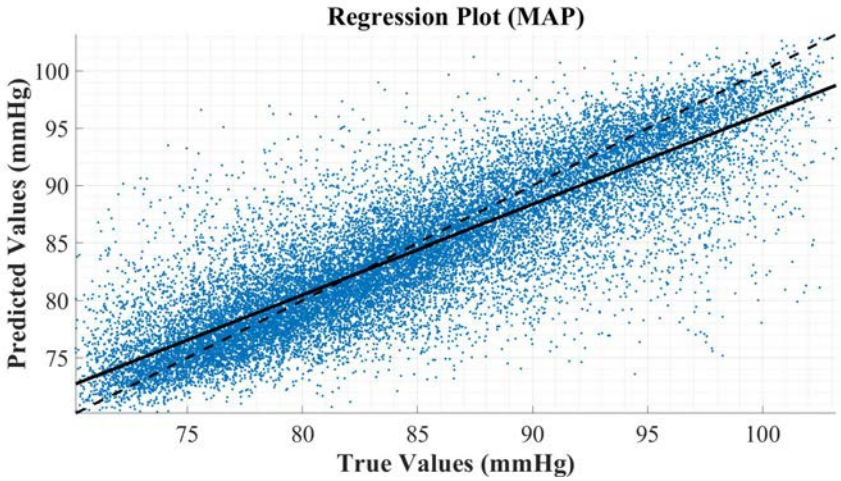


Figure 3.17: Regression plot for MAP (raw waveform scheme).

Each regression plot illustrates strong positive linear correlation between the "true" values and the actual predictions generated by the raw waveform CNN-LSTM model. The calculated correlation lines fall close to ideal correlation lines in all cases. To further analyse correlation, the correlation coefficients were calculated, with the results displayed in Table 3.5.

Table 3.5: Coefficients of correlation for the raw waveform scheme.

Blood Pressure Parameter	**Coefficient of Correlation**
SBP	0.8008
DBP	0.8482
MAP	0.8597

These results clearly confirm the strong positive linear relationships between the predictions for SBP, DBP, and MAP made by the proposed scheme when compared with the respective measurements acquired with invasive intra-arterial monitoring.

Overall, it is evident that there is a high level of agreement and strong correlation between the raw waveform scheme and the current gold-standard for blood pressure estimation. As the proposed scheme is entirely non-invasive, unlike intra-arterial monitoring, these results are extremely promising for the future of healthcare, especially for at-risk patients such as premature babies and the elderly.

Feature-Based Scheme

Bland-Altman analysis was also performed for the feature-based scheme, with Figs. 3.18-3.20 showing the Bland Altman plot for SBP, DBP, and MAP. As shown in these figure, there is a high level of agreement between the values calculated by the feature-based CNN-LSTM model and that of the intra-arterial monitoring, with narrow LOAs and clustering of data points around the mean difference line visible in all graphs.

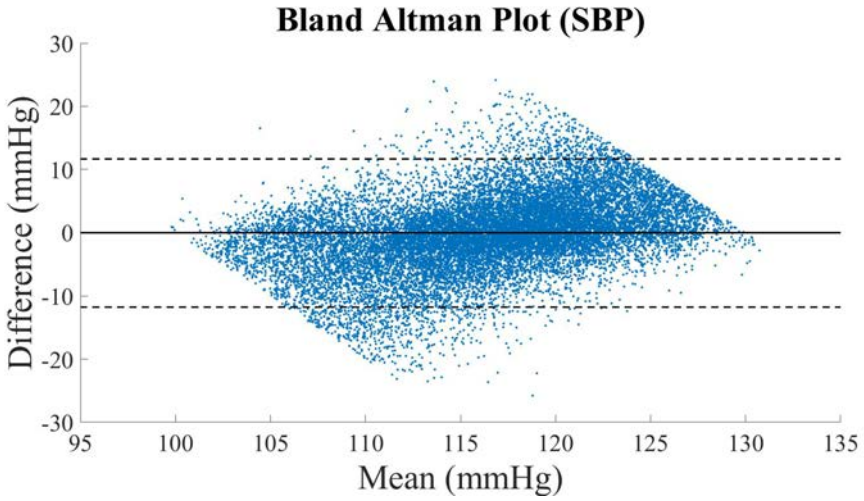


Figure 3.18: Bland Altman plot for SBP (feature-based scheme).

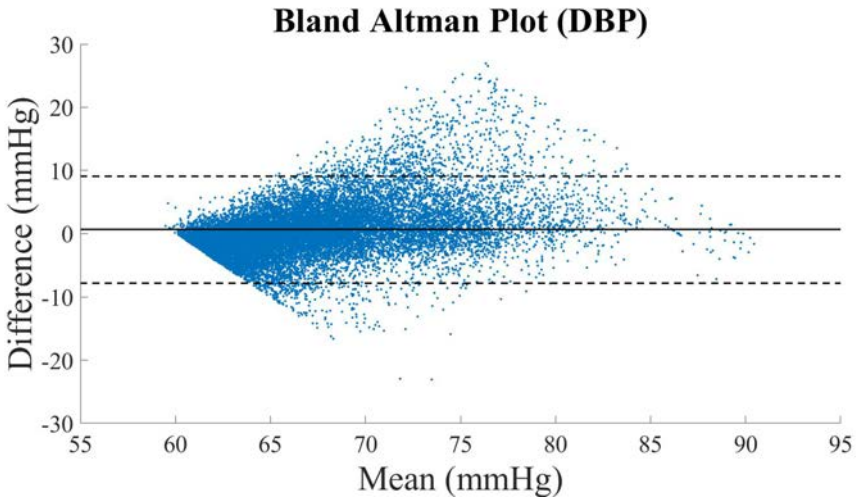


Figure 3.19: Bland Altman plot for DBP (feature-based scheme).

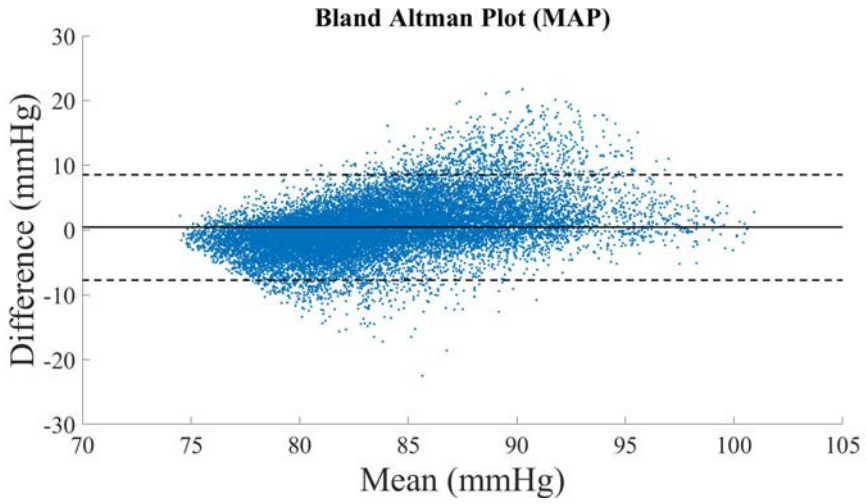


Figure 3.20: Bland Altman plot for MAP (feature-based scheme).

To further evaluate the level of agreement between intra-arterial BP and the feature-based scheme, regression plots were generated and the coefficients of correlation were calculated. In Figs. 3.21-3.23, the solid black line represents the line of best fit to the data, while the dashed black line illustrates what the "ideal" linear regression would have been. A reasonable level of agreement is seen between gold-standard intra-arterial monitoring and the feature-based scheme in each figure, however this level of agreement is not as strong as was seen between intra-arterial monitoring and the raw waveform scheme.

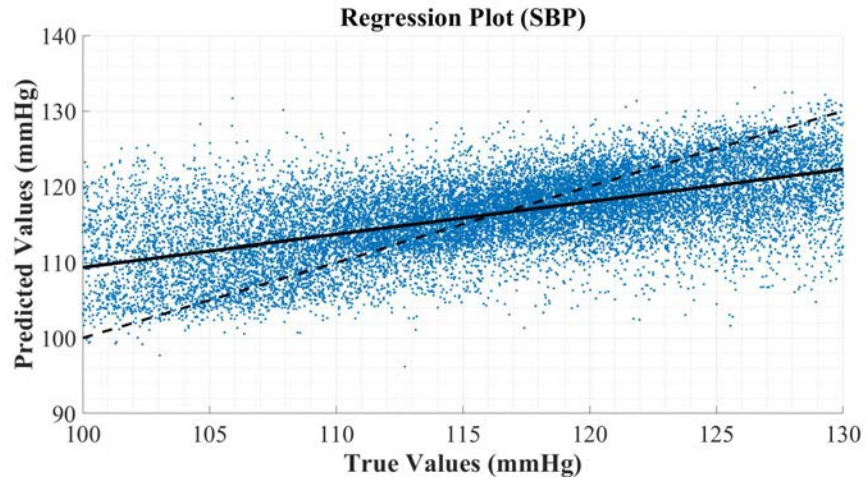


Figure 3.21: Regression plot for SBP (feature-based scheme).

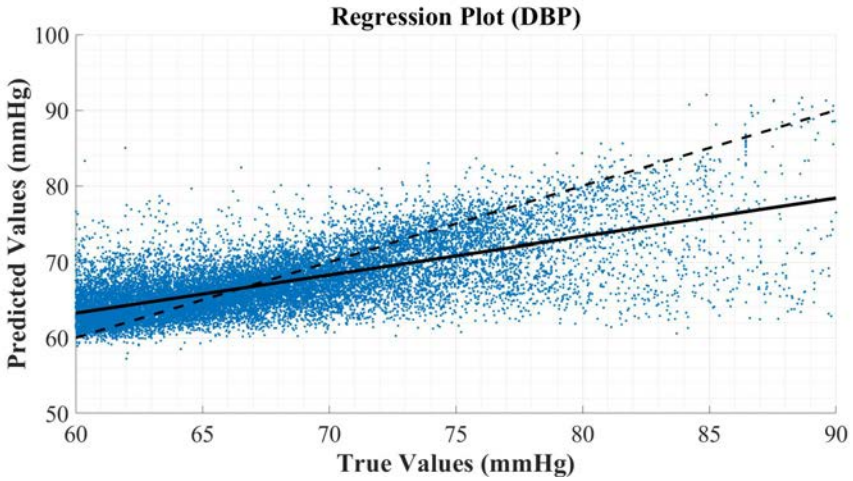


Figure 3.22: Regression plot for DBP (feature-based scheme).

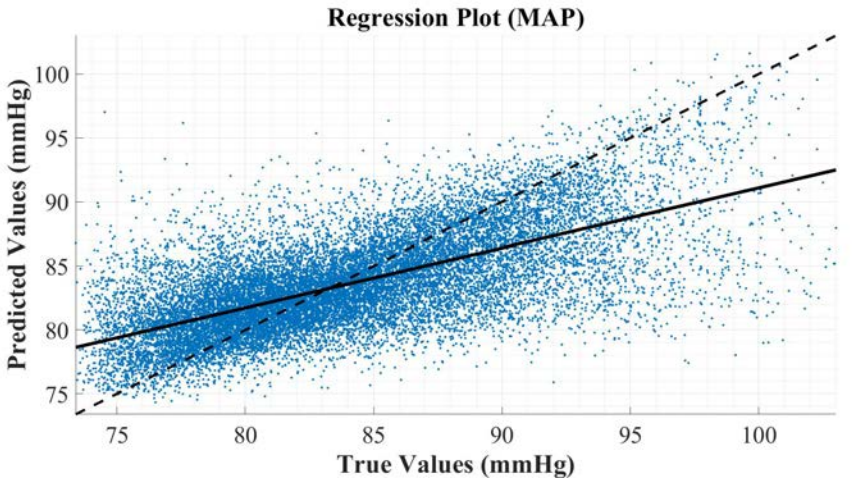


Figure 3.23: Regression plot for MAP (feature-based scheme).

The correlation coefficients are shown in Fig. 3.6, with all confirming the positive linear relationship between intra-arterial measurements and predictions made by the proposed CNN-LSTM networks in all categories of BP measurement. In each case, the correlation coefficient was lower than those achieved by the raw waveform scheme, but still reasonably strong.

Table 3.6: Coefficients of correlation for the feature-based scheme.

Blood Pressure Parameter	**Coefficient of Correlation**
SBP	0.6085
DBP	0.6808
MAP	0.6686

3.3.5 Comparison to Previous Works

Several works that are strongly related to this work are presented in [33, 34, 36–38, 98]. Each of these methods utilises PPG and ECG signals to predict blood pressure, though preprocessing and segment lengths vary. Several of these works utilise the MIMIC-II database [34, 38, 158] or MIMIC-III database [36], while others utilise small databases that were independently acquired [37, 98]. In this section, the results obtained by the NNs proposed in this chapter are compared with the best results achieved by the previous literature. The work in [36] presented results from several databases, and as such the results that they obtained using MIMIC-III data are used in comparisons, as these results are the most directly comparable to this work.

In Table 3.7, the models developed in previous works are compared to the proposed CNN-LSTM networks with respect to the BHS protocol. In terms of SBP, this table shows that the proposed raw waveform and feature-based schemes outperform the previous state-of-the-art works. The proposed schemes are the only ones to have achieved an 'A' grade for SBP estimation across all data analysed. This comparison also shows that the raw waveform scheme performs marginally better than the feature-based scheme in terms of the BHS protocol criteria, achieving more predictions under each error threshold.

Table 3.7: Comparison of schemes based on the BHS protocol.

		Absolute Difference (mmHg)			
		≤5	**≤10**	**≤15**	**Grade**
Kachuee [33]	SBP	34.1%	56.5%	72.7%	D
	DBP	62.7%	87.1%	95.7%	A
	MAP	54.2%	81.8%	93.1%	B
Mousavi [34]	SBP	71%	77%	84%	C
	DBP	84%	92%	97%	A
	MAP	79%	83%	93%	B
Miao [98]	SBP	51%	81%	94%	B
	DBP	62 %	92%	99%	A

Continued on next page

Table 3.7 – continued from previous page

		Absolute Difference (mmHg)			
		≤5	**≤10**	**≤15**	**Grade**
	MAP	60%	90%	98%	A
Song [37]	SBP	N/A	N/A	N/A	B
	DBP	N/A	N/A	N/A	A
	MAP	-	-	-	-
Miao [36]	SBP	50.07%	76.41%	90.39%	B
	DBP	65.66%	89.77%	96.63%	A
	MAP	65.14%	89.58%	96.61%	A
Raw Waveform Scheme	SBP	67.66%	89.82%	96.82%	A
	DBP	82.79%	96.12%	99.09%	A
	MAP	84.21%	97.38%	99.58%	A
Feature-Based Scheme	SBP	64.96 %	89.90%	97.76%	A
	DBP	82.08%	96.18%	98.86 %	A
	MAP	81.04%	96.91%	99.50%	A

The proposed schemes also perform extremely well for DBP estimation when compared to the previous works. As shown in Table 3.7, the raw waveform scheme again performs slightly better than the feature-based scheme. However, both schemes definitively outperform the DBP prediction networks presented in other works. While the work in [34] achieved a higher percentage of results with ≤5 mmHg error, the proposed schemes each had a higher percentage of errors ≤10 mmHg and ≤15 mmHg. This indicates that the work in [34] has overfit to the data, while the networks proposed by this chapter have not suffered from overfitting, and therefore generate fewer extremely high errors. While all schemes for DBP estimation achieved grades of 'A' and therefore could be recommended for use by the BHS for this particular parameter, the raw waveform scheme has achieved the lowest number errors greater than 10 mmHg and thus would be considered the most suitable for clinical use.

MAP prediction is also considered in Table 3.7, which compares the schemes proposed in this chapter with previous works. No results for MAP were presented by [37], but this important diagnostics parameter was examined in [33, 34, 36,

98]. As shown in Table 3.7, both of the proposed schemes for MAP estimation outperform the previous works. With 99.19% of errors falling under 15 mmHg for the raw waveform scheme, it is clear that this scheme has very few high-range errors when compared to previous state-of-the-art works. Only the schemes proposed by this chapter and the work presented in [36, 98] achieve the grade of 'A', however the schemes proposed here each had significantly fewer high-range errors exceeding 5 mmHg and thus would be the most suitable for clinical use. Once again, the raw waveform scheme slightly outperforms the feature-based scheme.

As shown in Table 3.8, the proposed schemes also compare favourably with previous works with respect to the AAMI standards. In all cases, the grade was determined based on MAE and standard deviation (SD). Table 3.8 shows that the schemes proposed in this chapter and the scheme presented in [37] are the only ones to achieve a grade of "Pass" for SBP estimation. However, the raw waveform scheme has a lower MAE and SD than that of [37], while the feature-based scheme achieves results comparable to those presented in [37]. The proposed schemes each have a marginally higher MAE and SD than the scheme in [34] for DBP and MAP, however this is likely due to the overfitting seen in Table 3.7 for [34].

Table 3.8: Comparison of schemes based on the AAMI standard.

		Error Metrics (mmHg)		
		MAE	**SD**	**Grade**
Kachuee [33]	SBP	11.80	9.88	Fail
	DBP	5.83	5.71	Fail
	MAP	5.92	5.25	Fail
Mousavi [34]	SBP	3.97	8.901	Fail
	DBP	2.43	4.173	Pass
	MAP	2.61	4.911	Pass
Miao [98]	SBP	6.13	7.76	Fail
	DBP	4.54	5.52	Pass
	MAP	4.81	6.03	Pass

Continued on next page

Table 3.8 – continued from previous page

		Error Metrics (mmHg)		
		MAE	**SD**	**Grade**
Song [37]	SBP	4.8	6.0	Pass
	DBP	4.8	6.0	Pass
	MAP	N/A	N/A	N/A
Sharifi [38]	SBP	7.83	9.1	Fail
	DBP	4.86	5.21	Pass
	MAP	N/A	N/A	N/A
Miao [36]	SBP	7.10	9.99	Fail
	DBP	4.61	6.29	Pass
	MAP	4.66	6.36	Pass
Raw Waveform Scheme	SBP	4.4097	6.1075	Pass
	DBP	2.9105	4.2347	Pass
	MAP	2.7663	3.8832	Pass
Feature-Based Scheme	SBP	4.5010	5.9678	Pass
	DBP	3.0167	4.2987	Pass
	MAP	3.0517	4.1357	Pass

Overall, the proposed raw waveform and feature-based schemes for the prediction of SBP, DBP, and MAP both perform strongly and have been shown to generalize well to new data. For SBP, DBP and MAP, the proposed models both achieved 'A' and "Pass" grades for the BHS and AAMI standard respectively. The schemes proposed in this chapter were the only schemes to achieve these high results for all three BP metrics.

Additionally, the schemes proposed by this chapter were the only schemes to achieve both 'A' and "Pass" grades for SBP measurement. The raw waveform scheme exhibited superior predictive performance when compared to the feature-based scheme, achieving lower MAE for SBP, DBP, and MAP, as well as stronger results with respect to the BHS criteria.

A strong level of agreement with gold-standard measurements was achieved by both schemes, as shown by the Bland-Altman and regression plots. However, the regression plots indicate stronger positive correlation between intra-arterial

monitoring and raw waveform scheme predictions than was seen when comparing intra-arterial monitoring with the feature-based scheme. These results suggest that both of the proposed schemes are highly suitable options for non-invasive BP estimation from ECG and/or PPG waveforms, with the raw waveform scheme achieving the best predictive performance and level of agreement with intra-arterial monitoring.

3.3.6 Computational Efficiency

To compare the computational efficiency of the two schemes proposed in this chapter, speed testing was conducted on a NVIDIA GeForce GTX 1070 graphics card. The time taken for the model in the feature-based scheme to make a single prediction was 3.86×10^{-6} seconds, while the raw waveform model took 1.20×10^{-4} seconds to make a prediction. This means that the proposed feature-based scheme is capable of making predictions $31\times$ faster than the raw waveform scheme. While the higher performance of the raw waveform scheme would make it more suitable for clinical applications, the significantly improved computational efficiency with slight reduction in performance would make the proposed feature-based scheme more suitable for wearable devices.

3.4 Conclusion

In this work, two schemes based on hybridized CNN-LSTM neural networks are proposed for the estimation of blood pressure. The first scheme utilized raw ECG and PPG waveforms as inputs, showing that with minimal data pre-preprocessing, accurate estimations of SBP, DBP, and MAP can be achieved with the CNN-LSTM network structure. In the second scheme, 12 straightforward features describing the shape of ECG and PPG waveforms are extracted and used as inputs to a shallower CNN-LSTM neural network, with strong performance again seen for the prediction of SBP, DBP, and MAP.

When compared to standards set by the reputable healthcare bodies of BHS and AAMI, the proposed networks performed extremely well. The two schemes each met the requirements set by AAMI and achieved grades of ‘A’ in accordance

with the BHS protocol for SBP, DBP, and MAP. This success indicates that a device implementing either of the proposed schemes would be recommended for clinical use by these professional bodies.

Furthermore, when the proposed models are compared to previous state-of-the-art schemes in the literature, the proposed schemes outperformed each previous work. Previous schemes performed well in measurement of certain BP parameters, but not in others. Additionally, overfitting was apparent in some schemes. Meanwhile, the schemes proposed in this chapter are shown to have fit the data extremely well through their high performance on previously unseen data. Additionally, the proposed schemes were the only ones to achieve grades of 'A' for the BHS protocol and 'pass' for the AAMI standard across all BP measurements.

When comparing the two proposed schemes directly, it was found that the raw waveform scheme showed stronger performance by the AAMI and BHS criteria than the feature-based scheme, as well as demonstrating a higher level of agreement with the gold-standard intra-arterial monitoring. However, the feature-based scheme still met the requirements of each standard and was found to be capable of making a prediction 31x faster than the raw waveform scheme. Overall, the raw waveform scheme would be more suitable where computational power is readily available, while the feature-based scheme would be suitable for wearable devices as it offers lower-power computation without significant compromise on predictive performance.

Overall, the performance of the proposed algorithms indicate that hybrid CNN-LSTM networks are highly suitable for blood pressure prediction. Implementing the proposed algorithms into appropriate healthcare monitoring devices would likely result in non-invasive, continuous, and highly accurate blood pressure measurements for a number of applications, from intensive care to smart watches.

The proposed algorithm addresses the first research problem of blood pressure monitoring (presented in Section 1.2.1) and leads to the first two original contributions presented in Section 1.3. This chapter offers substantial contribution to the literature through accurate schemes for both low-power wearables

and higher-powered devices such as medical equipment and computers. The high performance of CNN-LSTM networks in this chapter lead to the exploration of LSTM and CNN-LSTM networks in the remaining research chapters of this thesis.

Chapter 4

Machine Learning Approach to Calculating Respiratory Rate from Heart Rate Variations

This chapter contains materials included in the following manuscript, which has been submitted to *PLOS One*

[3] **S. Baker**, W. Xiang, and I. Atkinson, "Determining respiratory rate from photoplethysmogram and electrocardiogram signals using respiratory quality indices and neural networks," *PLoS One*, vol. 16, no. 4, p. e0249843, February 2021.

4.1 Introduction

Respiratory rate (RR) is a fundamental physiological parameter, and abnormality in this vital sign is one of the earliest indicators of critical illness. One recent study found that elevated respiratory rate was a key predictor of clinical deterioration within 48 hours of discharge from the emergency department (ED) [104]. Another classical study determined that the occurrence of at least one RR $\geq$ 27 breaths per minute (BrPM) in a 72 hour period was a strong predictor of cardiac arrest [105]. Elevated RR has also been linked to increased mortality [106], while relative changes in RR have been shown to indicate patient stability [107]. In children, elevated RR is a primary indicator of pneumonia, an infection that is the most common cause of death in children aged 0-5 [13, 14]. Clearly,

abnormalities or variations in the RR are key indicators of clinical deterioration.

Despite the clinical significance of RR, several studies have noted that it is historically less recorded than other vital signs [15, 104, 108, 109]. This has somewhat improved with the introduction of the Modified Early Warning Score (MEWS) [108], which incorporates measurement of RR. However, one study observed that nurses still don't measure RR in 50% of cases [15]. Time constraints and the lack of equipment for measuring RR were both cited as reasons for not monitoring this parameter.

This lack of recording can be partially attributed to the fact that there is a lack of tools available for automatically measuring RR. Currently, most common methods for automatic RR measurement rely on oronasal systems incorporating sensors including capnography, temperature, and moisture sensors [14]. However, these have not been widely adopted, with issues related to cost, wearability, and accuracy identified for existing automated devices [14].

Manual measurement remains the accepted method for determining RR. To obtain RR, it is recommended that healthcare staff count the number of breaths a patient takes over a one-minute period [13]. However, several studies have found that both doctors and nurses estimate respiratory rate over shorter time periods, or without counting the breath at all [16, 110]. Accuracy of manual RR calculations can be affected by patient awareness [15], as well as time constraints, interruptions from patients and other staff, and patient agitation [14, 110].

In addition to the complications associated with obtaining an accurate manual RR measurement, there is also a significant time cost. One study found that as much as 7.2% of nurses' time was spent performing patient assessment, including measurement of RR [17]. There are approximately 3 million registered nurses in America, earning an average of $75,510 USD per annum each as of May 2018 [159]. Thus, the total financial cost incurred by time nurses spend on patient assessment exceeds 16 billion USD per year.

Given the major limitations in measuring RR, it is clear that a reliable method of automatic and continuous monitoring of this vital sign in a non-invasive manner would significantly improve patient outcomes in hospitals. Additionally, given the usefulness of RR as an early indicator of critical illness, continuous

at-home measurement of RR could be lifesaving for at-risk patients living alone.

Several recent studies have investigated the use of photoplethysmogram (PPG) and echocardiogram (ECG) signals to derive RR in a wearable and non-invasive manner [39, 40, 42–45]. Respiration modulates the ECG and PPG signals in three main ways - baseline wander (BW) modulation, amplitude modulation (AM) and respiratory sinus arrhythmia (RSA) modulation, more commonly known as frequency modulation (FM). These modulations are caused by movement associated with breathing, and various responses to the change in intrathoracic pressure during respiration [160].

In order to accurately estimate RR, several recent studies have developed respiratory quality indices (RQIs) to determine which of the extracted modulations are of the highest quality [43, 44, 117]. This in turn allows for identification of which modulation-extracted RRs are realistic, thus allowing for more accurate estimation of actual RR.

Interestingly, there are very few studies that have attempted to estimate RR from PPG and ECG using machine learning (ML). The best performing ML-enabled technique was presented in [43], where a mean absolute error (MAE) of 0.71 BrPM was achieved using linear regression on a small database. While these are good results, this chapter demonstrates that they can be improved upon by instead using neural networks (NNs) in combination with the proposed novel RQI scheme.

In this chapter, an RQI scheme is developed for assessing the quality of modulation-extraction respiration signals. The proposed scheme uses statistics regarding the signal variation to assign 'good' or 'bad' ratings to RRs calculated from modulation-extracted signals. Bidirectional long short-term memory (BiLSTM) neural networks are trained and tested, comparing the performance in two scenarios: one where only RR features are used as features, and the other where both RR and corresponding RQIs are used.

The remainder of this chapter is structured as follows. Section 4.2. describes the methodology utilized for obtaining signal quality and an overall RR estimation using various NN structures. Section 4.3. presents results and discussion before Section 4.4. concludes the chapter and provides recommendations for

future work.

4.2 Methodology

4.2.1 Obtaining Data

Data for this work was obtained from the open-source Medical Information Mart for Intensive Care (MIMIC-III) database [151], which features an extremely large number of records from intensive care units (ICUs). To train the neural networks, ECG and PPG signals were needed to derive RR from the BW, AM, and FM modulations. Additionally, a reference "true" RR signal was needed to provide the neural networks with an expected output RR. As such, the PhysioBank ATM tool [161] was used to obtain a list of all records containing ECG, PPG, and respiratory waveforms from the MIMIC-III database. Then, a Python script was developed to download all relevant records as MATLAB-compatible files, utilizing several functions from the Waveform Database (WFDB) Toolbox [162]. After running this script, a total of 8,781 records were obtained. No exclusions were made based on patient demographics, diagnoses or treatments received, as the goal of this chapter was to develop an all-inclusive scheme that could measure respiratory rate irrespective of whether respiration was being affected by health conditions or respiratory support treatments.

4.2.2 Preprocessing Data

The primary preprocessing performed was the denoising of ECG and PPG signals. Many of the ECG and PPG signals were affected by baseline wander that could be attributed both to respiration and other movement. This BW prevents accurate derivation of respiration from amplitude and frequency modulations, and also inhibits signal quality assessment. To eliminate all BW from each individual ECG signal, a sixth-order polynomial was fitted to the ECG signal and then subtracted from it. Meanwhile, a low-pass Chebyshev filter was applied to each PPG signal and then subtracted from it to remove frequency components outside of the range of the heart rate. The order of the filter was determined

dynamically based on the sampling frequency of each signal. In all cases, the original ECG and PPG waveforms were retained for estimation of BW due to respiration later on.

After removing the low-frequency BW components from the signals, it was observed that many ECG signals still appeared noisy. To denoise the ECG signals, a seventh-order Savitsky-Golay filter was utilized. This filter type was chosen due as they are well-known to preserve small details of a waveform, such as the Q- and S-waves found in ECG signals.

After signals were denoised, all records including ECG, PPG and respiratory signals were split into segments. In this chapter, three different segment lengths are trialled to determine the most suitable length for accurate RR prediction. The segments chosen were 20, 30, and 60 seconds. These segment lengths are commonly used in the literature, allowing for fair comparison. They also each enable very frequent RR estimation, while also providing a wide enough window to accurately calculate even very low RRs. At this point, any segment with a missing signal or flat-lining signal was discarded.

For each 20-second segment, the R-waves (or peaks) of the ECG signals were found, as well as the peaks of the PPG and reference RR signals. Additionally, the beat-to-beat intervals were calculated for PPG and ECG signals, and the breath-to-breath (BrTBr) interval was calculated for RR signals. Heart rate (HR) was then calculated from both the PPG and ECG signals, before RR was calculated from the reference respiration signal. This extracted information was then used by a purpose-built signal quality index (SQI) as described in the next section, to determine the overall quality of the segment and thus the segment's suitability for training and testing the neural networks.

Furthermore, the RR of each 20-second segment was calculated by finding the average period between peaks of the respiration signal. This period represents one full breath, and thus the RR was calculated using the following formula:

$$RR_{\text{true}} = \frac{60}{mean(BrTBr_1, BrTBr_2, ..., BrTBr_n)} \tag{4.1}$$

where 'BrTBr' represents a breath-to-breath interval measured in seconds, 'n' is the number of BrTBr intervals within for the 20-second respiratory signal,

and the 'RR_{true}' is taken as the "true RR" for that segment.

4.2.3 Signal Quality Assessment

Signal quality assessment is vital to ensure that neural networks are learning from realistic data. One significant work [154] found that simple conditional statements can be used to effectively assess the quality of PPG, ECG, and blood pressure (BP) signals. In these works, various sanity checks were performed to determine the quality of a signal, such as ensuring that heart rate (HR) and beat-to-beat (BTB) intervals were within reasonable ranges. Reasonable range for RR were determined based on clinical medicine resources

In this work, PPG and ECG signals are considered with respect to calculating RR, and as such the quality of the respiration signal is also vital. As such, this work develops an SQI tool based on conditional statements relevant to the problem in order to successfully classify a record containing PPG, ECG and respiration signals as either "good" or "bad" based on a series of conditional statements. This is described by the following algorithm:

Algorithm 2 Signal Quality Index Algorithm

Input: hr_ppg, hr_ecg, ppg_peak_ratio, ecg_peak_ratio, ppg_btb_ratio, ecg_btb_ratio, true_rr, true_rr_peak_ratio, true_rr_brtbr_ratio

Output: signal_quality

1: **if** [(abs(hr_ppg - hr_ecg) < 10) & (hr_ppg > 40) & (hr_ppg < 180) & (ppg_peak_ratio < 1.5) & (ecg_peak_ratio < 1.5) & (ptp_btb_ratio < 1.5) & (ecg_btb_ratio < 1.5) & (true_rr > 8) & (true_rr < 35) & (true_rr_peak_ratio < 1.5) & (true_rr_brtbr_ratio < 1.5)] **then**
2: signal_quality = 1
3: **else**
4: signal_quality = 0
5: **end if**

In this algorithm, *hr_ppg* and *hr_ecg* are the HR values calculated from the

PPG and ECG signals, respectively. They are compared to each other to verify that they were acceptably similar, then *hr_ppg* was checked to ensure that HR was within the physiologically probable range of 40-180 bpm [78]. Meanwhile, *ppg_peak_ratio, ecg_peak_ratio* and *true_rr_peak_ratio* represent the ratio of the maximum to minimum peak heights for the PPG, ECG and reference RR signals respectively, and *ppg_btb_ratio, ecg_btb_ratio* and *true_rr_brtbr_ratio* represent the ratio of maximum to minimum PPG signal BTB intervals, ECG signal BTB intervals and reference RR signal BrTBr intervals respectively. It was checked that each of these ratios was <1.5 to ensure that there was acceptable consistency within each individual signal, as consistency is a strong indicator of signal quality. Lastly, *true_rr* represents the RR extracted from the reference signal using Eq. 4.1, and it was checked that this fell within the conservative range of 8-35, as the RR of a healthy adult would fall between 15-30 BrPM [163] but within an ICU environment it is likely that RRs across critically ill patients would be highly variable. Records that met all criteria were assigned a *signal_quality* of 1, meaning "good", while failure to meet any criteria resulted in a *signal_quality* of 0, or "bad".

After testing all 20-second segments with the SQI tool, a total of 19,084 "good" records were found. The next stage was to extract features from each of these signals for use in training the neural networks. This was a multi-step process, which begins with the extraction of respiration-induced modulations from the ECG and PPG signal as discussed in the following subsection.

4.2.4 Extracting Respiratory Signals from ECG and PPG

If PPG and ECG signals were recorded with no interference from respiration or movement, they would appear as is shown in Fig. 4.1. However, this is not the reality. Recall that respiration can modulate the ECG and PPG signals in three key ways - baseline wander (BW) modulation, amplitude modulation (AM) and frequency modulation (FM) caused by respiratory sinus arrhythmia. As previously discussed, one or more respiratory modulations may be absent from the PPG and ECG signals of some patients. As such, endeavouring to extract all three key modulations from both the ECG and PPG signal will greatly enhance

a neural network's ability to estimate true RR.

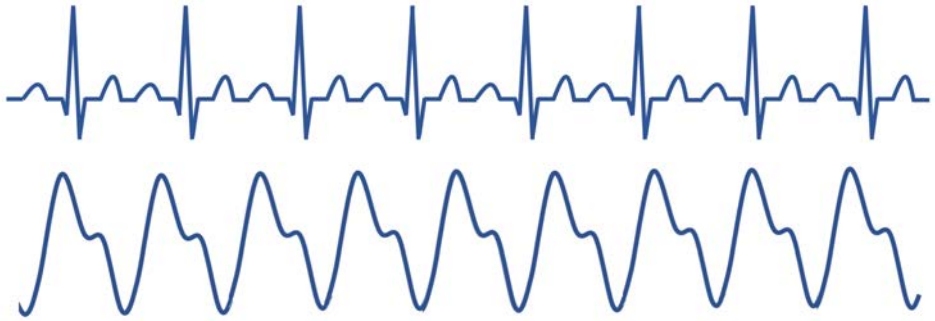


Figure 4.1: Sample ECG and PPG unaffected by respiration.

Extracting Respiratory Signals

In the context of respiration, BW is the overall shift in the baseline of an ECG or PPG signal due to respiration, as is shown in Fig. 4.2. BW was obtained by low-pass filtering the ECG and PPG signals. Hereafter the BW signals extracted from the PPG and ECG signals are denoted as PPG-BW and ECG-BW, respectively. Meanwhile, AM presents as the variation in peak heights in the ECG and PPG signals, after BW has been removed, as shown in Fig. 4.3. Finally, FM presents in ECG or PPG signals as varying beat duration, as shown in Fig. 4.4. Thus, AM and FM respiration signals are easily derived from the peak heights and BTB intervals of the waveforms, respectively. The AM and FM signals extracted from PPG and ECG are henceforth denoted as PPG-AM, PPG-FM, ECG-AM, and ECG-FM.

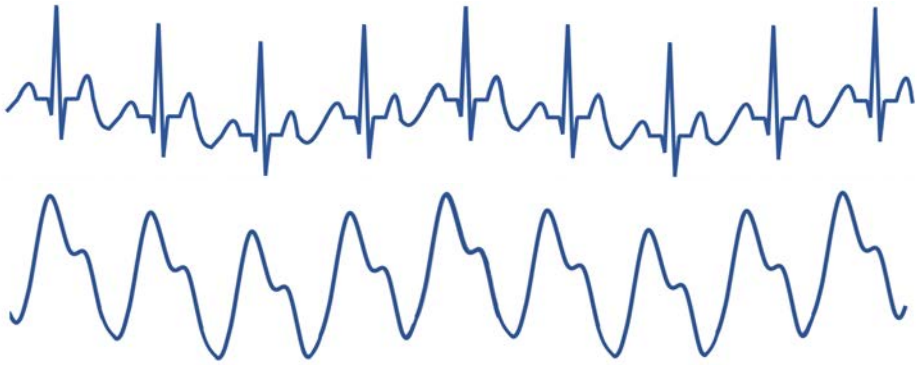


Figure 4.2: BW in the ECG and PPG signals.

Figure 4.3: AM in the ECG and PPG signals.

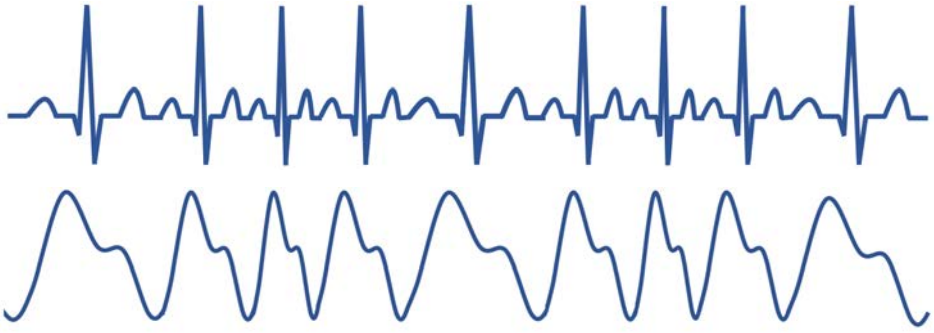


Figure 4.4: FM in the ECG and PPG signals.

After BW, AM and FM signals were extracted from the PPG and ECG signal, peaks and troughs of each signal were calculated and stored in six separate vectors. Breath-to-breath intervals, as well as the intervals between trough locations, were also calculated and stored in six additional vectors. These parameters were then used by the developed respiratory quality index (RQI) tool described in the following subsection.

Finally, a possible respiratory rate was derived from each signal by finding the average period between peaks (the breath-to-breath interval), and thus determining the number of breaths per minute. This process is mathematically defined as:

$$RR_{\text{signal}} = \frac{60}{mean(BrTBr_1, BrTBr_2, ...BrTBr_n)} \tag{4.2}$$

where 'BrTBr' is a breath-to-breath interval, 'n' is the number of BrTBr intervals within the extracted signal, and the 'signal' of RR_{signal} is the PPG-BW, PPG-AM, PPG-FM, ECG-BW, ECG-AM or ECG-FM.

4.2.5 Respiratory Quality Assessment

The development of an RQI scheme that assigns each modulation-extracted respiratory signal a quality rating on some scale could improve RR estimation algorithms, as knowledge about the quality of each estimated RR can enhance the networks ability to determine true RR based.

This chapter proposes an efficient and effective RQI scheme that considers the variance in peak heights (ph), trough depths (td), and the distances between peak pairs (p-p) and trough pairs (t-t) for any given extracted RR signal.

Consistency is a key indicator of respiratory signal quality, and as such a metric called the differential coefficient of variation (DCV) metric is developed, a variation on the the coefficient of variation (CV), to quantify how much variation is in the signal. The DCV is calculated as follows:

$$DCV = 1 - \frac{\sigma}{\mu} \tag{4.3}$$

where σ represents the standard deviation (SD) and μ represents the mean of the vector of data. The DCV is calculated for each of the four properties of interest - peak height, trough depths, distance between peak pairs, and distance between trough pairs. These are denoted as $\mathrm{DCV_{ph}}$, $\mathrm{DCV_{td}}$, $\mathrm{DCV_{p\text{-}p}}$ and $\mathrm{DCV_{t\text{-}t}}$ in Equation (4.4), respectively.

As is shown in Equation (4.4), the RQI is then calculated by finding the average of the four DCVs.

$$RQI = \sum \frac{DCV_{\mathrm{ph}} + DCV_{\mathrm{td}} + DCV_{\mathrm{p\text{-}p}} + DCV_{\mathrm{t\text{-}t}}}{4} \tag{4.4}$$

The calculated RQI will be 0 in the case where there is no consistency, and 1 in the case where there is perfect consistency. As consistency is the best indicator of signal quality, higher RQI values indicate better quality signals.

This scheme was used to calculate an RQI for each of the six modulation-extracted respiratory signals in every 20-second record; PPG-BW, ECG-BW, PPG-AM, ECG-AM, PPG-FM, and ECG-FM.

4.2.6 Feature Selection

Two separate feature vectors are developed to analyse the performance of neural networks with and without the RQI features as inputs. For the first test, only the modulation-extracted RRs are selected, resulting in a six-feature input vector as follows:

$$[RR_{\text{ECG-BW}}, RR_{\text{PPG-BW}}, RR_{\text{ECG-AM}},$$
$$RR_{\text{PPG-AM}}, RR_{\text{ECG-FM}}, RR_{\text{PPG-FM}}]$$

For the second test, the feature vector included RQIs calculated using the scheme proposed in this chapter, along with the modulation-extracted RRs. The resultant twelve-feature vector is as follows:

$$[RQI_{\text{ECG-BW}}, RR_{\text{ECG-BW}}, RQI_{\text{PPG-BW}}, RR_{\text{PPG-BW}},$$
$$RQI_{\text{ECG-AM}}, RR_{\text{ECG-AM}}, RQI_{\text{PPG-AM}}, RR_{\text{PPG-AM}},$$
$$RQI_{\text{ECG-FM}}, RR_{\text{ECG-FM}}, RQI_{\text{PPG-FM}}, RR_{\text{PPG-FM}},]$$

These two feature vectors were constructed for every record that was classified as 'good' by the SQI tool.

4.2.7 Neural Network Structure

In this work, a bidirectional long short-term memory (BiLSTM) network structure is used to predict respiratory rate from the input features. BiLSTM cells are updated using the same mathematical structure as unidirectional long short-term memory cells, but the data is passed through the network both as-is (forwards) and in reversed order (backwards). The results of these operations is then concatenated before passing to the next layer. The mathematical structure of a single forward or backwards pass is described by the following equations.

$$\tilde{c}_{\text{t}} = \tanh(w_{\text{c}}[\, a_{(\text{t-1})}, x_{\text{t}}] + b_{\text{c}}) \tag{4.5}$$

$$f_{\text{t}} = \sigma(w_{\text{f}}[\, a_{(\text{t-1})}, x_{\text{t}}] + b_{\text{f}}) \tag{4.6}$$

$$u_{\text{t}} = \sigma(w_{\text{u}}[\, a_{(\text{t-1})}, x_{\text{t}}] + b_{\text{u}}) \tag{4.7}$$

$$o_{\mathrm{t}} = \sigma(w_{\mathrm{o}}[\, a_{(\mathrm{t\text{-}1})}, x_{\mathrm{t}}] \, + b_{\mathrm{o}}) \tag{4.8}$$

$$c_{\mathrm{t}} = u_{\mathrm{t}} \bullet \tilde{c}_{\mathrm{t}} + f_{\mathrm{t}} \bullet c_{(\mathrm{t\text{-}1})} \tag{4.9}$$

$$a_{\mathrm{t}} = o_{\mathrm{t}} \bullet \tanh(c_{\mathrm{t}}) \tag{4.10}$$

where , $\mathrm{w_c}$, $\mathrm{w_f}$, $\mathrm{w_u}$ and $\mathrm{w_o}$ refer to the learned weights for their respective operations, while $\mathrm{b_c}$, $\mathrm{b_f}$, $\mathrm{b_u}$ and $\mathrm{b_o}$ are the learned biases. Once again, these are learnt during training using the Adam optimization algorithm. Additionally, the parameter $a_{(\mathrm{t\text{-}1})}$ refers to the output of the previous layer, while x_{t} is the input for timestep t. Eqn. (4.9) utilizes the results of (4.5) through (4.7) as well as the cell state of the previous time step, $c_{(\mathrm{t\text{-}1})}$ to update the cell state, and (4.10) uses the resultant c_{c} as well as the output gate results. The '•' symbol in (4.9) and (4.10) represents element-wise matrix multiplication.

The neural network structure utilised in this work includes three hidden BiLSTM layers each comprised of the forward and backwards passes followed by the concatenation operation. The first two hidden layers return a sequence of all hidden cell states, hence the high number of concatenation operations. The third hidden layer outputs only the final state of each cell from both the forward and backwards pass, and these are then concatenated. The network structure is illustrated in Fig. 4.5 below.

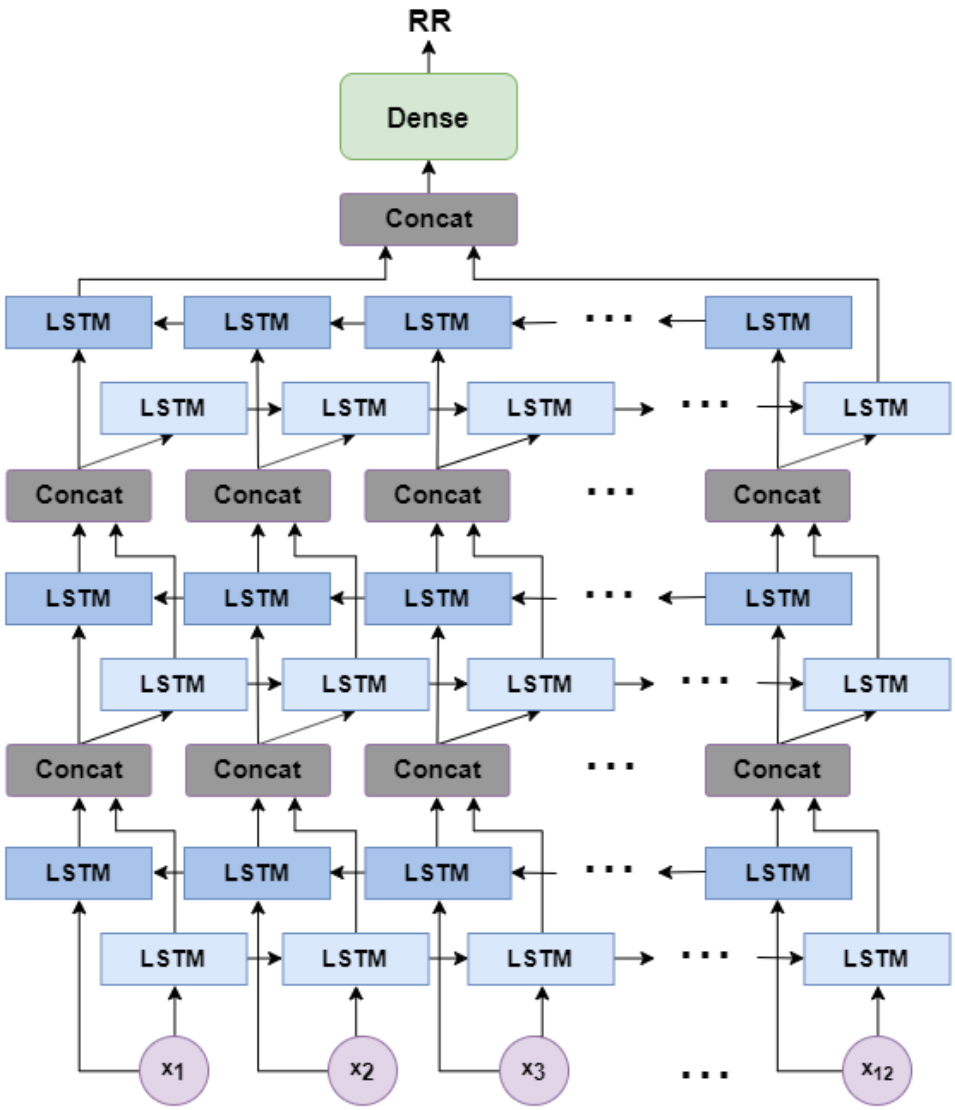


Figure 4.5: Structure of the BiLSTM model.

The NN structure included 128 hidden units per hidden layer and a batch size of 1024 to enable good generalization without overfitting. Adam optimization [157] is used to update weights and biases during training, while the mean absolute error (MAE) is used as the loss function.

4.2.8 Training & Testing the Algorithms

In this work, the NN structure was trained six times to compare the performance of the network using the six different feature vectors, as follows:

- All 12 features, as calculated from 20-second segments
- The 6 RR features only, as calculated from 20-second segments

- All 12 features, as calculated from 30-second segments
- The 6 RR features only, as calculated from 30-second segments
- All 12 features, as calculated from 60-second segments
- The 6 RR features only, as calculated from 60-second segments

The data was pseudorandomly shuffled before being split into subsets for training, validating, and testing. 80% of the data was used for training the NNs, 10% was used for fine-tuning hyperparameters through the validation process, and the remaining 10% of unseen data was utilized to fairly test the models.

4.3 Results & Discussions

After training and testing all of the NN configurations, statistical and graphical analysis was conducted to assess the performance of each network. In terms of statistical analysis, several informative metrics were considered: mean absolute error (MAE), root mean square error (RMSE), and Pearson's correlation coefficient (PCC). Furthermore, Bland Altman analysis was conducted by calculating the bias or mean difference (MD) and the width between the limits of agreement (LOAs).

Segment Length	**Features**	**MAE (BrPM)**	**RMSE (BrPM)**	**PCC**	**MD**	**LOA Width**
20 seconds	RR & RQIs	0.821	2.236	0.891	-0.08	8.76
	RR Only	1.301	2.776	0.829	-0.16	10.87
30 seconds	RR & RQIs	0.747	1.926	0.901	0.14	7.54
	RR Only	1.116	2.430	0.839	-0.04	9.53
60 seconds	RR & RQIs	0.638	1.575	0.932	-0.15	6.17
	RR Only	0.711	1.731	0.919	-0.14	6.79

Table 4.1: Performance of BiLSTM NN using various feature vectors for estimating respiratory rate

MAE gives key insight into how skilled the network is at producing a reasonable prediction for RR. RMSE is indicative of how many high-range errors there

are, and thus provides information about whether the network has fit appropriately to the data. PCC indicates the level of linear correlation, and will give a result between 0 and ±1, representing no correlation and total positive/negative correlation respectively.

In terms of the Bland Altman analysis metrics, a low MD along with narrow LOAs is a good indicator of strong agreement between the two methods of measurement. In Bland Altman analysis, each data point is the result of comparing the mean of the two measurement methods with the difference between their predictions. As such, a high-performing network would have low MD and low LOA width.

The results of calculating these metrics for the BiLSTM NNs trained using each feature vector are shown in Table 4.1. These results clearly indicate that the inclusion of RQIs calculated using the proposed scheme greatly improves the success of machine learning in estimating true RR. Table 4.1 shows that the inclusion of RQI features reduced the MAE by up to 36.89% when compared to the equivalent networks that were trained using solely the modulation-extracted RRs. Significant improvements RMSE and PCC are all also visible across all NN structures considered. In all cases, including RQI features increased the level of agreement between true and predicted RR measurements, narrowing the LOA width. MDs were extremely small across all networks.

Table 4.1 also shows that the BiLSTM network model performs strongly regardless of the segment length used to derive the RR and RQIs, however MAE is shown to decrease as segment length is increased. The overall lowest MAE was 0.638, achieved by the network trained on RRs & RQIs extracted from 60 second segments. As the inclusion of RQI features is shown to reduce MAE, the remainder of this analysis will focus on the networks trained with both RR & RQI features.

To further analyse the predictive performance of the BiLSTM network, the following error histograms were created to graphically investigate the spread of errors in RR predictions. To create these figures, all errors were rounded to the nearest 0.25 to allow for better visualisation. These figures reiterate the high accuracy of the systems trained using both RR and RQI features.

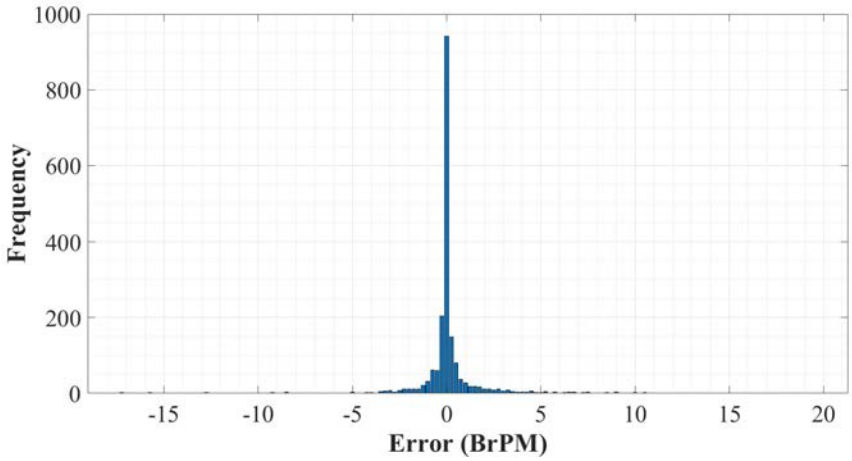


Figure 4.6: Error Histogram for RR Estimation using RR & RQI features derived from 20-second PPG & ECG segments.

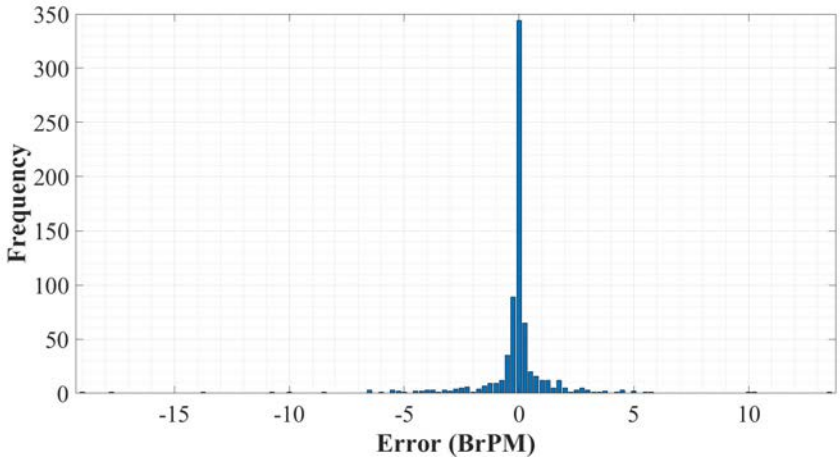


Figure 4.7: Error Histogram for RR Estimation using RR & RQI features derived from 30-second PPG & ECG segments.

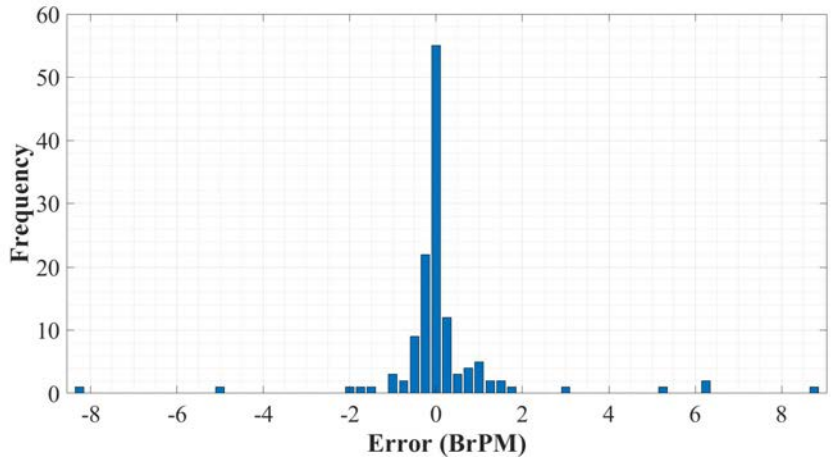


Figure 4.8: Error Histogram for RR Estimation using RR & RQI features derived from 60-second PPG & ECG segments.

The performance of the BiLSTM network is further analysed via the Bland Altman plots in Figs. 4.9-4.11. Bland Altman plots are used to assess the level of agreement between two measurement methods - in this case, comparisons are made between the proposed BiLSTM model against the reference RR measurement from the MIMIC-III database. The difference between the two measurements is plotted against the mean of the two measurements, and as such a high density around the central 'mean difference' line within the 'limits of agreement' indicates strong agreement between two schemes. In each plot, the difference vs. mean results were often extremely close together and appeared to overlap. As such, a density color scale is included in Figs. 4.9-4.11 to better illustrate the concentration of points. As can be seen from these plots, there is a high density of points along the mean difference line, with LOA widths falling below 10 BrPM for all parameters. This indicates a strong correlation between the true RRs and those predicted by the proposed network, regardless of the segment length used for feature extraction.

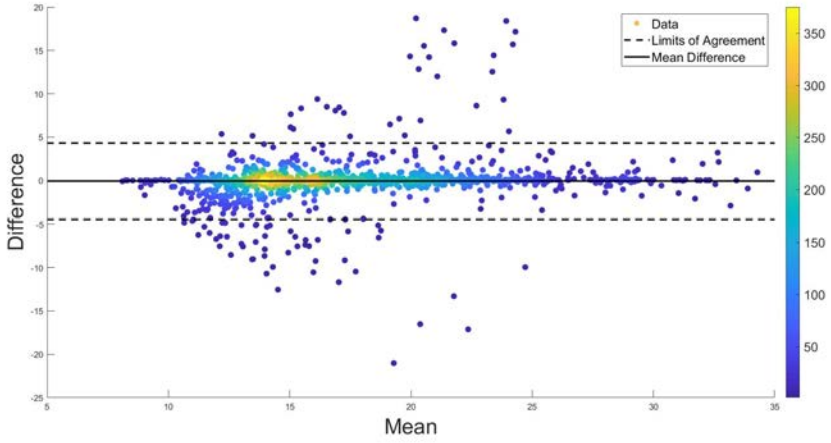


Figure 4.9: Bland Altman Plot for RR Estimation using RR & RQI features derived from 20-second PPG & ECG segments.

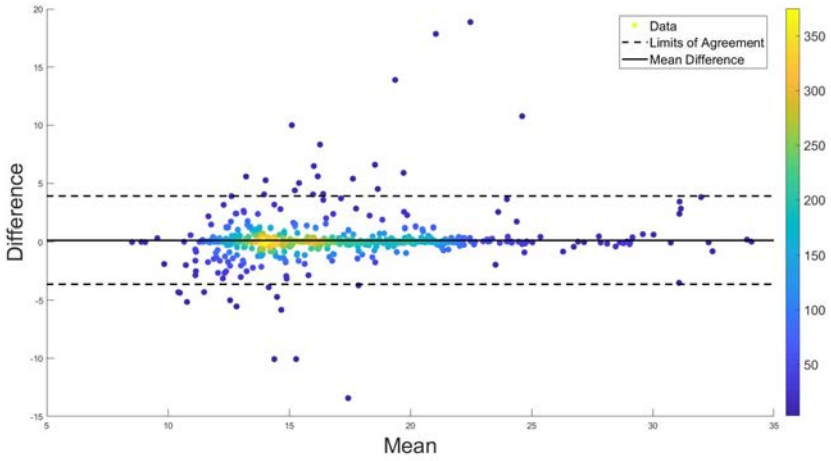


Figure 4.10: Bland Altman Plot for RR Estimation using RR & RQI features derived from 30-second PPG & ECG segments.

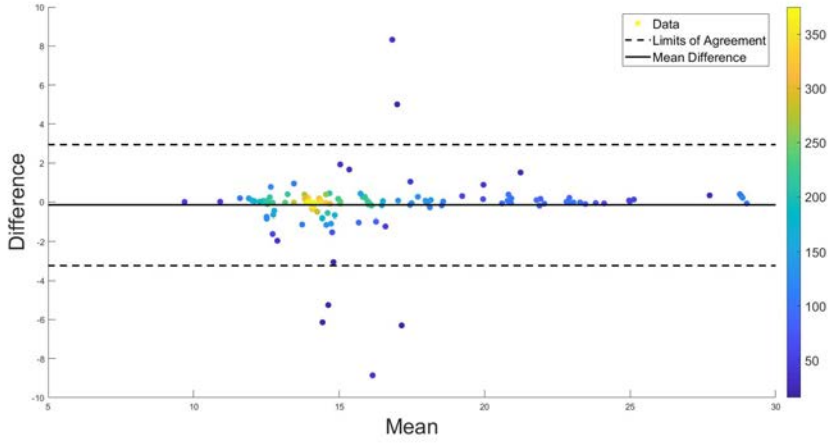


Figure 4.11: Bland Altman Plot for RR Estimation using RR & RQI features derived from 60-second PPG & ECG segments.

To further assess the correlation between the true and predicted values for RR, the regression plots in Figs. 4.12-4.14 were constructed. In each figure, the thick black line represents what 'perfect' correlation would look like, while the dashed black line is the actual correlation achieved by the network. From this regression plot, it is clear that there is a strong correlation between the predictions made by the BiLSTM model and the reference RRs obtained from the MIMIC-III database, regardless of the segment length used to derive the features. In each plot, the actual correlation line falls very close to the ideal correlation line, and very few data points are outliers in the trend.

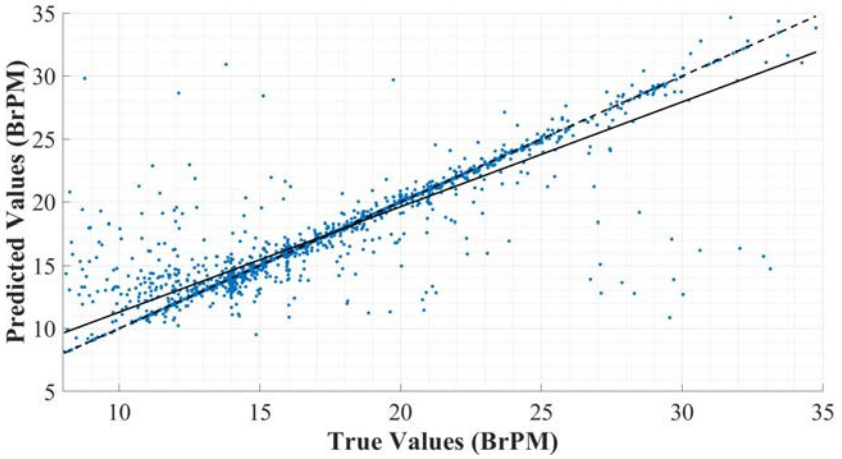


Figure 4.12: Regression Plot for RR Estimation using RR & RQI features derived from 20-second PPG & ECG segments.

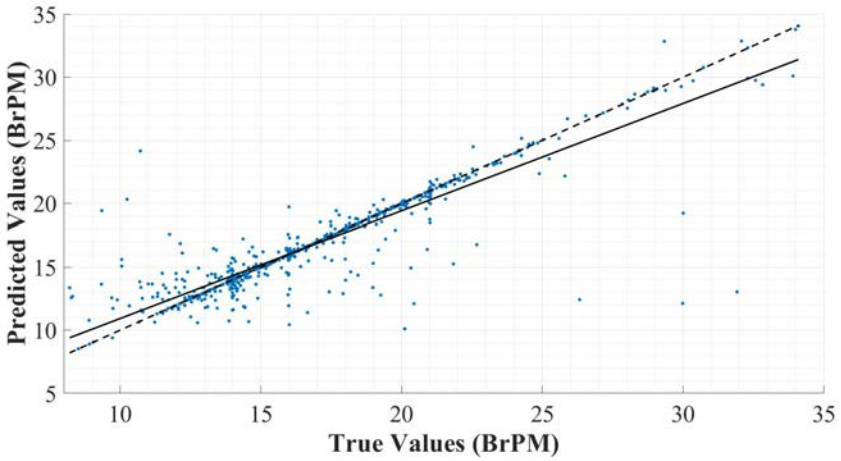


Figure 4.13: Regression Plot for RR Estimation using RR & RQI features derived from 30-second PPG & ECG segments.

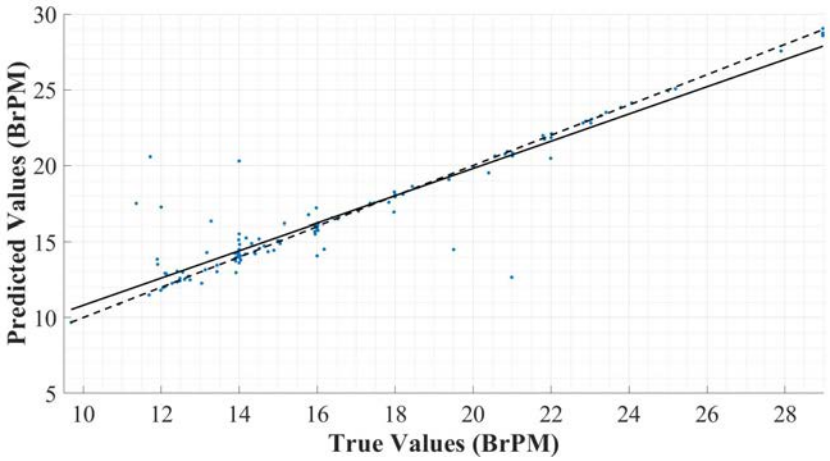


Figure 4.14: Regression Plot for RR Estimation using RR & RQI features derived from 60-second PPG & ECG segments.

Overall, the BiLSTM model shows low error and a high level of agreement with gold-standard measurement, regardless of which segment length is used for feature extraction. Performance increased as segment length increased, but even shorter segments showed strong results. In all cases, the inclusion of features calculated based on the proposed RQI scheme greatly improves the performance of the BiLSTM neural network. Therefore, it is clear that a BiLSTM model utilising extracted RRs and the proposed RQIs would significantly improve RR calculation in clinical and at-home environments, with longer ECG and PPG segments for feature extraction leading to the most accurate predictions.

4.3.1 Comparison to Previous Works

The results obtained by the BiLSTM model compares well to previous works when the feature vectors with both modulation-extracted RRs and corresponding RQIs were used, regardless of segment length. This is shown in Table 4.2. It is clear that both models outperform the previous state-of-the-art schemes for RR estimation from ECG and PPG signals, achieving significantly better MAE and comparable RMSE.

		Error Metrics (BrPM)	
	Segment Length (s)	**MAE (BrPM)**	**RMSE (BrPM)**
Orphanidou [39]	60	1.80	N/A
Karlen [40]	60	N/A	2.3
Birrenkott [43]	32	0.71[1], 3.12[2]	N/A
Pirhonen [45]	N/A	1.764	3.996
BiLSTM + RQI	20	0.821	2.236
BiLSTM + RQI	30	0.747	1.926
BiLSTM + RQI	60	0.638	1.575

Table 4.2: Comparison to previous works

[1] Based on testing against 42 Capnobase [164] records

[2] Based on testing against 53 records MIMIC-II [165] records

[3] Based on testing against 42 Capnobase [164] records, results varied based on window length selected and on signal used (PPG or ECG)

Compared to the works presented in Table 4.2, the BiLSTM model with RR and RQI features perform extremely strongly regardless of segment length used to extract these parameters. The RMSEs of all models were lower than the previous works in the literature. In terms of MAE, the model trained using 60s segments outperformed all previous works. One work [43] reported a lower MAE of 0.71 BrPM on the Capnobase database than was achieved by the proposed models based on 20s and 30s signal segments, however the MAE of [43] rose to 3.12 BrPM when the scheme was applied to the larger and more comprehensive MIMIC database. As this chapter is based on MIMIC data, the latter result

is more comparable. Overall, the BiLSTM model outperforms the literature in terms of MAE regardless of segment length.

Interestingly, enhanced results were achieved even where the short window length of 20 seconds was used. Accuracy increased with time, however the risk of artefacts impacting the signal quality also increases with the length of the segment. This suggests that the proposed scheme could predict RR faster, while also achieving a lower error.

It is also worth noting that the previous works largely relied on very small datasets. Through using a large database for this work, it has been possible to thoroughly validate the performance of the network across a large and diverse set of patients. The results presented in this chapter were obtained through training and testing the proposed schemes on the large and diverse MIMIC-II database, compared to other recent works such as [43, 45] where 95 and 29 records were used to obtain the results in Table 4.2, respectively. This ultimately means that the proposed BiLSTM network with RQI features is more likely to translate to real-world application with success, while many of the previous works would need to be validated on larger databases.

4.4 Conclusion

In this work, an RQI scheme was developed to enhance the performance of neural networks utilizing the respiratory modulations of ECG and PPG signals to estimate true RR. The proposed RQI scheme was implemented and tested to evaluate improvements in the performance of NNs in predicting RR from modulation-extracted RR estimates, with exceptional results.

When RQIs were used alongside modulation-extracted RRs as input features, a bidirectional LSTM model was able to achieve the low MAE of 0.821 BrPM. This is a significant improvement when compared to other works in the literature, and proves that RQIs can greatly enhance the performance of neural networks.

The results of this chapter show that a device implementing the proposed RQI scheme with a BiLSTM NN would be suitable for continuous and non-invasive monitoring of respiratory rate, using hardware that is already in place

in many healthcare environments. This algorithm would likely be suitable for clinical use due to the low error and strong agreement with current gold-standard measurements. It also offers ease of implementation that is achieved through utilising sensor data already available in clinical and at-home healthcare settings. With further validation on persons outside of ICU, it would also be suitable for at-home health monitoring. This scheme could greatly improve early prediction of potentially fatal conditions, enhance remote healthcare, and ultimately improve patient outcomes.

This chapter addresses the research problem of respiratory rate monitoring that was presented in Section 1.2.1 and provides the third original contribution listed in Section 1.3. The high performance of LSTM networks in this chapter further supported their use as part of the hybrid networks used in subsequent research chapters. Additionally, the development of this scheme alongside the blood pressure schemes presented in Chapter 3 supports the development of enhanced diagnostics and prognostics tools. The following chapters therefore focus on developing prognostics tools that assess mortality risk using only vital signs and basic demographics.

Chapter 5

Continuous and Automatic Mortality Risk Prediction for Adult Patients using Vital Signs

This chapter contains material that has been published in the following article:

[2] **S. Baker**, W. Xiang, and I. Atkinson, "Continuous and Automatic Mortality Risk Prediction using Vital Signs in the Intensive Care Unit: A Hybrid Neural Network Approach," *Scientific Reports*, vol. 10, pp. 21282, December 2020.

5.1 Introduction

Intensive care units (ICUs) treat the most critically ill patients, and as a result are known to have the highest mortality rate of hospital units [9]. In 2001-2012, mortality rates across several ICUs in the United States (US) ranged from 11.3%-12.6% [118]. As such, typical ICUs have high staff-to-patient ratios, and it has been found that outcomes are improved where there are a higher number of nurses and consultants per bed [10]. However, providing a high standard of critical care comes at a cost, with $108 billion spent on critical care medicine in the US in 2010, accounting for 0.72% of gross domestic product (GDP) [166]. The use of mortality risk assessment tools can aid in resource allocation and treatment decisions, potentially reducing costs while continuing to provide a high standard of care to critically ill patients.

There are several points-based schemes currently used to quantify mortality risk in ICUs today, including multiple iterations of the Acute Physiology and Chronic Health Evaluation (APACHE) score and Simplified Acute Physiology Score (SAPS). While newer versions exist, APACHE-II [119] and SAPS-II [120] remain the most commonly used mortality risk assessment tools worldwide [122]. Another commonly used tool is the Sequential Organ Failure Assessment (SOFA) score [121], which was developed to assess sepsis risk but has since been found to be a relatively good predictor of mortality. Unfortunately, there are several limitations associated with these tools. Firstly, it has been found that their performance decreases fairly rapidly over time, with Kramer [30] indicating that SAPS II was out of calibration by 2005. Several subsequent studies have also identified calibration problems with APACHE, SOFA, and SAPS [123–125]. Calibration can be lost over time due to changing patient populations and medical treatments, and typically results in overestimation of mortality [30]. Aside from the effect of time, Sakr *et al.* [123] and Lew *et al.* [125] noted that the schemes performed poorly for European and Singaporean cohorts, respectively. This indicates that insufficient consideration of diverse patient cohorts has also affected performance. Aside from calibration issues, these schemes rely on variables that can be time-consuming and difficult to obtain, such as pathological laboratory test results and patient medical history.

The limitations of existing scoring systems have lead to a rise in researchers exploring machine learning techniques for mortality prediction [35, 55–61], as well as the related issues of predicting the onset of various intervention methods [167, 168] detecting the risk of sepsis [46–48, 126] and other clinical deterioration events [49, 50]. Machine learning approaches have the advantage of being relatively easy to continuously update and recalibrate, with algorithms able to be configured in a way that enables continuous training based on new data obtained while it is being used in clinical environments. This in turn enables machine learning techniques to better generalise to current local or global populations, even as treatments and outcomes change with time. A recurring theme in these papers is a dependence on features including complex laboratory results, existing health conditions, and other patient history. Of the aforementioned studies that consider mortality

risk prediction, the majority depend heavily upon laboratory results [35, 56–60], which include values obtained from extensive blood, urine, breath, and other clinical analysis, and are often complex and time-consuming to obtain and then enter in patient's medical records.

A common metric for assessing the discrimination performance of diagnostic tools is the area under the receiver-operator curve (AUROC). In terms of this metric, the highest performing mortality prediction systems were presented by Johnson *et al.* [59] and Delehanty *et al.* [60] with AUROCs of 0.927 and 0.94, respectively. However, the system presented by Johnson *et al.* [59] is dependent on 148 features comprised predominantly of complex laboratory results. This limits the usefulness of the system, as medical staff would need to measure and enter a massive number of variables to receive an accurate prediction.

Meanwhile, in the work presented by Delahanty *et al.* [60], only 17 variables were used - however, over 50% of the decision made by their system is based on All Patients Refined Diagnosis Related Groups (APR-DRG) risk of mortality and severity of illness, as well as Glasgow Coma Score (GCS), and the cost-weight index based on Medicare Diagnosis Risk Groups (MS-DRG). In short, this system depends heavily on diagnoses being made manually by doctors based on data available close to the time of ICU admission. This introduces potentially heavy human bias, and would not be functional for hospitals where the APR-DRG and MS-DRG diagnosis coding schemes aren't used.

Conversely, Deliberato *et al.* [55] investigated the use of features extracted from only vital signs, achieving a relatively low AUROC of 0.65, indicating low ability to distinguish between mortality and non-mortality cases. The authors also considered using vital signs in combination with other parameters, achieving a much higher AUROC of 0.84 when vitals data was combined with the GCS, SAPS-II score, patient demographics and information obtained about the patient during their hospital stay prior to ICU admission. This is certainly an improved performance; however it depends upon significant lab results and data from pre-admission. It would be preferable to use only vital signs and basic demographics, but with much higher performance than the 0.65 AUROC achieved by this work.

Another common theme in the literature is that of mortality prediction at

admission. While there are advantages of early mortality risk prediction, this method is inflexible and does not consider how a patient might respond to treatments after ICU admission. This was identified in a recent work by K. Yu *et al.* [61], where a bidirectional long short-term memory network was trained on multiple windows to identify mortality risk at any given time. This scheme uses the same features such as the SAPS-II score, which includes many laboratory values, GCS, demographic information, admission type, and comorbidities. As such, laboratory measurements would need to be repeated regularly for the system to predict effectively. Additionally, it requires 48 hours of data to predict future mortality risk effectively.

In the literature, there are many machine learning (ML) techniques considered for prediction of mortality. However, there have been relatively few that investigate the use of neural networks (NNs) specifically. Early works investigating NNs for mortality prediction focused on simple feed-forward neural networks [128–130], achieving comparable performance to scoring schemes such as APACHE. Works in recent years have begun to focus on NNs that are more advanced, such as long short-term memory (LSTM) NNs and convolutional NNs (CNNs). Several works [56, 61, 131] have identified long short-term memory (LSTM) networks as candidates for mortality prediction. LSTM has also proven successful in predicting septic shock [46] and other clinical deterioration events [50]. The primary advantage of LSTM is that it has the ability to 'remember' information that it has already seen, allowing it to identify relationships between different variables within the sequence.

Meanwhile, convolutional neural networks (CNNs) have also proven powerful in solving many medical problems such as detecting heart anomalies [51–53], identifying variations in Korotkoff sounds [54] and gait detection [169]. CNNs are exceptional at identifying the importance of certain features with respect to one another, and thus adding CNN layers prior to LSTM layers can greatly improve the predictive ability compared to pure LSTM. This was attempted by Alvis *et al.* [57], with their scheme achieving an AUROC of 0.836 when predicting ICU mortality from a 48-hour window of features including vital signs and laboratory values. This strongly suggests that a well-designed CNN-LSTM network would

be a good candidate for mortality prediction, combining the benefits of both network types for a resultant network that can identify important variables and any relationships that exist between them.

Aside from the model itself, another critical factor in the success of a neural network is the selection of features. To develop a system that could automatically update mortality risk throughout a patient's stay, it is essential to choose features that are both meaningful and easy to measure, ideally without any manual measurement. One recent work by Giannini *et al.* [126] considered hundreds of features for predicting septic shock, a strong risk factor for mortality. In a retrospective analysis, the authors found that 10 out of 20 of the most important features were derived from vital signs, while another was age. This finding is significant but largely unsurprising, given vital signs measure the most critical functions of the human body [127]. Fortunately, vital signs are regularly recorded, either through automatic measurements or regular manual measurements. These advantages mean that vital signs and statistics derived from them are ideal features for use in mortality prediction.

Another important factor in the real-world success of a NN for medical prediction problems is the adoption of clinicians. In recent years, many studies have identified the need for artificial intelligence to be interpretable, especially for healthcare applications [170–173]. Primarily, interpretability involves making the system easier to understand and therefore to trust. There are many ways to achieve this, including selecting features that are simple to understand from the perspective of domain experts [172].

As such, this chapter aims to develop a neural network approach for mortality using straightforward features with clear ties to patient health. Our work contributes to the literature through the development of the novel Artificial Intelligence Mortality Score (AIMS) scheme, a mortality risk classifier based on a hybridized CNN-LSTM network that uses only age, gender, and statistical parameters derived from a 24-hour window of vital sign measurements as features. AIMS is capable of continuously-updating prediction of the risk of mortality within 3-day, 7-day, and 14-day windows. Much of the previous literature focuses on predicting mortality events within the entire stay [57–60], however the

average length of stay in ICU in America is only 3.8 days [166]. Our analysis of the patients from the MIMIC-III database who met our selection criteria revealed that 65% of patients stayed in ICU for ≤ 3 days, 87% for ≤ 7 days, and 95% for ≤ 14 days. The models in the literature that focus on the entire stay would likely form a bias towards data obtained from shorter stays, given that these form the majority of cases. This in turn introduces the risk that models would not perform as well on longer-term patients. However, mortality risk assessment needs to be reliable for even the longest staying patients to ensure that they receive the appropriate care should they begin to stabilize or deteriorate. As our selection of 3-day, 7-day and 14-day windows encompasses the entire stay for the majority of patients, it enables fair comparison to the literature. In clinical practice, it offers the clear advantage of predicting risk within a clear time frame, with the score able to be easily and continuously recalculated continuously throughout the stay so that mortality risk is able to be quantified for all patients at all times, including for those with longer stays in ICU.

The remainder of this chapter is structured as follows; Section 5.2. describes the methodology used for extracting and processing data from the MIMIC-III database, as well as the structure of the AIMS network. Section 5.3. presents results and discussion, including comparison to currently used schemes and other novel schemes in the literature. Finally, Section 5.4. concludes the chapter and presents recommendations for future work.

5.2 Methodology

5.2.1 Selection of Data

The large amount of data used in this work was obtained from the Medical Information Mart for Intensive Care (MIMIC-III) clinical database [151]. The MIMIC-III database is comprised of deidentified data from over 60,000 ICU stays, including both adult and neonatal patients.

This study focuses on adult patients admitted to ICU for any reason, and thus the only criterion when selecting patient records was that the patient must be $\geq$18 years old. To be able to extract data for 3-day, 7-day, and 14-day mortality

risk prediction, we obtained a 14-day window for each patient. Our method for selecting data is outlined as follows:

- Where the patient survived their ICU stay, and their stay exceeded 14 days, the first 14 days of data after ICU admission were obtained;
- Where the patient died during their ICU stay, and their stay exceeded 14 days in length, the 14 days of data prior to their death time were obtained;
- Where the patient stay was shorter than 14 days, all data from ICU admission to discharge from ICU were obtained.

For all patients, some fundamental information was recorded during the data selection process - namely their age, gender, and time of death where applicable.

The events that were obtained from the database for our AIMS scheme were the heart rate (HR), systolic BP (SBP), diastolic BP (DBP), mean arterial pressure (MAP), respiratory rate (RR), blood oxygen levels (SpO_2), and temperature. All events matching this description were obtained, as our AIMS scheme depends upon statistical analysis of the variation of events such as HR and temperature. Vital signs were chosen as features for two main reasons: interpretability, and ease of measurement in the ICU. Vital signs are the most fundamental indicator of health, and are readily understood by all healthcare professionals. Most vital signs are easily measured using non-invasive equipment, enabling continuous measurement and thus data streams rich with information. Perhaps the most challenging to measure are BP and RR, with continuous methods currently either invasive or uncomfortable. However, recent research in measuring these parameters has focused on non-invasive, continuous methods [1], and as such it is likely that data streams for BP and RR measurement will become increasingly data rich as this technology is adopted into clinical practice. Richer data streams enable better quantification of the variability of vital signs, and thus would further improve the predictive performance of our network.

5.2.2 Feature Selection

For the development of our AIMS scheme, features were selected or derived from the commonly recorded parameters in the ICU. The first two features selected

are those that provide basic information about the patient: their age and gender. Age is recorded in years as an integer, while gender is recorded as a binary value where '1' and '0' represent female and male patients, respectively.

All other features selected were chosen to represent the vital signs of interest - HR, SBP, DBP, MAP, RR, SpO_2 and temperature. These parameters were regularly recorded in the MIMIC-III database, however the recording was often inconsistent with the frequency of measurement varying throughout the patient's stay. This resulted in highly variant quantities of data available for different patients. However, it has previously been shown that trends in vital signs can assist in identification of clinical deterioration in hospital settings [174]. As such, we apply statistical analysis to quantify the variability of each vital sign over the 24-hour window. This has the benefit of representing the inconsistent data within the database in a consistent manner, and also has the secondary benefit of improved computational efficiency.

We limit the acquisition window to 24 hours to ensure that the network is considering the patient's current health status. In our own experimentation, narrower windows reduced performance, while broader windows of 48 hours did not significantly improve performance. Additionally, if a wider window were used then the network may be prone to under- or over-estimating the severity of the patient's current condition based on their previous condition. For example, if all data from admission onward were used, then the patient might remain relatively stable for the first 9 days of their stay before showing signs of deterioration on the 10th day. If a risk window from admission onward was used, then overall the variation in the patients health would appear low, and thus the deterioration may not be noticed. Similarly, a patient who is highly unstable at the start of the admission but stabilises as a result of treatment might incorrectly be identified as a mortality risk. Considering a 24-hour risk window avoids this problem, as AIMS could be regularly and automatically recalculated throughout the entire stay, and thus would be more capable of identifying deterioration or stabilisation during long stays.

From the 24-hour window, the first and last values were taken to indicate how the vital sign changed from the beginning to the end of the window, while the

minimum and maximum values were recorded to show the most extreme events during the window. Mean and median were both chosen to provide an accurate representation of the average event for the vital sign - the use of both helps to reduce the risk of unusual data distribution skewing the result. Finally, the standard deviation (STD) is used to quantify the variability of the events for that vital sign. Where any result was NaN, it was replaced with a 'numerical NaN' chosen to be the extremely negative value of -999. If more than two vital signs were completely absent from a patient's records, then that record was discarded and not used for training or testing of the AIMS model.

After extracting the relevant data, the final feature array included the fundamental patient information, as well as the variability statistics for each vital sign. This resulted in a total of 51 features, including age, gender, and 7 statistical features for each of the 7 vital signs.

Our model was trained to predict three different cases; risks of mortality within 3 days, 7 days, and 14 days respectively. These models are hereafter referred to as AIMS-3, AIMS-7 and AIMS-14 respectively. These windows were selected to consider both immediate mortality risk, and longer term mortality risk. The feature vectors for risk of mortality within 3-day, 7-day and 14-day, risk windows considered the same features, however they were calculated from varying 24-hour windows in each case.

Where the patient survived their ICU stay, the relevant 24-hour window used for training and testing the model was always the first 24 hours after admission. Where the patient did die, the 24-hour window selected was dependent on the length of their stay. Where the patient died within less than the 3-day, 7-day, or 14-day risk period, the first 24 hours of data post-admission were used. Where their death occurred after a longer stay than the risk window, then their time of death was set as the end time, and the 24-hour prediction window was chosen to start from *(end time - size of the risk window)*. For example, where 3-day mortality risk was considered, the death time would be set as '72 hours'. Then, 72 hours prior to death would be chosen as relative 0, with the prediction window thereafter being data recorded between hours 0 to 24.

After data and feature selection, there were 3 distinct cohorts. This was

largely as a result of inconsistencies in data richness, with certain variables not recorded for some patients in different windows. This lead to some patients having data available for one or two risk windows, but not the remainder. The cohorts for AIMS-3, AIMS-7, and AIMS-14 are illustrated in Tables 1-3 below. Ages have been clustered by ranges, as ages exceeding 89 in the MIMIC-III database were set to values exceeding 300 for de-identification purposes [151].

Table 5.1: Characteristics of patient cohort for AIMS-3

Characteristic	**All patients (n = 51279)**	**Survived (n = 45863)**	**Died (n = 5416)**
Female	22415 (43.71%)	19888 (43.36%)	2527 (46.66%)
Age (Years)			
18-39	4896 (9.55%)	4687 (10.22%)	209 (3.86%)
40-59	14204 (27.70%)	13147 (28.67%)	1057 (19.52%)
60-79	21230 (41.40%)	19040 (41.51%)	2190 (40.44%)
≥ 80	10949 (21.35%)	8989 (19.60%)	1960 (36.19%)

Table 5.2: Characteristics of patient cohort for AIMS-7

Characteristic	**All patients (n = 51455)**	**Survived (n = 45863)**	**Died (n = 5592)**
Female	22483 (43.69%)	19888 (43.36%)	2595 (46.41%)
Age (Years)			
18-39	4906 (9.53%)	4687 (10.22%)	219 (3.92%)
40-59	14244 (27.68%)	13147 (28.67%)	1097 (19.62%)
60-79	21305 (41.41%)	19040 (41.51%)	2265 (40.50%)
≥ 80	11000 (21.38%)	8989 (19.60%)	2011 (35.96%)

Table 5.3: Characteristics of patient cohort for AIMS-14

Characteristic	**All patients (n = 51639)**	**Survived (n = 45863)**	**Died (n = 5776)**
Female	22560 (43.69%)	19888 (43.36%)	2672 (46.41%)
Age (Years)			
18-39	4916 (9.52%)	4687 (10.22%)	229 (3.96%)
40-59	14282 (27.66%)	13147 (28.67%)	1135 (19.65%)
60-79	21397 (41.44%)	19040 (41.51%)	2357 (40.81%)
≥ 80	11044 (21.39%)	8989 (19.60%)	2055 (35.58%)

5.2.3 Neural Network Structure

Hybrid NNs offer the advantages of multiple standard NN types. In this application, we develop a hybridized CNN-LSTM network, as shown in Fig. 5.1. CNNs are widely used to identify patterns and important features, while LSTM networks are known for their "memory", which enables them to remember which information in a sequence is the most important. Combining the two network structures results in a powerful hybrid NN with strong pattern and sequence recognition abilities, which is highly beneficial in an application where patterns and feature importances are not easily identified.

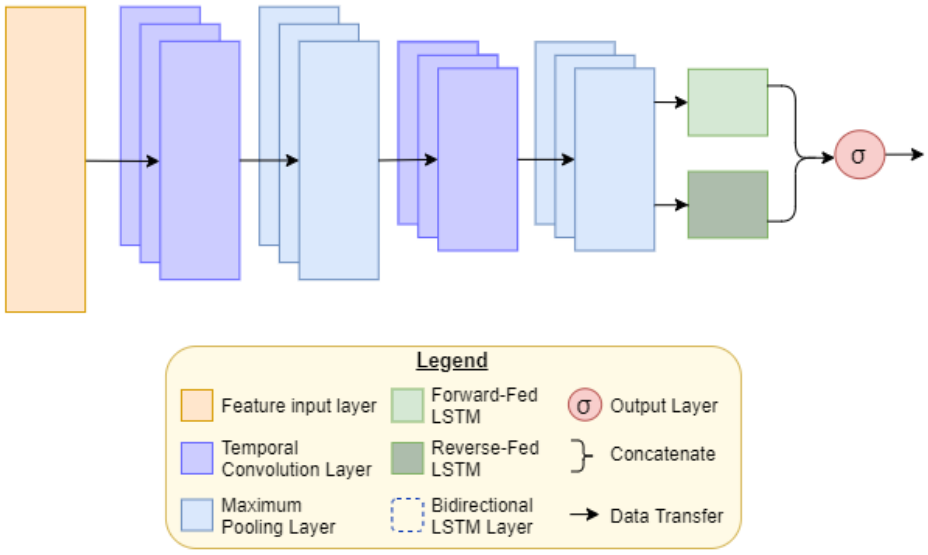


Figure 5.1: AIMS network structure.

As shown in Fig. 5.1, our network includes two temporal (one-dimensional) CNN layers with 128 hidden units each. Both CNN layers utilize rectified linear unit (ReLU) activation, described mathematically as

$$relu(z) = max(0, z) \tag{5.1}$$

The CNN layers can thereafter be mathematically described as follows:

$$y_{\mathrm{j}}^{\mathrm{i}} = relu(\sum_{n=1}^{N} w_{\mathrm{jn}}^{\mathrm{i}} * x_{\mathrm{m}}^{(\mathrm{i\text{-}1})} + b_{\mathrm{j}}^{\mathrm{i}}) \tag{5.2}$$

where y_j^i represents the output jth feature map of the ith layer. Convolution is indicated by the $*$ symbol. Weights are denoted using the term $w_{\mathrm{jn}}^{\mathrm{i}}$, which describes the nth weight of the jth feature map from the $(i-1)$th layer, where $n = 1, ..., N$. The parameter $x_{\mathrm{m}}^{(\mathrm{i\text{-}1})}$ represents the outputs of the $(i-1)$th layer, and finally b_j is the jth bias term of the ith layer. Weight and bias terms are all initialized to zero and updated using the Adam optimization algorithm [157] throughout training.

Temporal average pooling layers with a pool size of 2 and a stride size of 2 follow each of the CNN layers. This operation sweeps through the output of the CNN layers, taking the average of each pool it sees and outputting that value. Effectively, this downsamples the data by a factor of 2, helping prevent overfitting of the network. Average pooling layers are denoted in Fig. 5.1 as AvgPool-1 and AvgPool-2, respectively.

A bidirectional LSTM layer with 128 hidden units follows the final temporal average pooling layer. Bidirectional LSTMs (BiLSTMs) have the same mathematical structure as unidirectional LSTMs, but data is passed through the network in both the original and reversed orders. This allows for learning from both past and future values in the sequence. The results of both the forward and reversed passes are then concatenated to form the final output. The mathematical theory of LSTM networks is described by Hochreiter *et al.* [175].

The final layer of our AIMS network is a simple densely-connected unit utilizing sigmoid activation, which outputs a value between 0-1. If the result is $<$ 0.5, the network predicts that the patient will survive. Conversely, if the result is $\geq$ 0.5, the network predicts that the patient will die. The further away from

0.5 the output is, the more confident the network is in its prediction, and higher confidence typically corresponds with higher accuracy.

For the purposes of training and testing the performance of AIMS, thresholds are used to predict either 'mortality' or 'no mortality'. However, in terms of interpreting this result in a clinical sense, the overall prediction of 'mortality' or 'no mortality' could be provided alongside a confidence metric that indicates how certain the neural network was of its prediction. The raw 0-1 value outputted by the final layer of the network indicates the level of confidence the network has in its prediction. This confidence metric could be modified for easier understanding by the clinicians using the following equation:

$$ConfidencePercentage = \frac{|0.5 - output|}{0.5} \times 100 \tag{5.3}$$

Using this equation, a score of 0.14 would be interpreted as 'no mortality - 72% confident' while a score of 0.78 would be interpreted as 'mortality - 56% confident'. This easy-to-understand strategy would further increase the likelihood that clinicians would trust the model, as they would be able to better understand the severity of the patient's condition and it would be clearer that AIMS is not simply making binary decisions with full confidence. This metric would also give clinicians more insight into the path of treatment that would be most appropriate. For example, a patient who was predicted as 'mortality - 96% confident' may require more rapid and extreme treatment than a patient who was predicted as 'mortality - 2% confident'; effectively on the cusp of the two prediction classes.

5.2.4 Training & Testing the Algorithms

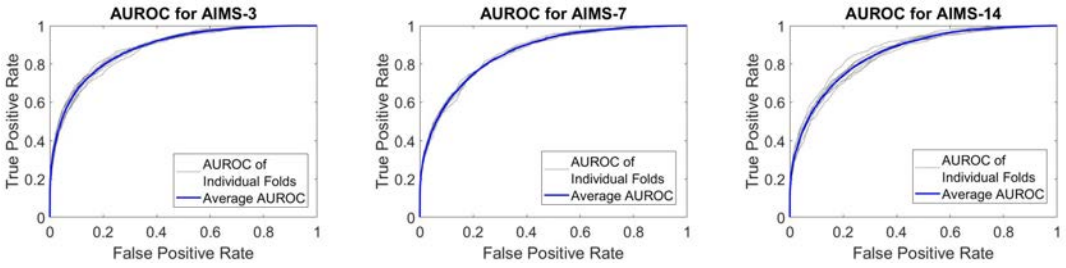
To train and test the proposed AIMS network, we used stratified k-fold cross-validation with 10 folds. Using this method, the data is split in 10 different ways, with all data being used as the testing set during one fold. Stratification ensures that each class is represented roughly equally in all splits. Cross-validation using k-fold gives a more realistic idea of the performance of a network.

After preprocessing, records from 51,279 unique patient stays were available for use in training and testing AIMS-3. There were slightly higher record numbers

of 51,455 and 51,639 for AIMS-7 and AIMS-14 respectively. The slight differences in record numbers are caused by a higher degree of missingness in some windows, leading to more data that was excluded in those cases. As we were using 10 folds for cross validation, 10% of the data was used for testing purposes and thus was unseen to the model for that fold. A further 80% of the data was used as the training set for AIMS, while the final 10% was used for validation, which improves fine-tuning of hyperparameters and allows for assessment of the "best" model during training.

The data available for this task was highly unbalanced, with mortality events only occurring in up to 11.19% of cases. To ensure that the model did not achieve high accuracy simply by overfitting to the majority class, a weighting of 9 was placed on the importance of learning the minority case. This value is reflective of the mortality rate within the ICUs included in the MIMIC-III database; it was chosen based on the fact that there were approximately 9 non-mortality cases to each mortality case within each cohort. This weighting ensured that the network considered the two classes to be equally important, and placed approximately equal emphasis on accurately predicting both mortality and survival. If the network is to be trained on an ongoing basis in the future, this weighting may need to be adjusted based on the mortality rates within the training set of data, but this could be done programmatically.

For each fold, the AIMS model was trained over 100 epochs with a batch size of 1024. These values were found to be optimal for ensuring that the model is capable of generalizing well, rather than overfitting to the training data. Binary crossentropy was used as the loss function, due to its clear suitability for this binary classification problem. For each fold, the "best" model weights were determined to be those that resulted in the lowest validation set loss during the 100 epochs of training; these weights were saved and used to test the model. The loss function used was binary cross-entropy, due to the binary classification nature of the model.



(a) AIMS-3 *(b) AIMS-7* *(c) AIMS-14*

Figure 5.2: Average ROC and ROC of each fold for 10-fold cross validation.

5.3 Results & Discussions

Following the training and testing of AIMS-3, AIMS-7 and AIMS-14 with 10-fold cross-validation, an extensive statistical analysis was conducted to assess the performance. The most commonly used metric in assessing the performance of a diagnostic tool is AUROC, which plots the true positive rate against the false positive rate. Figs. (5.2a-5.2c) illustrate the receiver-operator curves (ROCs) for each of the trained networks. It is clear from these figures that the ROC curve is highly consistent across all 10 folds, with none deviating far from the calculated average. This cross-validation confirms that the results obtained from the AIMS-3, AIMS-7, and AIMS-14 models are a realistic representation of how the network would perform in reality.

Fig. 5.3 further illustrates the differences between the ROC curves for the three networks. As can be observed from this figure, the AIMS-3 model achieves the highest AUROC, followed by AIMS-7 and then AIMS-14. This is a largely unsurprising result as AIMS-3 predicts the risk of death within the shortest window of 3 days. However, this figure also clearly demonstrates that all three models have high AUROC, meaning that they are able to distinguish between mortality and non-mortality cases very well.

A numerical summary of the AUROC results obtained across the 10-fold cross-validation of each model is presented in Table 5.4. This table highlights the minimum, maximum, and average AUROC obtained across the 10 folds when training each of the three models. These values again indicate the strong consistency across all 10 folds of cross-validation, further suggesting that this model

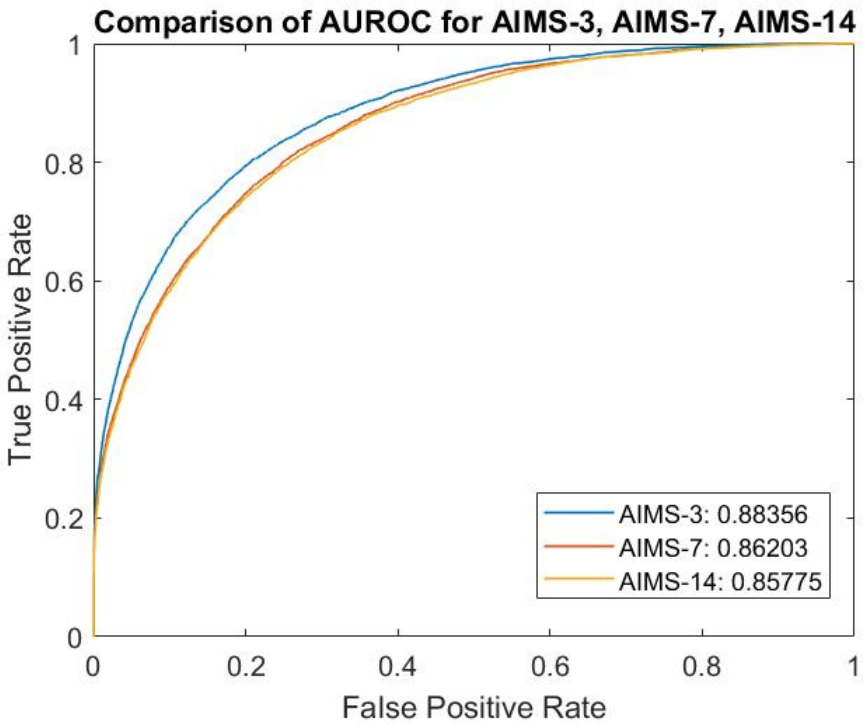


Figure 5.3: Comparison of ROCs for all AIMS schemes

is realistic and suitable for use as a diagnostic tool. The highest variance in AUROCs across folds is seen in AIMS-14, which is to be expected given that accurate prediction becomes more challenging across longer windows. However, the variance is still very low and all folds achieved strong AUROC values.

Table 5.4: AUROC statistics over 10 folds

Model	**AUROC**		
	Minimum	*Average*	*Maximum*
AIMS-3	0.8741	0.8835	0.8926
AIMS-7	0.8587	0.8619	0.8676
AIMS-14	0.8399	0.8577	0.8826

An alternative metric to the AUROC considered by some works in previous literature is the area under the precision-recall curve (AUPRC). This metric can be useful where data is imbalanced, as it was within our chosen database. Fig. 5.4 illustrates the precision-recall curves (PRCs) for each of the AIMS schemes. As is evident in Fig. 5.4, each of the AIMS models performs strongly and is well above the baseline. Once again, we see that AIMS-3 has the best curve, achieving

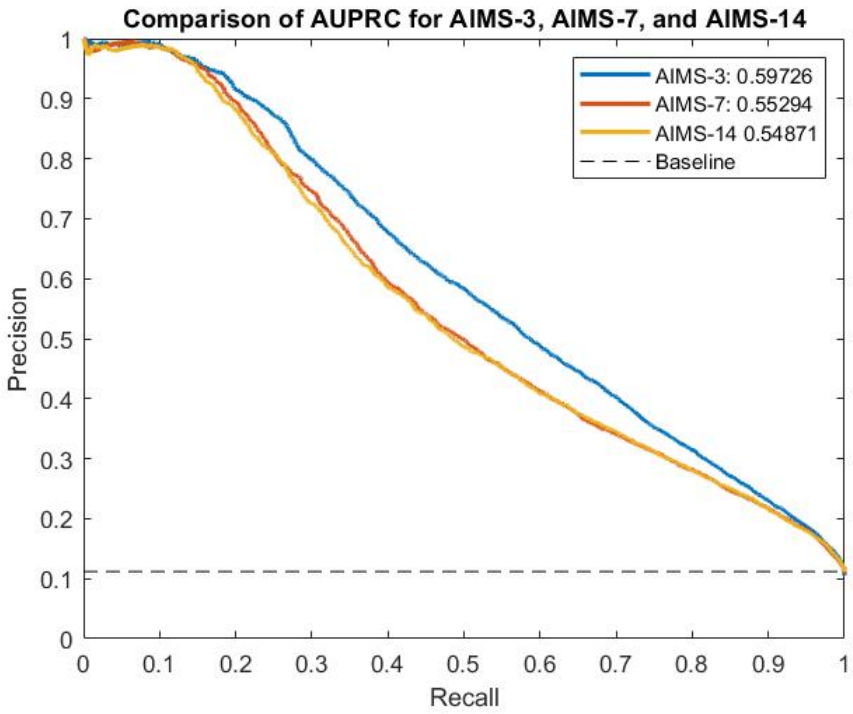


Figure 5.4: Comparison of PRCs for all AIMS schemes.

an average AUPRC of 0.5973 across 10 folds. Meanwhile AIMS-7 and AIMS-14 achieve AUPRCs of 0.5529 and 0.5487, respectively. This again indicates that there is more certainty regarding mortality risk prediction for smaller windows of time.

Other features that are important for any classification problem are accuracy (ACC), specificity or true negative rate (TNR), and sensitivity or true positive rate (TPR). ACC, TNR and TPR provide insight into the overall accuracy, the accuracy for the negative class, and the accuracy for the positive class, respectively. We summarise these parameters, as well as those of AUROC and AUPRC, in Table 5.5. Each of the presented value is the average, taken across the 10 folds of cross-validation.

Table 5.5: Results obtained by AIMS

Model	**ACC (%)**	**TNR**	**TPR**	**AUROC**	**AUPRC**
AIMS-3	80.07	0.802	0.792	0.884	0.597
AIMS-7	77.07	0.770	0.780	0.862	0.553
AIMS-14	76.22	0.765	0.779	0.858	0.549

Table 5.5 clearly shows that each of the networks achieves high accuracy. High accuracy alone cannot be used to assess the results where data is so heavily imbalanced, so sensitivity (TPR) and specificity (TNR) are also considered. Both of these parameters are measured between 0-1. High sensitivity and specificity indicate the ability of the network to correctly predict the mortality and non-mortality events, respectively. For each of the AIMS models, TPR and TNR are nearly equal, indicating that they can predict both the mortality and non-mortality cases with similar accuracy. Once again, AIMS-3 performs best in all categories, but AIMS-7 and AIMS-14 still show strong performance despite their longer prediction window.

5.3.1 Comparison to Previous Works

After performing thorough statistical analysis on AIMS-3, AIMS-7 and AIMS-14, we now consider how these networks perform compared to other state-of-the-art systems presented in the literature. The comprehensive Table 5.6 compares our three AIMS schemes to schemes presented by several recent papers. Unfortunately, many papers only presented AUROC, but where other metrics were available we have included these for comparison also. Accuracy and TNR have not been included, as these have unfortunately not been presented in any of the works we consider.

We include columns presenting the number of features, measurement window, and description of features. The description of features column is used to broadly explain what types of features were used by each network. 'Vital signs' refers to raw vital signs and any statistics derived from them, 'GCS' is the manually determined Glasgow Coma Score, 'demographics' refers to information about the patient's background, 'comorbidities' are existing diagnosed conditions, and 'medications' are those administered during the ICU stay. 'Laboratory results' refers to results obtained from blood, urine, and other laboratory analysis, including but not limited to: bilirubin, creatinine, hematocrit, blood urea nitrogen, white blood cell count, and many more.

Table 5.6: Performance of AIMS-3, AIMS-7, AIMS-14 and other schemes from the literature

Model	**No. Features**	**Description of Features**	**Measurement Window (hrs)**	**TPR**	**AUROC**	**AUPRC**
Alves [57]	37	Vital Signs, Laboratory Results	48 (from admission)	-	0.836	-
Delahanty [60]	17	APR-DRG Codes, MS-DRG Cost Index, GCS, Vital Signs, Laboratory Results	48 (24 hours pre- and post-ICU admission)	-	0.94	-
Deliberato [55] (Best Model)	14	Vital Signs, Demographics, GCS	Varies - 1 hour from admission, plus pre-admission data and SAPS-II	-	0.84	-
Deliberato [55] (Vitals Model)	6	Vital Signs	1 (from admission)	-	0.65	-
Johnson [59]	148	Vital Signs, GCS Laboratory Results	24 (from admission)	-	0.927	-
Thorsen-Meyer [131]	44	SAPS-III features (Vital Signs, GCS, Laboratory Values, Co-morbidities, Demographics, Patient History)	Various (from admission)	-	0.73-0.88	-

Continued on next page

Table 5.6 – continued from previous page

Model	No. Features	Description of Features	Measurement Window (hrs)	TPR	AUROC	AUPRC
Miao [35]	32	Demographics, Comorbidities, Laboratory Values, Medications	N/A - used first measurements after admission	-	0.821	-
Yu, K. [61]	Varies	Bag-of-words representation	48 (any window)	-	0.8854	0.3184
Yu, R. [56]	15	Vital Signs, GCS, Laboratory Results	24 (from admission)	0.503	-	0.520
Zahid [58]	79	Vital Signs, Laboratory Results, Demographics, GCS	24 (from admission)	-	0.86	-
AIMS-3	51	Age, Gender, Vital Signs	24 (any window)	0.792	0.884	0.597
AIMS-7	51	Age, Gender, Vital Signs	24 (any window)	0.780	0.862	0.553
AIMS-14	51	Age, Gender, Vital Signs	24 (any window)	0.779	0.858	0.549

Our AIMS-3 and AIMS-7 networks exceed the performance of all other schemes except those presented by Johnson *et al.* [59], Delahanty *et al.* [60], Thorsen-Meyer *et al.* [131], and K. Yu *et al.* [61] in all measured metrics. Compared to the scheme presented by K. Yu *et al.* [61], our AIMS-3 scheme achieves an AUROC that is less than 0.002 lower, while AIMS-7 and AIMS-14 achieve slightly lower AUROCs. However, our models perform much more strongly in terms of the AUPRC. This indicates that our models have far stronger precision and recall, which in turn indicates a higher accuracy, and stronger performance on the mortality event class. Additionally, our scheme depends on far simpler features that are easily interpreted by the user.

The model proposed by Thorsen-Meyer *et al.* [131] for 90-day mortality from

admission time was based on iterative improvement to the measurement throughout the hospital stay. It is based on the same features as SAPS-III, with these features recalculated every hour and used to update the model's prediction. At admission, the AUROC was at its lowest - 0.73. This then steadily increased throughout the stay, with AUROC reaching 0.82 after 24 hours, then 0.85 after 72 hours, and finally 0.88 at the time of discharge from ICU. This is an interesting approach for long-term mortality prediction, however it differs from our own scheme in several ways. Firstly, our scheme provides mortality risk prediction within shorter windows, which is more suitable for supporting immediate treatment decisions. Accurate prediction of 90-day mortality is certainly commendable, but the width of the window would limit the usefulness in making treatment decisions during a patient's stay. Secondly, the dependence of the Thorsen-Meyer *et al.* model [131] on SAPS-III parameters introduces a higher burden on healthcare workers, with a number of pathological tests needing to be constantly re-run to keep this information up-to-date. Our scheme uses only vital signs, which are simple to record even without automatic equipment. As such, our scheme would place less additional burden on healthcare providers and would be more suitable for low-resource hospitals. Finally, the Thorsen-Meyer *et al.* model [131] has substantially lower AUROC during the early stages of admission, reaching only 0.82 with 24 hours of data. All three of our AIMS schemes achieve a higher AUROC using 24 hours of data. Furthermore, within 24 hours our AIMS-3 scheme achieves a marginally better AUROC than that of the Thorsen-Meyer *et al.* model [131] at time of discharge. Overall, the AIMS scheme would be more suitable for short-term mortality prediction in ICU environments.

The AUROC of our strongest network, AIMS-3, also compares favourably to the high-performing scheme presented by Johnson *et al.* [59]. Unfortunately there are no other metrics presented in this paper to which we can compare. However, it is clear that this model depends on a feature vector nearly three times larger than our own. Additionally, the feature vector used by Johnson *et al.* [59] depended heavily upon laboratory values. Of the 148 variables considered, only 20 were vital signs - the rest were the results of laboratory tests

and the GCS. As such, it would be extremely difficult to implement this scheme in reality, and even more challenging to update it regularly during the stay, as medical professionals would have to undertake the laborious task of extensive data entry. Meanwhile, our scheme depends primarily on vital signs that are either automatically recorded or manually measured simply and regularly, greatly reducing the demand on healthcare workers.

Minimal statistics were presented by Delahanty *et al.* [60], but they do achieve the highest AUROC of the papers we consider. Unfortunately, it has many limitations. While it depends on 17 features directly, three of those features are based on manual diagnosis. As previously discussed, the APR-DRG Risk of Mortality and APR-DRG Severity of Illness are determined based on the diagnosis of the patient, which requires that all diagnoses are known, and also that the hospital uses this particular coding scheme. The scheme also demonstrated strong dependency on MS-DRG codes to determine the Medicare cost-weight index, and these codes are certainly not used in all hospitals. Despite depending on 17 features directly, the dependency of the scheme presented by Delahanty *et al.* [60] on these 3 diagnoses-based features means that there is a much greater true dependence based on the many parameters that are used to determine the diagnosis. Additionally, the scheme presented Delahanty *et al.* [60] depends on a 48-hour window including 24 hours pre-ICU admission and 24 hours following admission. Therefore, it could not be used in any window other than at ICU admission, and would likely not perform well where the patient is admitted directly to ICU without having first stayed elsewhere in the hospital. Conversely, our AIMS schemes depend on just 24 hours of data, and can be easily and automatically updated throughout the patient's stay.

Overall, our AIMS schemes - particularly AIMS-3 - perform strongly as opposed to the comparative schemes in the literature. Two papers - those by Johnson *et al.* [59] and Delahanty *et al.* [60] achieved higher AUROC, but presented no other statistics. Additionally, there are strong limiting factors that would prevent their adoption into healthcare environments. Meanwhile, our AIMS networks depend solely upon statistics derived from vital signs and two simple demographics - age and gender. These parameters are regularly recorded

in ICUs, often automatically. Additionally, the selection of vital signs as parameters would ensure that even low-resource healthcare environments would be able to utilise our scheme. The ease of measurement would also be extremely valuable in times of crises where a high number of patients may be admitted to intensive care, such as following a natural disaster or during a pandemic like COVID-19. Other schemes in the literature that rely on time-consuming and laborious collection of pathology results and/or patient histories would place high burden on an already strained system, making them challenging and impractical to use in such situations.

Our AIMS models also have the advantage of being easy to calculate in any 24-hour window due to the regular and often continuous recording of vital signs. This is a significant advantage over other schemes that have only considered calculation of mortality during a single window immediately following admission. We therefore conclude that our scheme is a strong candidate for predicting real-time mortality in ICU environments, however we acknowledge that this study has been conducted retrospectively on a single popular ICU database. We aim to further verify the performance of AIMS through clinical trials, with the aim of ensuring that it will perform as strongly on other patient populations. Further improvements will be made as necessary to ensure that the AIMS scheme performs equally well across all populations.

5.4 Conclusion

In this chapter, we have presented AIMS, a hybrid neural network structure that combines temporal convolution layers with long short-term memory layers. We have then trained and tested three instances of AIMS: AIMS-3, AIMS-7, and AIMS-14, which predict the risk of a mortality event within the following 3, 7, and 14 days, respectively.

AIMS-3 was the highest performing instance of the network, however AIMS-7 and AIMS-14 also perform strongly and compare well to other schemes in the literature. All three schemes could be used simultaneously in hospitals, giving healthcare workers a clear picture of both short-term and longer-term mortality

risk to the patient.

The AIMS scheme is dependent only on age, gender, and statistics derived from vital signs. These features are all readily and regularly recorded in ICU environments, with minimal effort required by healthcare workers. The simplicity of the features chosen also ensures that AIMS could be recalculated on a continuous and automatic basis during a patient's stay. This would provide invaluable information about whether a patient is responding to treatment or not, thus allowing medical professionals to modify their treatment plan more readily. Furthermore, the simplicity of both inputs and outputs to AIMS improves the interpretability of the overall model, thus improving the likelihood that healthcare providers would place their trust in its predictions.

One limiting factor for this work was that the MIMIC-III records for individual patients are not equally data-rich for all windows considered. This leads to some patients being included in the cohort for one or two schemes, but not the remainder. In turn, this prevented robust analysis of stability between the systems - that is, analysis of how similar the predictions of AIMS-3, AIMS-7, and AIMS-14 were for a single patient. We aim to address this in the clinical trial phase, where we will be able to ensure that data is recorded with consistent frequency throughout trials.

Overall, AIMS is an easy-to-interpret and powerful tool for mortality risk prediction in the ICU. The results presented in this chapter indicate that the three AIMS networks may be suitable for clinical implementation. In our own future works, we aim to conduct clinical trials using the AIMS networks to further analyse and improve upon its performance. We will also seek to assess the interpretability of the system during the clinical trial process, improving upon it as necessary.

The AIMS scheme addresses the research problem of mortality risk assessment in adults, as described in Section 1.2.3. It leads to the fourth original contribution of this work. Furthermore, this chapter shows that vital signs can be used to predict mortality risk, which informed the feature selection and overall methodology of the following and final research chapter.

Chapter 6

Non-invasive and Continuous Neonatal Mortality Risk Assessment using Respiratory and Heart Rate Variations

This chapter extends upon the adult mortality risk prediction scheme presented in Chapter 6, developing a neural network technique for mortality risk prediction in neonatal patients. A separate risk prediction scheme is necessary for this group, as 'normal' health parameters differ greatly between newborns and adults. The scheme presented in this chapter quantifies mortality risk for neonates using fewer vital signs and alternative age demographics that quantify the prematurity of the infant.

This chapter contains materials included in the following manuscript, which has been submitted to *Computers in Biology and Medicine*

[5] **S. Baker**, W. Xiang, and I. Atkinson, "Non-invasive and Continuous Neonatal Mortality Risk Assessment using Respiratory Rate and Heart Rate," in revisions with *Computers in Biology and Medicine*.

6.1 Introduction

Complications resulting from premature birth are the leading cause of death in children under 5 [132], and over 50% of neonatal deaths occur in preterm infants

[176]. Child deaths due to preterm birth are in excess of 1.1 million per year globally [133]. Recent data shows that preterm birth rates are increasing in 62 of the 65 countries with reliable trend data, indicating that this is a growing problem throughout the world.

Preterm infants are regularly cared for in Neonatal Intensive Care Units (NICUs). A recent study in the United States found that 84.41% of very low birthweight infants (those weighing 500-1499 g) and 41.18% of low birthweight infants (those weighing 1500-2499 g) are admitted to NICU, respectively [177]. In the NICU, assessment of mortality risk assists medical specialists in making difficult decisions regarding which treatments should be used and when, and whether initiated treatments are working effectively. It has been identified that precise mortality prediction would ease the process of making such decisions [178].

Currently, there are several scoring schemes used in NICUs for mortality risk assessment. One commonly used score is the updated Clinical Risk Index for Babies (CRIB-II) [134], which is a recalibrated and simplified iteration of the original CRIB score [135]. Another family of scores that are routinely used are the Score for Neonatal Acute Physiology (SNAP) [137] and it's derivatives, which include the expanded SNAP Perinatal Expansion (SNAPPE) [136], and the simplified versions of SNAP-II and SNAPPE-II [179]. The Berlin score [138] and Neonatal Mortality Prognostic Index (NMPI) [139] are also used, albeit to a lesser extent.

There are several limitations with the existing scores. Firstly, all aforementioned scores include parameters that require complex manual measurement. CRIB-II is the simplest and relies on just five parameters, however one of these is base excess. The small number of variables is good for quick calculation, however may limit the ability of the CRIB-II score to identify response to treatment throughout the NICU stay. The SNAP-II is more complex to calculate, relying on parameters such as PO_2/FiO_2, serum pH, presence of seizures, and urine output. SNAPPE-II expands SNAP-II by adding gestational age, birthweight, and Apgar score. The Berlin score includes similar complex variables to CRIB-II and SNAPPE-II, including Apgar score, base excess, severity of respiratory distress syndrome, and use of artificial ventilation. Finally, NMPI utilises variables

including PO_2/FiO_2, congenital malformations, base excess, and septicaemia.

Furthermore, these scores were all developed over 15 years ago. The most recent score is CRIB-II, which was formulated in 2003. There have been significant advancements in neonatal intensive care thereafter, and a recent extensive review of the scoring systems has identified the need for updated and enhanced scores based on more recent cohorts [31]. This is further highlighted by a recent study [140], which found that the SNAPPE-II score achieved an AUROC of 0.849 on babies admitted to a Bangladesh hospital between 2012-2013. This is significantly lower than the AUROC of 0.91 that was reported in the 2001 paper in which SNAPPE-II was proposed [179]. Similarly, a recent paper [178] conclude that CRIB-II does not adequately account for advancements in neonatal care, after finding that CRIB-II achieved AUROCs of 0.667 and 0.708 for mortality cases in $\leq$7 days >7 days, respectively on babies admitted to the NICU of Samsung Medical Centre between 2001-2011. This is undoubtedly a drastic decrease from the AUROC of 0.92 reported in the 2003 paper that proposed CRIB-II [134].

With recent studies identifying the weaknesses of existing scores, there has some renewed interest in developing updated neonatal mortality risk scores using new techniques, however this field is in its infancy when compared to the field of adult mortality risk prediction. Several studies have done this using techniques including logistic regression [141, 142], densely-connected neural networks [63], random forest [143], and fusion of multiple machine learning algorithms into a superlearner [62].

In one work [141], neonatal mortality risk prediction is considered for three cases: pre-birth, at start of delivery, and 5 minutes post-birth. The latter was the post-birth scenario, where variables used include mode of delivery, delivery complications, size of baby, condition of the baby at 5 minutes, and several more. Some of these variables are subjective, and were reported by mothers or family members. Logistic regression with these parameters showed reasonably good ability to distinguish between the mortality and non-mortality cases, as measured by the area under the receiver-operator curve (AUROC) of 0.85.

Another work [142] analysed 18 candidate variables to develop a logistic re-

gression model, ultimately selecting three parameters - birthweight, admission oxygen saturation, and highest level of respiratory support within 24 hour of birth. This relatively simple score achieved AUROCs of 0.8903 and 0.8082 on UK and Gambian cohorts, respectively. The simplicity of these scheme is a large advantage, however it focused only on babies who weighed less than 2000 g. Low birthweight is defined by WHO as <2500 g, so this study exclusionary of some babies who fall into this risk category.

Another recent work [143] considered multiple machine learning techniques, namely logistic regression, linear and quadratic discriminant analysis, k-nearest neighbor (KNN), support vector machine (SVM), random forest (RF), and three Gaussian processes. The features selected incorporated vital signs, birthweight, gestational age, and the SNAP-II and SNAPPE-II scores. The highest AUROC of 0.922 was achieved using the random forest classifier. This shows strong ability to distinguish between mortality and non-mortality cases, however the dependence on SNAP-II and SNAPPE-II leads to a dependence on the complex variables that these scores use.

Several machine learning techniques were also considered in another recent work [63], including logistic regression, KNN, RF, Gradient Boosting Machine, SVM, and densely-connected neural networks (NNs). Features used included birthweight, gestational age, and other basic demographics, as well as more complex variables such as presence of chorioamniotis, prenatal care, administration of antenatal steroids, maternal hypertension, and more. The neural network outperformed the other schemes, achieving an AUROC of 0.9136. This indicates that NNs are a stronger candidate for solving the problem of neonatal mortality than other machine learning techniques, however the complex variables used in this work limit the usefulness of the scheme. Choosing simpler features would greatly improve usability.

Another work [62] used a superlearner approach to predict mortality for post-operative neonatal patients, creating a fusion of 14 machine learning techniques to determine a best estimate. The algorithms included in the fusion were predominantly regression and RF algorithms. Extensive variables including demographics, existing conditions, prior treatments, congenital malformations, and

incidence of sepsis were used. This work achieved an AUROC of 0.91, again highlighting the strength of machine learning. However, the usefulness is once again limited by complexity of variables. This particular work is also specifically focused on postoperative neonates, rather than extending to all NICU patients.

The literature on predicting adult mortality is far more extensive, and many studies have investigated machine learning for prediction of mortality in adult ICU [35, 55–61], with most achieving reasonable ability to distinguish between mortality and non-mortality. Of particular interest are long short-term memory (LSTM) networks, which were identified to be suitable for mortality prediction and the related problem of sepsis prediction in several works [46, 55, 61]. Additionally, one work [57] identified that the hybridisation of LSTM networks with convolutional neural networks (CNN) can enhance predictive performance, although their own work was again based on complex variables.

A recurring limitation in the literature for both neonatal and adult mortality prediction is the selection of variables that are tedious or difficult to measure regularly. This limits the usefulness of such schemes, as often the acquisition of these parameters would increase the burden on neonatal healthcare staff. Conversely, several other studies were limited by their selection of variables that do not change - such as the scheme [142] that used birthweight, blood oxygen at admission, and respiratory support within the first 24 hours from birth. This prevents recalculation of the infant's risk on a continuous or ongoing basis, and does not allow for assessment of response to treatments.

An ideal mortality risk prediction scheme would be one that uses fundamental demographics and routinely measured parameters to provide continuous mortality risk assessment, allowing for assessment of changing risk throughout the NICU stay without placing unreasonable additional burden on NICU staff.

In this chapter, the Neonatal Artificial Intelligence Mortality Score (NAIMS) is proposed. NAIMS is a hybrid CNN-LSTM neural network that relies on simple demographics and trends in vital signs to determine mortality risk in the NICU for short- and long-term risk windows. Using 12 hours of data from any window, NAIMS shows strong performance in predicting an infant's risk of mortality within 3, 7, or 14 days. Due to the simplicity of the proposed scheme, NAIMS

could readily be continuously and automatically recalculated, enabling analysis of a NICU baby's responsiveness to treatment and other health trends.

The remainder of this chapter is structured as follows: Section 6.2. presents the methodology used for selecting, extracting, and processing data from an online, open-source database. It also discusses the structure of the NAIMS network. Section 6.3. includes results and discussion, including comparison to the aforementioned schemes for neonatal mortality risk prediction. Section 6.4. concludes the chapter.

6.2 Methodology

6.2.1 Data Selection

The data used in this study was obtained from the Medical Information Mart for Intensive Care (MIMIC-III) clinical database [151]. This database includes records from 7870 neonates admitted between 2001-2008. As this study focuses on all infants admitted to the NICU for any reason, the criterion used to select patients was that the first care unit was the NICU. No exclusions were made based on birthweight, gestational age, or other factors.

At this stage, length of stay for the mortality cases was evaluated to determine the most useful windows for mortality prediction. It was found that the average length of stay (LOS) was 8.08 days with a high standard deviation (SD) of 16.75 days. This served as the motivation for considering several risk windows of varying lengths, namely 3-day, 7-day, and 14-day windows. Assessment of mortality risk within these three windows would enable assessment of immediate risk, as well as longer-term survival prospects.

Given the largest window of interest is 14 days, there were 14 days of data acquired for each patient. If the NICU stay exceeded 14 days, the first 14 days were obtained for non-mortality cases, while the 14 days prior to death time were used for the mortality cases. Where any patient stay was less than 14 days, all data from NICU admission to discharge or death were obtained.

Information obtained from this database included gestational age, birthweight, gender, time of death (where applicable) and available chart events.

Feature selection was then performed, as discussed in the following subsection.

6.2.2 Feature Selection

In selecting features for the proposed NAIMS scheme, there were several major considerations. Firstly, to prevent placing additional burden on healthcare staff, it was determined that features must be based on parameters that are easy to measure. Ideally, dynamic parameters would also be able to be measured automatically. Secondly, feature selection was supported by recent findings in the literature.

Demographics features that describe fundamental information about the patient were selected - namely birthweight, gestational age, and gender. Birthweight and gestational age have repeatedly been shown in the literature to be strong indicators of mortality risk, and have been used by most existing schemes in the literature for this reason. Birthweight is a static variable, and thus the first birthweight in the patient's record was used. For most patients, gestational age was recorded in MIMIC-III as a range (i.e. 26-28) weeks. As such, took the middle value of the provided range was taken to be the gestational age. Where the infant was older than 40 weeks, their gestational age was recorded as "40" in MIMIC-III; thus any patients in this age group had their gestational age recorded as simply 40.

Sex has also been used as it has been long known that physiological differences between the genders lead to differing normal ranges for vital signs [180]. During preprocessing, the sex of babies was set to either '1' or '0', corresponding to the 'F' or 'M' classification in MIMIC-III, respectively.

Next, features were selected from commonly recorded parameters in the NICU. A recent comprehensive review paper [181] concluded that the current techniques of intermittent vital sign measurement fail to capture health trends, and that continuous analysis of vital sign trends would likely improve outcomes for NICU patients. Another work [182] identified that short-term variability of heart rate (HR) and respiratory rate (RR) are strong predictors of high morbidity. As such, focus was placed on HR and RR in this work. These two metrics are readily available in the MIMIC-III database, indicating that they are currently

recorded routinely and readily in NICU environments.

To capture information about the trends in these two vital signs, a 12-hour period at the beginning of the relevant risk window was selected. All HRs and RRs during this 12-hour window were recorded, and then statistical analysis was applied to quantify the variation of each vital sign during the 12-hour window.

For both HR and RR, the first value, last value, minimum value, maximum value, mean value, median value, and standard deviation was calculated for inclusion in the feature vector. The first and last values were chosen as these can highlight major changes in the vital sign during the 12-hour window. Minimum and maximum are used to show the most extreme values during the considered window. To represent the average vital sign, both mean and median were recorded. While mean is typically more useful, median is helpful in the case where there are significant outliers. Finally, the standard deviation is used as it is a strong indicator of variability. Where either HR or RR measurements were completely absent from a patient's record, that record was discarded and not used for training or testing.

The final feature array was as follows: *birthweight, gestational age, gender, first value for HR, last value for HR, minimum HR, maximum HR, mean HR, median HR, standard deviation of HR, first value for RR, last value for RR, minimum RR, maximum RR, mean RR, median RR, and standard deviation of RR.*

These features were calculated from the first 12 hours for each of the considered risk windows; 3-day, 7-day, and 14-day. Cohorts varied in size for each considered risk window, due to differing levels of missingness in the data for different windows.

6.2.3 Balancing the Dataset

Following data and feature selection, it was clear that the data was strongly unbalanced. In the cohort that met all criteria for inclusion in training and testing the 3-day NAIMS scheme, only 1.02% of the 2,751 cases ended in mortality. Similarly, for 7-day and 14-day NAIMS, the mortality rates were 1.02% of 2,751 cases and 1.09% of 2,753 cases, respectively. The level of imbalance can cre-

ate significant overfitting issues when training a neural network, and as such the non-mortality cases were undersampled by saving only 150 eligible non-mortality records.

Following undersampling, the mortality rate in the 3-day and 7-day cohorts was 15.64% of 179 cases, while for 14-day NAIMS the mortality rate was 16.47% of 181 cases. Further statistical analysis of the cohorts used for training and testing each version are outlined in Tables 6.1 and 6.2 below, with 3-day, 7-day, and 14-day names hereafter denoted as NAIMS-3, NAIMS-7 and NAIMS-14 respectively.

Table 6.1: Characteristics of patient cohort for NAIMS-3 and NAIMS-7

Characteristic	**All patients (n = 179)**	**Survived (n = 151)**	**Died (n = 28)**
Birthweight (kg)	2.13 (0.46-4.76)	2.27 (0.61-4.76)	1.39 (0.46-3.64)
Female	77 (43.02%)	71 (47.02%)	6 (21.43%)
Gestational age at birth (weeks)			
≤24	16 (8.94%)	5 (3.31%)	11 (39.29%)
25-28	22 (12.29%)	14 (9.27%)	8 (28.57%)
29-32	15 (8.38%)	13 (8.61%)	2 (7.14%)
33-36	87 (48.60%)	85 (56.29%)	2 (7.14%)
≥40	39 (21.79%	34 (22.52%)	5 (17.86%)

Table 6.2: Characteristics of patient cohort for NAIMS-14

Characteristic	**All patients (n = 181)**	**Survived (n = 151)**	**Died (n = 30)**
Birthweight (kg)	2.12 (0.46-4.76)	2.27 (0.61-4.76)	1.37 (0.46-3.64)
Female	79 (43.65%)	71 (47.02%)	8 (26.67%)
Gestational age at birth (weeks)			
≤24	16 (8.84%)	5 (3.31%)	11 (36.67%)
25-28	24 (13.26%)	14 (9.27%)	10 (35.72%)
29-32	15 (8.29%)	13 (8.61%)	2 (6.67%)
33-36	87 (48.07%)	85 (56.29%)	2 (6.67%)
≥40	39 (21.55%	34 (22.52%)	5 (6.67%)

6.2.4 Neural Network Structure

Hybrid networks have previously been used in mortality prediction for adults in one work that used extensive laboratory values and vital signs over a 48-hour window, achieving reasonable AUROC of 0.834 [57]. While this shows good ability to distinguish between mortality and non-mortality cases, the dependence on long measurement windows and laboratory measurements limits the usability of the scheme for adult patients, let alone neonatal patients.

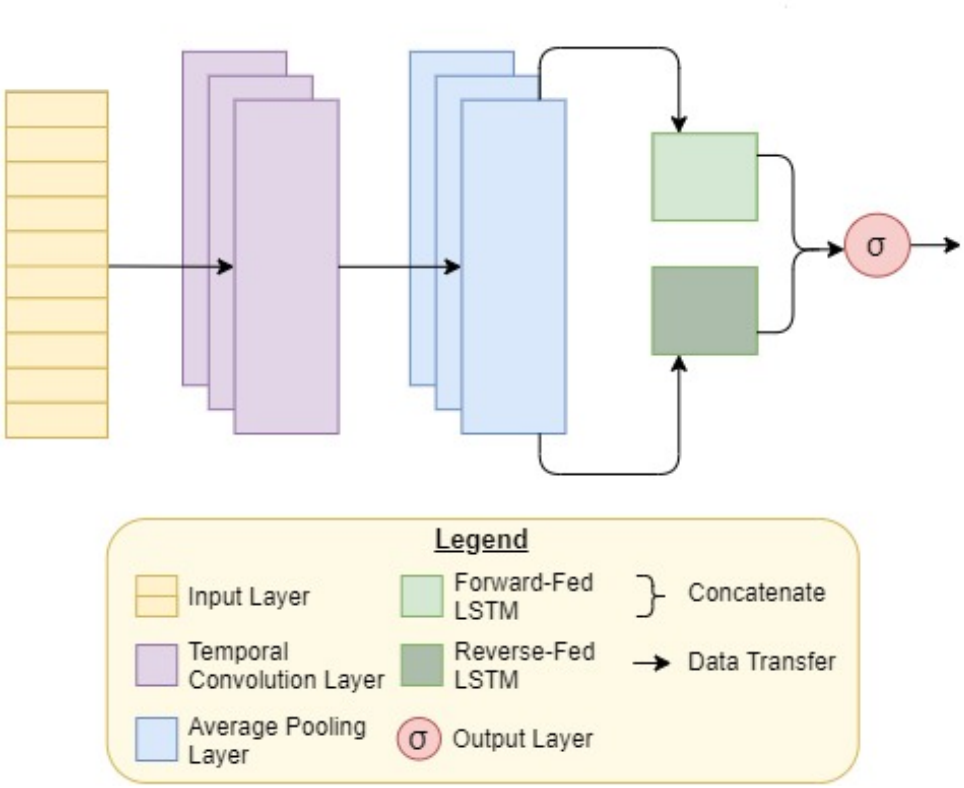


Figure 6.1: Neural network structure for NAIMS.

As such, in this chapter a shallow CNN-LSTM hybrid neural network is proposed, as illustrated in Fig. 6.1. The hybridisation of these two NN types combines the benefits of both. CNNs are well known for their ability to identify important features, while LSTM networks are well known for their ability to remember previous data in the sequence. These attributes are important in health applications, where the most important features are often not known. Deeper networks were trialled, however no improvement in performance was seen as a result of adding further layers. As such, the shallow network was used for

maximum efficiency.

The proposed NAIMS network uses the 17-feature vector outlined in the previous subsection as the input. This input vector is passed to the first layer, a temporal CNN layer with 128 hidden units that can be mathematically denoted as follows.

$$y_{\mathrm{j}}{}^{\mathrm{i}} = max(0, \sum_{n=1}^{N} w_{\mathrm{jn}}{}^{\mathrm{i}} * x_{\mathrm{m}}{}^{(\mathrm{i\text{-}1})} + b_{\mathrm{j}}{}^{\mathrm{i}}). \tag{6.1}$$

where y_j^i is the jth output feature map from the ith layer. The term $w_{\mathrm{jn}}{}^{\mathrm{i}}$, denotes the nth weight of the jth output feature map from the $(i-1)$th layer, with $n = 1, ..., N$. The bias term b_j is the jth bias term of the ith layer. Weights and biases are updated during training using the Adam optimization algorithm. The outputs of the $(i-1)$th layer are denoted as $x_{\mathrm{m}}{}^{(\mathrm{i\text{-}1})}$ represents the outputs of the $(i-1)$th layer. Finally, the convolution operation itself is denoted by the asterisk symbol $(*)$.

The temporal CNN layer is then followed by a temporal average pooling layer, with pool and stride sizes of 2. This operation steps through the output of the CNN layer and takes the average of each pool. This results in a undersampled output, which aids in prevention of overfitting without needing to use dropout.

The output from the pooling layer is then passed to a bidirectional LSTM (BiLSTM) layer with 128 hidden units. The mathematical structure of the layer is shown in Eqs. (6.2)-(6.7). As the layer is bidirectional, the data is passed through this mathematical process in both original and reversed orders. The benefit of bidirectionality is that the layer can learn from both past and future values in the sequence.

$$\tilde{c}_{\mathrm{t}} = \tanh(w_{\mathrm{c}}[\, a_{(\mathrm{t\text{-}1})}, x_{\mathrm{t}}] \, + b_{\mathrm{c}}) \tag{6.2}$$

$$f_{\mathrm{t}} = \sigma(w_{\mathrm{f}}[\, a_{(\mathrm{t\text{-}1})}, x_{\mathrm{t}}] \, + b_{\mathrm{f}}) \tag{6.3}$$

$$u_{\mathrm{t}} = \sigma(w_{\mathrm{u}}[\, a_{(\mathrm{t\text{-}1})}, x_{\mathrm{t}}] \, + b_{\mathrm{u}}) \tag{6.4}$$

$$o_{\mathrm{t}} = \sigma(w_{\mathrm{o}}[\, a_{(\mathrm{t\text{-}1})}, x_{\mathrm{t}}] \, + b_{\mathrm{o}}) \tag{6.5}$$

$$c_{\mathrm{t}} = u_{\mathrm{t}} \bullet \tilde{c}_{\mathrm{t}} + f_{\mathrm{t}} \bullet c_{(\mathrm{t\text{-}1})} \tag{6.6}$$

$$a_t = o_t \bullet \tanh(c_t) \tag{6.7}$$

where weights are indicated by w_c, w_f, w_u and w_o, respectively. Biases are indicated by b_c, b_f, b_u and b_o, respectively. Again, biases and weights are updated using the Adam optimization algorithm. Outputs of the previous layer are denoted as $a_{(t-1)}$, while x_t is the input to time t. Equations 6.6 and 6.7 are the updated cell state and layer output respectively. Element-wise multiplication is denoted by '•', while σ is the sigmoid activation function.

The final layer of NAIMS is a densely-connected node utilizing signmoid activation. Where the result of this activation is ≥ 0.5, the patient is predicted to die within the 3-day, 7-day, or 14-day window of the respective networks. Conversely, a result < 0.5 indicates survival for that period.

6.2.5 Training & Testing the Algorithms

The NAIMS network was trained using stratified k-fold cross-validation with 5 folds, a method that splits data in 5 different ways while ensuring consistent ratios of the positive to negative cases in each split. All data is used as part of the testing set in only one of the five folds. Results obtained via cross-validation provide a more realistic view of the network performance. Due to using five folds for cross-validation, 20% of the data was used for testing in each fold, and thus remained unseen to the network while training for that fold. For training and validation, 60% and 20% of the data was used, respectively.

Even after undersampling the non-mortality cases, the remaining data was unbalanced, with mortality occurring in up to 16.47% of cases. To prevent overfitting to the majority case of non-mortality, heavier weightings were placed on the importance of learning the mortality case until their relative importances were roughly equivalent. This ensured that the network would consider accurate prediction of the death and survival cases as equally important, which is essential to prevent overfitting to either case.

For each of the five folds, NAIMS was trained for 75 epochs with a batch size of 2048 with binary cross-entropy used as the loss function. This combination was found to enable good generalization. During each fold, the weights that resulted in the lowest validation loss were used for testing.

6.3 Results & Discussions

In analysing the performance of the NAIMS networks, the key metric considered was area under the receiver-operator curve (AUROC). AUROC is the most common metric used for analysing diagnostics tools, and is calculated from the receiver-operator curve (ROC). Higher AUROC values indicate stronger ability to distinguish between the mortality and non-mortality case. Fig. 6.1 plots the ROC curves for NAIMS-3, NAIMS-7 and NAIMS-14, with the AUROC shown in the legend. From this graph, it is clear that NAIMS-3 has achieved the highest AUROC, followed by NAIMS-7 and then NAIMS-14.

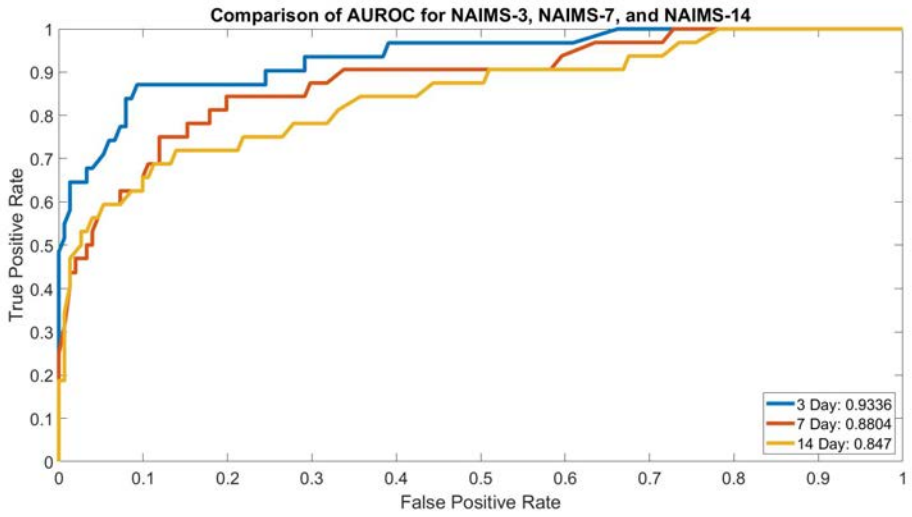


Figure 6.2: Comparison of ROCs for all NAIMS schemes.

Area under the precision-recall curve (AUPRC) is also often considered, particularly where data is imbalanced and the predictive performance on the positive cases is highly important. AUPRC is considered with respect to the performance of a baseline random classifier, which would vary in performance depending upon the imbalance of the data. The higher the AUPRC is above the random classifier, the better its ability to distinguish between the two classes. The precision-recall curves for all NAIMS schemes are shown in Fig. 6.3, with AUPRC values presented in the legend. This figure clearly shows that the AUPRC for NAIMS-3 is very strong, with NAIMS-7 and NAIMS-14 also performing quite strongly. Overall, this indicates that all three models are distingushing well between mortality and non-mortality cases.

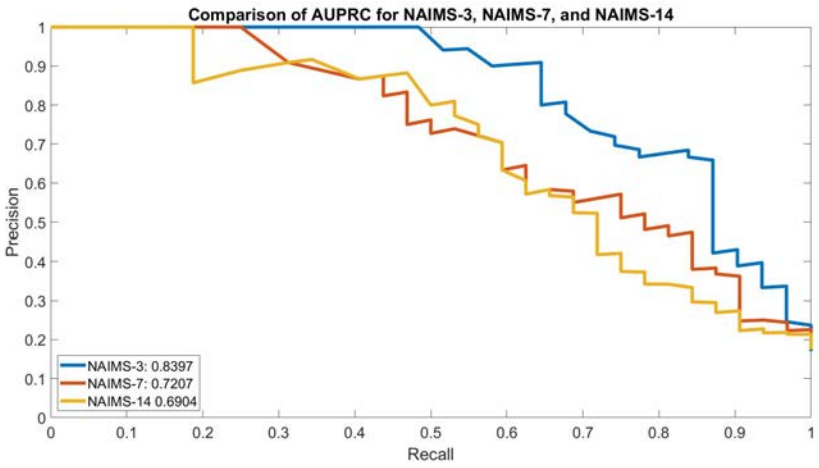


Figure 6.3: Comparison of PRCs for all NAIMS schemes.

The results obtained by NAIMS-3, NAIMS-7 and NAIMS-14 during the testing phase are further numerically summarised in Table 6.3. In addition to AUROC and AUPRC, several metrics were considered to evalutate the predictive accuracy. Overall accuracy (ACC), sensitivity or true positive rate (TPR), and specificity or true negative rate (TNR) were calculated for each version of NAIMS. TPR and TNR are useful metrics for an unbalanced data set, as they show the accuracy for the positive (mortality) and negative (survival) cases, respectively. This allows for analysis of the fit; if ACC, TPR and TNR are all similar values, then the network has fit equally well to both cases despite the imbalance of the data.

Table 6.3: Results obtains by NAIMS, presented as the average across the 5 folds with standard deviation in parantheses.

Scheme	**ACC (%)**	**TPR**	**TNR**	**AUROC**	**AUPRC**
NAIMS-3	86.81 (7.88)	0.8710 (0.0726)	0.8675 (0.0941)	0.9336 (0.0337)	0.8397 (0.0356)
NAIMS-7	80.33 (5.74)	0.8125 (0.0568)	0.8013 (0.0787)	0.8804 (0.0471)	0.7207 (0.1511)
NAIMS-14	73.77 (6.09)	0.7500 (0.0916)	0.7351 (0.0894)	0.8470 (0.0259)	0.6904 (0.0824)

As shown in Table 6.3, NAIMS-3 performed extremely strongly. NAIMS-3 achieved an excellent AUROC of 0.9336 with strong overall accuracy and highly

similar performance on both the positive and negative classes, as shown by the TPR and TNR values. Both NAIMS-7 and NAIMS-14 also achieved good results, with strong AUROC values and reasonable accuracy. Unsurprisingly, NAIMS-3 was the strongest performer across all metrics. This is likely due to it's shorter predictive window, as the patient is more likely to be showing signs of deterioration when mortality risk is assessed for a shorter window. Meanwhile NAIMS-14 had the lowest performance, likely due to the longer window. This trend further emphasises the need for continuous short-term mortality risk assessment as an alternative or compliment to the single mortality assessment that is routinely performed at admission time in NICUs today.

It is also worth noting that the models could readily be tuned to focus more heavily on predicting the positive case, however this would lead to reduced predictive performance for the negative case and thus an increase in false alarms. This work aimed for roughly equal ability to predict both the positive and negative case with the intention of reducing alarm fatigue, a well documented phenomenon in NICU and general hospital environments [183–185] wherein healthcare workers are overwhelmed by the large number of patient health alarms and thus become desensitized to them. Alarm fatigue leads to serious risk of missing significant alarms, which has previously lead to deaths in hospitals [184].

6.3.1 Comparison to Previous Works

In this section, the results achieved by the NAIMS schemes are compared to works presented in the literature. Table 6.4 compares the AUROCs of recent works, and includes descriptions of the features and measurement windows considered in each of the included works. The majority of previous works did not include ACC, TPR, TNR and/or AUPRC values in their analyses, so these have been excluded from the table.

Table 6.4: Performance of NAIMS-3, NAIMS-7, NAIMS-14 and other schemes from the literature

Scheme	**No. Features**	**Description of Features**	**Data Acquisition Window**	**Algorithm Type**	**AUROC**
Cooper [62]	284	Birth characteristics, laboratory test results, treatments received, existing conditions	Varied - all available data from the patient stay to time of mortality risk assessment used	Superlearner (14 ML algorithms)	0.91 -
Podda [63] (Best Model)	12	Birth characteristics, demographics, existing conditions, treatments received, maternal characteristics, maternal treatments received	Varied - used values regarding maternal health pre-birth, plus measurements from the first 5 minutes post-birth	Densely-Connected Neural Network	0.9136
Houweling [141] (Post-Birth Model)	10	Birth characteristics, maternal characteristics, condition of the baby by visual inspection	Varied - some information obtained pre-birth and during birth, plus 5 minutes post-birth	Logistic regression	0.85
Medvedev [142] (UK Cohort)	3	Birth weight, admission oxygen saturation, highest respiratory support within 24 hours	24 (from admission)	Logistic regression	0.8903

Continued on next page

Table 6.4 – continued from previous page

Scheme	No. Features	Description of Features	Data Acquisition Window	Algorithm Type	AUROC
Medvedev [142] (Gambia Cohort)	3	Birth weight, admission oxygen saturation, highest respiratory support within 24 hours	24 hours	Logistic regression	0.8082
Jaskari [143] (Best Model)	14	Vital signs, demographics, SNAP-II and SNAPPE-II scores	36 hours	Random Forest	0.922
NAIMS-3	17	Gestational age, birthweight, gender, vital signs	12 (any window)	CNN-LSTM	0.9336
NAIMS-7	17	Gestational age, birthweight, gender, vital signs	12 (any window)	CNN-LSTM	0.8804
NAIMS-14	17	Gestational age, birthweight, gender, vital signs	12 (any window)	CNN-LSTM	0.8470

The results presented in Table 6.4 indicate that NAIMS-3 outperforms all previous works in the literature, achieving a significantly higher AUROC than all previous works. This high-performing network also has several other advantages over existing schemes, including the ability to perform mortality risk assessment based on any 12-hour window of data during the patient's stay.

NAIMS-7 performs comparably to previous works, outperforming multiple schemes. NAIMS-14 performs comparably to the work presented by Houweling, et al. [141], however does not perform as strongly as much of the literature. It is likely that access to additional training data would improve the performance

of all NAIMS networks, and indeed this would be the next step required to work towards implementation of these schemes in real healthcare environments.

While the proposed NAIMS networks depend upon more features than some previous works, 14 of the 17 features are easily derived from temporal HR and RR data. Conversely, several works [63, 141] depend on variables that are completely distinct from each other and thus require more extensive acquisition and calculation. Furthermore, the work presented by Houweling, et al. [141] depends upon subjective metrics assessing the baby's appearance, rather than on tangiable measurements. In the work presented by Jaskari, et al. [143] only 14 variables are directly mentioned, however the dependence on SNAP-II and SNAPPE-II introduce many additional dependencies. Of the previous works included in Table 6.4, only the scheme presented by Jaskari, et al. [143] could be updated on an ongoing basis during the patient stay. All other works depend on variables that are static and are measured immediately post-birth.

Another significant advantage of all NAIMS schemes is the ability to be updated regularly and automatically, which allows for easier identification of trends in the patient's health. Gender and birthweight are fixed at birth, and gestational age could be automatically updated as the baby ages. HR and RR statistical values can be automatically calculated from monitoring equipment, or from manual data entries made by healthcare staff. Meanwhile, the only other scheme in the literature that is designed in such a way that mortality risk could be updated during the stay [143] still depends upon variables that would realistically make this challenging. Namely, it depends upon SNAP-II and SNAPPE-II scores which introduce a direct dependency on parameters such as PO_2/FiO_2, base excess, and urine output. Such parameters are substantially more complex to measure than vital signs, and thus would introduce a higher burden on healthcare workers.

Overall, the NAIMS schemes perform well when compared to existing schemes in the literature. In particular, NAIMS-3 outperforms all works in the literature, highlighting the benefit of shorter-term mortality risk assessment. NAIMS can be recalculated regularly and automatically, allowing for ongoing analysis of the patient's condition during the NICU stay. Furthermore, NAIMS uses a short, 12-hour window of temporal data to make its predictions, allowing the first pre-

diction to be made within half a day of admission, without needing knowledge of maternal condition prior to birth. As a result of these benefits and the strong ability to distinguish between mortality and non-mortality cases, it is suggested that the NAIMS schemes are suitable for use in predicting mortality risk in NICU environments. In particular, the NAIMS-3 scheme outperforms all existing works in the literature and would thus be highly suitable for use as a continuously-updating short-term mortality risk prediction tool.

6.4 Conclusion

In this chapter, the NAIMS shallow hybrid neural network is proposed, utilizing temporal convolution, pooling, and long short-term memory layers. NAIMS was then trained and tested for predicting mortality risk within the following 3, 7, and 14 day periods, resulting in NAIMS-3, NAIMS-7, and NAIMS-14, respectively.

It was shown that NAIMS-3 outperformed all other schemes in the literature, with NAIMS-7 and NAIMS-14 performing comparably to several state-of-the-art works. The high performance of NAIMS-3 indicates that this network would be suitable for use in NICU environments.

NAIMS also depends only upon simple features that are readily available in the NICU environment already. This simplicity enables regular and automatic recalculation of mortality risk during the stay, which in turn enables healthcare workers to monitor a patient's health trends and response to any treatments.

The primary limitation of this work was the low availability of data. While MIMIC-III contains extensive records for adult patients, relatively few neonatal patients were included. Furthermore, the mortality rate amongst NICU patients in MIMIC-III was 1.02%, resulting in a need to undersample the non-mortality cases to prevent overfitting in training. Nonetheless, this work serves as a proof-of-concept that neonatal mortality risk can be predicted from easily obtained vital signs in the NICU.

Overall, the NAIMS scheme performs strongly in mortality prediction for shorter risk windows. The presented results indicate that NAIMS-3 outperforms previous works in the literature, while NAIMS-7 and NAIMS-14 perform

comparably. The NAIMS schemes could readily be implemented in healthcare environments due to the high availability of the vital sign data it depends on in healthcare environments. Further clinical testing would be required to validate the performance of the network before widespread implementation could occur.

This chapter and the NAIMS scheme address the research problem of mortality risk assessment in a neonatal cohort, as discussed in Section 1.2.3. It also provides the fifth and final original contribution of this thesis. This chapter has further validated that vital sign information can be used to assess mortality risk across multiple critical care cohorts. It has also reiterated the strong performance of CNN-LSTM hybrid neural networks in healthcare applications.

Chapter 7

Conclusion

7.1 Summary

Patient outcomes in intensive care units can be vastly improved through continuous and non-invasive monitoring of vital signs and quantification of mortality risk. Existing algorithms in the literature for monitoring the vital signs of blood pressure and respiratory rates are unsuitable for clinical implementation, while existing schemes for determining patient risk are dependent on many complex health metrics that are challenging to recalculate throughout a patient's hospital stay. This thesis addresses these issues through the development of machine learning schemes for the measurement of blood pressure and respiratory rate from heart activity waveforms, before moving on to the assessment of mortality risk using vital sign data and basic demographics.

The research problem presented in Section 1.2.1 is addressed through the development of two schemes for non-invasive and continuous blood pressure measurement. Firstly, a hybrid CNN-LSTM neural network was developed for the calculation of SBP, DBP, and MAP using five-second segments of ECG and PPG waveforms as inputs. This minimised preprocessing and reduced the risk of introducing human bias to the neural network. The proposed scheme outperforms all other schemes in the literature, and meets the standards set by both the AAMI and BHS. Based on the results presented in this chapter, it is highly likely that this scheme would be suitable for clinical use.

Next, this work expands upon the investigation of suitable blood pressure monitoring techniques to develop a more computationally efficient alternative.

Twelve simple features that describe the structure of the ECG and PPG waveforms are extracted and used as features, minimising the risk of human bias impacting the learning of the neural network while improving efficiency. A shallow CNN-LSTM network was then used to predict SBP, DBP, and MAP. The performance of this scheme was slightly lower than that of the scheme presented in the previous chapter, however still performed strongly compared to the literature and comfortably met the AAMI and BHS standards for blood pressure devices. This scheme would serve as a suitable alternative in low-powered devices.

The second research problem, presented in Section 1.2.2, was considered next. A scheme was developed to measure respiratory rate continuously and non-invasively, with ECG and PPG signals again used to derive inputs. Respiratory rate modulates these heart activity signals in three ways, and thus each modulation was extracted from both signals. A respiratory quality index (RQI) tool was then used to quantify the quality of the extracted modulation waveform and a candidate respiratory rate was derived. Where the six candidate RRs and corresponding RQIs were used as inputs to a bidirectional LSTM neural network, the results outperformed all previous works. It was found that the inclusion of the proposed RQI scheme greatly enhanced the neural network's ability to learn from the data. Overall, the results of this chapter indicate that this scheme could be implemented in clinical environments.

This concluded investigations of enhanced vital sign monitoring. The proposed schemes for blood pressure and respiratory rate fill a significant gap in the literature; of the five vital signs, only respiratory rate and blood pressure were previously unable to be monitored continuously and non-invasively. The schemes presented in this thesis also depend only upon waveforms that are already recorded in hospitals. The ability to measure all five vital signs continuously has significant diagnostic benefit to environments such as intensive care, leading to the second theme of this thesis - mortality risk assessment in intensive care. This research problem was described in Section 1.2.3, with consideration of both adult and neonatal cohorts.

Mortality risk prediction for adult patients was considered first, with variations in vital signs over a 24-hour period quantified with straightforward statis-

tics and used as inputs to a CNN-LSTM hybrid neural network model. The model was trained to predict mortality risk within 3-day, 7-day and 14-day periods, enabling quantification of mortality risk within several risk windows. The proposed model performed strongly compared to the literature, indicating that vital signs offer viable metrics for the accurate prediction of mortality in critical care settings. The simplicity of the features ensures that mortality risk can be continuously updated throughout the stay, providing invaluable information about whether a patient is responding to treatment or not, thus allowing medical professionals to modify their treatment plan more readily.

Following the success of the adult mortality risk prediction scheme, neonatal mortality risk assessment was considered using a similar approach. A shallow CNN-LSTM neural network was developed and trained using information regarding the variation of heart rate and respiratory rate within a 12-hour period, with the scheme for predicting mortality over a 3-day window outperforming all existing works in the literature. The scheme also performed strongly when quantifying 7-day and 14-day mortality risk. This further indicates that vital signs are suitable candidates for mortality prediction, and that short-term mortality risk prediction is more suitable than existing schemes based on whole-stay mortality risk assessment at time of admission.

This concluded the investigation on mortality risk assessment, filling a significant gap in the literature by devising first-of-a-kind schemes for continuously updating mortality risk prediction that relies only on vital sign features. Previous schemes in the literature were non-continuous and depended on complex laboratory testing, limiting their usefulness in the real world.

Overall, this thesis has addressed the research problems outlined in Section 1.2 through developing enhanced methods for non-invasive measurement of blood pressure and respiratory rate, enabling all five vital signs to now be measured continuously. This thesis then investigated the use of vital signs for prognostic purposes, namely quantifying the severity of illness in critical care patients through mortality risk assessment. With the implementation of the proposed techniques for measuring vital signs, the mortality risk prediction schemes would become increasingly powerful.

The significance of this work lies in the development of tools that address key needs in the medical industry. Vital signs are key indicators of overall health, and this work presents novel and accurate methods for measuring these parameters non-invasively and continuously from sensors which are readily available in clinical settings and wearable devices. The importance of measuring vital signs accurately and continuously is further emphasised by the successful development of prognostics tools for mortality risk assessment presented in this thesis. The mortality risk prediction schemes are also significant in the healthcare field. The schemes proposed in this work require little to no manual input from healthcare workers, minimizing time wasted on data entry and analysis. They also would improve patient experience, as it eliminates the dependency on extensive laboratory tests that is present in existing schemes. Overall, this work has significant implications for device manufacturers, healthcare staff, and end-users in applications ranging from at-home fitness tracking to critical care units.

7.1.1 Recommendations for Future Work

Healthcare is a broad field, and while this thesis has addressed many key gaps in the literature, there are certainly still opportunities for future work. Suggested areas for future directions include:

- *Clinical trials for the developed algorithms* - the blood pressure and respiratory rate measurement techniques presented in this thesis are dependent on ECG and PPG, both of which are readily available in hospital settings. Conducting clinical trials in hospital settings is a vital but significant task in moving towards real-world implementation of the proposed algorithms.

- *Development of wearable hardware devices* - the development and testing of highly wearable devices implementing ECG and PPG sensors would enable the use of these algorithms in applications ranging from fitness tracking to telehealth, and would enable continuous health monitoring outside of hospital settings.

- *Investigation of multiple PPG as a replacement for ECG and PPG* - while ECG is acquirable via wearables, it is more challenging to do so contin-

uously than it is to measure PPG waveforms continuously. Therefore it would be worthwhile exploring whether multiple PPG signals, perhaps acquired from different locations on the body, could be used to achieve comparable performance to the current combination of ECG and PPG. This would require extensive hardware design and data acquisition.

Bibliography

[1] S. Baker *et al.*, "Internet of Things for Smart Healthcare: Technologies, Challenges, and Opportunities," *IEEE Access*, vol. 5, pp. 26 521–26 544, 2017.

[2] S. Baker *et al.*, "Continuous and automatic mortality risk prediction using vital signs in the intensive care unit: a hybrid neural network approach," *Sci. Rep.*, vol. 10, no. 1, p. 21282, 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-78184-7

[3] S. Baker *et al.*, "Determining respiratory rate from photoplethysmogram and electrocardiogram signals using respiratory quality indices and neural networks," *PLOS ONE*, vol. 16, no. 4, p. e0249843, apr 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0249843

[4] S. Baker *et al.*, "A Hybrid Neural Network for Continuous and Non-Invasive Estimation of Blood Pressure from Raw Electrocardiogram and Photoplethysmogram Waveforms," *Comput. Methods Programs Biomed.*, p. 106191, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260721002650

[5] S. Baker *et al.*, "Non-invasive and continuous neonatal mortality risk assessment using respiratory rate and heart rate," *Submitted to The Journal of Paediatrics*, 2020.

[6] Australian Institute of Health and Welfare, "Australia's Health," 2014. [Online]. Available: http://www.aihw.gov.au/WorkArea/DownloadAsset.aspx-?id=60129548150

[7] E. Perrier, *Positive Disruption: Healthcare, Ageing & Participation in the Age of Technology.* Australia: The McKell Institute, 2015.

[8] Victorian Intensive Care Data Review Committee, "Ten years of intensive care in Victoria (2001-02 to 2010-11)," State of Victoria, Melbourne, Tech. Rep., 2014. [Online]. Available: https://www2.health.vic.gov.au/about/publications/researchandreports/Ten-years-of-intensive-care-in-Victoria-2001-02-to-2010-11

[9] UCSF Philip R. Lee Insititute for Health Policy Studies, "ICU Outcomes," University of California San Francisco, Tech. Rep., 2011. [Online]. Available: https://healthpolicy.ucsf.edu/icu-outcomes

[10] E. West *et al.*, "Nurse staffing, medical staffing and mortality in Intensive Care: An observational study," *International Journal of Nursing Studies*, vol. 51, no. 5, pp. 781–794, Mar. 2014. [Online]. Available: https://doi.org/10.1016/j.ijnurstu.2014.02.007

[11] P. Hicks *et al.*, "The financial cost of intensive care in Australia: a multicentre registry study," *Medical Journal of Australia*, vol. 211, no. 7, pp. 324–325, Oct. 2019. [Online]. Available: https://doi.org/10.5694/mja2.50309

[12] Australian Institute of Health and Welfare, "Australia's mothers and babies data visualisations," Australian Institute of Health and Welfare, Canberra, ACT, Tech. Rep., 2020. [Online]. Available: https://www.aihw.gov.au/reports/mothers-babies/australias-mothers-babies-data-visualisations

[13] W. Karlen *et al.*, "Improving the accuracy and efficiency of respiratory rate measurements in children using mobile devices," *PLoS ONE*, vol. 9, no. 6, pp. e99 266–e99 266, Jun. 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0099266

[14] A. S. Ginsburg *et al.*, "A Systematic Review of Tools to Measure Respiratory Rate in Order to Identify Childhood Pneumonia," *American Journal of Respiratory and Critical Care Medicine*, vol. 197, no. 9, pp.

1116–1127, Feb. 2018. [Online]. Available: https://doi.org/10.1164/rccm.201711-2233CI

[15] J. Hogan, "Why don't nurses monitor the respiratory rates of patients?" *British Journal of Nursing*, vol. 15, no. 9, pp. 489–492, May 2006. [Online]. Available: https://doi.org/10.12968/bjon.2006.15.9.21087

[16] K. Philip *et al.*, "Staff perceptions of respiratory rate measurement in a general hospital," *British Journal of Nursing*, vol. 22, no. 10, pp. 570–574, May 2013. [Online]. Available: https://doi.org/10.12968/bjon.2013.22.10.570

[17] A. Hendrich, "A 36-Hospital Time and Motion Study: How Do Medical-Surgical Nurses Spend Their Time?" *The Permanente Journal*, vol. 12, no. 3, pp. 25–34, Jan. 2008. [Online]. Available: https://doi.org/10.7812/tpp/08-021

[18] S. Romagnoli *et al.*, "Accuracy of invasive arterial pressure monitoring in cardiovascular patients: an observational study," *Critical Care*, vol. 18, no. 6, p. 644, Nov. 2014. [Online]. Available: http://doi.org/10.1186/s13054-014-0644-4

[19] C. Park and B. Lee, "Compressed estimation of heart and respiratory rates from a photoplethysmogram," in *Proc. 2017 IEEE Biomedical Circuits and Systems Conference*, Turin, Italy, Mar. 2018, pp. 1–4.

[20] S. Boccia *et al.*, "What Other Countries Can Learn From Italy During the COVID-19 Pandemic," *JAMA Internal Medicine*, vol. 180, no. 7, pp. 927–928, Jul. 2020. [Online]. Available: https://doi.org/10.1001/jamainternmed.2020.1447

[21] H. Salje *et al.*, "Estimating the burden of SARS-CoV-2 in France," *Science*, vol. 369, no. 6500, pp. 208 – 211, Jul. 2020. [Online]. Available: http://doi.org/10.1126/science.abc3517

[22] D. Ponce, "The impact of coronavirus in Brazil: politics and the

pandemic," *Nature Reviews Nephrology*, Jul. 2020. [Online]. Available: https://doi.org/10.1038/s41581-020-0327-0

[23] J. Willan *et al.*, "Challenges for NHS hospitals during COVID-19 epidemic," *BMJ*, vol. 368, p. m1117, Mar. 2020. [Online]. Available: http://doi.org/10.1136/bmj.m1117

[24] I. F. Miller *et al.*, "Disease and healthcare burden of COVID-19 in the United States," *Nature Medicine*, Jun. 2020. [Online]. Available: https://doi.org/10.1038/s41591-020-0952-y

[25] J. Xie *et al.*, "Critical care crisis and some recommendations during the COVID-19 epidemic in China," *Intensive Care Medicine*, vol. 46, no. 5, pp. 837–840, Mar. 2020. [Online]. Available: https://doi.org/10.1007/s00134-020-05979-7

[26] E. Litton *et al.*, "Surge capacity of intensive care units in case of acute increase in demand caused by COVID-19 in Australia," *Medical Journal of Australia*, vol. 212, no. 10, pp. 463–467, Jun. 2020. [Online]. Available: https://doi.org/10.5694/mja2.50596

[27] M. Vergano *et al.*, "Clinical ethics recommendations for the allocation of intensive care treatments in exceptional, resource-limited circumstances: the Italian perspective during the COVID-19 epidemic," *Critical Care*, vol. 24, no. 1, p. 165, Apr. 2020. [Online]. Available: https://doi.org/10.1186/s13054-020-02891-w

[28] E. J. Emanuel *et al.*, "Fair Allocation of Scarce Medical Resources in the Time of Covid-19," *New England Journal of Medicine*, vol. 382, no. 21, pp. 2049–2055, May 2020. [Online]. Available: https://doi.org/10.1056/NEJMsb2005114

[29] T. W. Farrell *et al.*, "Rationing Limited Healthcare Resources in the COVID-19 Era and Beyond: Ethical Considerations Regarding Older Adults," *Journal of the American Geriatrics Society*, vol. 68, no. 6, pp. 1143–1149, Jun. 2020. [Online]. Available: https://doi.org/10.1111/jgs.16539

[30] A. A. Kramer, "Predictive mortality models are not like fine wine," *Critical Care*, vol. 9, no. 6, pp. 636–637, Oct. 2005. [Online]. Available: https://doi.org/10.1186/cc3899

[31] B. Garg *et al.*, "Assessment of sickness severity of illness in neonates: review of various neonatal illness scoring systems," *The Journal of Maternal-Fetal & Neonatal Medicine*, vol. 31, no. 10, pp. 1373–1380, May 2018. [Online]. Available: https://doi.org/10.1080/14767058.2017.1315665

[32] E. Baldi *et al.*, "Out-of-Hospital Cardiac Arrest during the Covid-19 Outbreak in Italy," *New England Journal of Medicine*, vol. 383, no. 5, pp. 496–498, Apr. 2020. [Online]. Available: https://doi.org/10.1056/NEJMc2010418

[33] M. Kachuee *et al.*, "Cuffless Blood Pressure Estimation Algorithms for Continuous Health-Care Monitoring," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 4, pp. 859–869, Apr. 2017. [Online]. Available: http://doi.org/10.1109/TBME.2016.2580904

[34] S. S. Mousavi *et al.*, "Blood pressure estimation from appropriate and inappropriate PPG signals using A whole-based method," *Biomedical Signal Processing and Control*, vol. 47, pp. 196–206, Jan. 2019. [Online]. Available: https://doi.org/10.1016/j.bspc.2018.08.022

[35] F. Miao *et al.*, "Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest," *IEEE Access*, vol. 6, pp. 7244–7253, Jan. 2018. [Online]. Available: https://doi.org/10.1109/ACCESS.2018.2789898

[36] F. Miao *et al.*, "Continuous Blood Pressure Measurement from One-Channel Electrocardiogram Signal Using Deep-Learning Techniques," *Artificial Intelligence in Medicine*, p. 101919, Aug. 2020. [Online]. Available: https://doi.org/10.1016/j.artmed.2020.101919

[37] K. Song *et al.*, "Cuff-less Deep Learning-Based Blood Pressure Estimation for Smart Wristwatches," *IEEE Transactions on Instrumentation and*

Measurement, pp. 4292–4302, 2019. [Online]. Available: https://doi.org/10.1109/tim.2019.2947103

[38] I. Sharifi *et al.*, "A novel dynamical approach in continuous cuffless blood pressure estimation based on ECG and PPG signals," *Artificial Intelligence in Medicine*, vol. 97, pp. 143–151, Jun. 2019. [Online]. Available: https://doi.org/10.1016/j.artmed.2018.12.005

[39] C. Orphanidou, "Derivation of respiration rate from ambulatory ECG and PPG using Ensemble Empirical Mode Decomposition: Comparison and fusion," *Computers in Biology and Medicine*, vol. 81, pp. 45–54, Dec. 2017. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2016.12.005

[40] W. Karlen *et al.*, "Estimation of respiratory rate from photoplethysmographic imaging videos compared to pulse oximetry," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 4, pp. 1331–1338, May 2015. [Online]. Available: https://doi.org/10.1109/JBHI.2015.2429746

[41] A. M. Chan *et al.*, "Ambulatory respiratory rate detection using ECG and a triaxial accelerometer," in *Proc. of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Osaka, Japan, Sep. 2013, pp. 4058–4061. [Online]. Available: https://doi.org/10.1109/EMBC.2013.6610436

[42] M. A. F. Pimentel *et al.*, "Toward a Robust Estimation of Respiratory Rate From Pulse Oximeters," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1914–1923, Aug. 2017. [Online]. Available: https://doi.org/10.1109/TBME.2016.2613124

[43] D. A. Birrenkott *et al.*, "A Robust Fusion Model for Estimating Respiratory Rate From Photoplethysmography and Electrocardiography," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 9, pp. 2033–2041, Sep. 2018. [Online]. Available: https://doi.org/10.1109/TBME.2017.2778265

[44] S. Khreis *et al.*, "Breathing rate estimation using Kalman smoother with electrocardiogram and photoplethysmogram," *IEEE Transactions on*

Biomedical Engineering, vol. 67, no. 3, pp. 893–904, Mar. 2020. [Online]. Available: https://doi.org/10.1109/TBME.2019.2923448

[45] M. Pirhonen and A. Vehkaoja, "Fusion enhancement for tracking of respiratory rate through intrinsic mode functions in photoplethysmography," *Biomedical Signal Processing and Control*, vol. 59, p. 101887, Feb. 2020. [Online]. Available: https://doi.org/10.1016/j.bspc.2020.101887

[46] J. Fagerström *et al.*, "LiSep LSTM: A Machine Learning Algorithm for Early Detection of Septic Shock," *Scientific Reports*, vol. 9, no. 1, p. 15132, Oct. 2019. [Online]. Available: https://doi.org/10.1038/s41598-019-51219-4

[47] S. P. Shashikumar *et al.*, "Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics," *Journal of Electrocardiology*, vol. 50, no. 6, pp. 739–743, Aug. 2017. [Online]. Available: https://doi.org/10.1016/j.jelectrocard.2017.08.013

[48] C. R. Yee *et al.*, "A Data-Driven Approach to Predicting Septic Shock in the Intensive Care Unit," *Biomedical Informatics Insights*, vol. 11, p. 117822261988514, Nov. 2019. [Online]. Available: https://doi.org/10.1177/1178222619885147

[49] F. E. Shamout *et al.*, "Deep Interpretable Early Warning System for the Detection of Clinical Deterioration," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 437–446, Feb. 2020. [Online]. Available: https://doi.org/10.1109/JBHI.2019.2937803

[50] D. A. Kaji *et al.*, "An attention based deep learning model of clinical events in the intensive care unit," *PLoS ONE*, vol. 14, no. 2, pp. e0 211 057–e0 211 057, Feb. 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0211057

[51] X. Zhai and C. Tin, "Automated ECG Classification Using Dual Heartbeat Coupling Based on Convolutional Neural Network," *IEEE Access*, vol. 6, pp. 27 465–27 472, May 2018. [Online]. Available: https://doi.org/10.1109/ACCESS.2018.2833841

[52] X. Fan *et al.*, "Multi-Scaled Fusion of Deep Convolutional Neural Networks for Screening Atrial Fibrillation from Single Lead Short ECG Recordings," *IEEE Journal of Biomedical and Health Informatics*, pp. 1744–1753, Nov. 2018. [Online]. Available: https://doi.org/10.1109/JBHI.2018.2858789

[53] S. Kiranyaz *et al.*, "Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 664–675, Mar. 2016. [Online]. Available: https://doi.org/10.1109/TBME.2015.2468589

[54] F. Pan *et al.*, "Variation of the Korotkoff Stethoscope Sounds During Blood Pressure Measurement: Analysis Using a Convolutional Neural Network," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 6, pp. 1593–1598, Nov. 2017. [Online]. Available: https://doi.org/10.1109/JBHI.2017.2703115

[55] R. O. Deliberato *et al.*, "SEVERITAS: An externally validated mortality prediction for critically ill patients in low and middle-income countries," *International Journal of Medical Informatics*, vol. 131, p. 103959, Sep. 2019. [Online]. Available: http://doi.org/10.1016/j.ijmedinf.2019.103959

[56] R. Yu *et al.*, "Using a Multi-Task Recurrent Neural Network with Attention Mechanisms to Predict Hospital Mortality of Patients," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 486–492, Feb. 2020. [Online]. Available: https://doi.org/10.1109/JBHI.2019.2916667

[57] T. Alves *et al.*, "Dynamic Prediction of ICU Mortality Risk Using Domain Adaptation," in *Proc. 2018 IEEE International Conference on Big Data*, Seattle, WA, USA, Jan. 2019, pp. 1328–1336. [Online]. Available: https://doi.org/10.1109/BigData.2018.8621927

[58] M. A. Zahid and J. Lee, "Mortality prediction with self normalizing neural networks in intensive care unit patients," in *Proc. 2018 IEEE EMBS International Conference on Biomedical and Health Informatics*, Las Vegas, NV, USA, Jan. 2018, pp. 226–229. [Online]. Available: https://doi.org/10.1109/BHI.2018.8333410

[59] A. E. Johnson and R. G. Mark, "Real-time mortality prediction in the Intensive Care Unit," *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2017, pp. 994–1003, 4 2017. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29854167https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977709/

[60] R. J. Delahanty *et al.*, "Development and Evaluation of an Automated Machine Learning Algorithm for In-Hospital Mortality Risk Adjustment Among Critical Care Patients," *Critical care medicine*, vol. 46, no. 6, pp. e481–e488, Jun. 2018. [Online]. Available: https://doi.org/10.1097/CCM.0000000000003011

[61] K. Yu *et al.*, "Monitoring ICU Mortality Risk with A Long Short-Term Memory Recurrent Neural Network," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, vol. 25, pp. 103–114, 2020. [Online]. Available: https://doi.org/10.1142/9789811215636{_}0010

[62] J. N. Cooper *et al.*, "Postoperative neonatal mortality prediction using superlearning." *The Journal of Surgical Research*, vol. 221, pp. 311–319, Aug. 2017. [Online]. Available: https://doi.org/10.1016/j.jss.2017.09.002

[63] M. Podda *et al.*, "A machine learning approach to estimating preterm infants survival: development of the Preterm Infants Survival Assessment (PISA) predictor," *Scientific Reports*, vol. 8, no. 1, p. 13743, Sep. 2018. [Online]. Available: https://doi.org/10.1038/s41598-018-31920-6

[64] J. Kellett and F. Sebat, "Make vital signs great again - A call for action," *European Journal of Internal Medicine*, vol. 45, pp. 13–19, Sep. 2017. [Online]. Available: https://doi.org/10.1016/j.ejim.2017.09.018

[65] J. Ženko *et al.*, "Pulse rate variability and blood oxidation content identification using miniature wearable wrist device," *Proc. 2016 International Conference on Systems, Signals and Image Processing*, Jul. 2016. [Online]. Available: https://doi.org/10.1109/IWSSIP.2016.7502766

[66] E. E. Dooley *et al.*, "Estimating Accuracy at Exercise Intensities: A Comparative Study of Self-Monitoring Heart Rate and Physical Activity

Wearable Devices," *JMIR mHealth and uHealth*, vol. 5, no. 3, p. e34, Mar. 2017. [Online]. Available: http://doi.org/10.2196/mhealth.7043

[67] D. Giles *et al.*, "Validity of the Polar V800 heart rate monitor to measure RR intervals at rest," *European Journal of Applied Physiology*, vol. 116, no. 3, pp. 563–571, Mar. 2016. [Online]. Available: https://doi.org/10.1007/s00421-015-3303-9

[68] J. Claes *et al.*, "Validity of heart rate measurements by the Garmin Forerunner 225 at different walking intensities," *Journal of Medical Engineering & Technology*, vol. 41, no. 6, pp. 480–485, Aug. 2017. [Online]. Available: https://doi.org/10.1080/03091902.2017.1333166

[69] R. Delgado-Gonzalo *et al.*, "Evaluation of accuracy and reliability of PulseOn optical heart rate monitoring device," in *Proc. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Nov. 2015, pp. 430–433. [Online]. Available: https://doi.org/10.1109/EMBC.2015.7318391

[70] A. Henriksen *et al.*, "Using Fitness Trackers and Smartwatches to Measure Physical Activity in Research: Analysis of Consumer Wrist-Worn Wearables," *Journal of Medical Internet Research*, vol. 20, no. 3, pp. e110–e110, Mar. 2018. [Online]. Available: https://doi.org/10.2196/jmir.9157

[71] V. P. Rachim and W. Y. Chung, "Wearable Noncontact Armband for Mobile ECG Monitoring System," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 6, pp. 1112–1118, Dec. 2016. [Online]. Available: https://doi.org/10.1109/TBCAS.2016.2519523

[72] W. Von Rosenberg *et al.*, "Smart Helmet:Wearable Multichannel ECG and EEG," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 4, p. 2700111, Nov. 2016. [Online]. Available: https://doi.org/10.1109/JTEHM.2016.2609927

[73] E. Spanò *et al.*, "Low-Power Wearable ECG Monitoring System for Multiple-Patient Remote Monitoring," *IEEE Sensors Journal*, vol. 16,

no. 13, pp. 5452–5462, Jul. 2016. [Online]. Available: https://doi.org/10.1109/JSEN.2016.2564995

[74] P. Aqueveque *et al.*, "Monitoring Physiological Variables of Mining Workers at High Altitude," *IEEE Transactions on Industry Applications*, vol. 53, no. 3, pp. 2628–2634, May 2017. [Online]. Available: https://doi.org/10.1109/TIA.2017.2675360

[75] P. Narczyk *et al.*, "Precision human body temperature measurement based on thermistor sensor," *Pro. 2016 IEEE 19th International Symposium on Design and Diagnostics of Electronic Circuits and Systems*, pp. 1–5, Jun. 2016. [Online]. Available: https://doi.org/10.1109/DDECS.2016.7482451

[76] T. Nakamura *et al.*, "Development of flexible and wide-range polymer-based temperature sensor for human bodies," *Proc. 3rd IEEE EMBS International Conference on Biomedical and Health Informatics*, pp. 485–488, Apr. 2016. [Online]. Available: https://doi.org/10.1109/BHI.2016.7455940

[77] A. Eshkeiti *et al.*, "A novel self-supported printed flexible strain sensor for monitoring body movement and temperature," *Proc. 2014 IEEE Sensors*, vol. 2014-Decem, pp. 1615–1618, Dec. 2014. [Online]. Available: https://doi.org/10.1109/ICSENS.2014.6985328

[78] N. J. Talley and S. O'Connor, *Clinical Examination - A Systematic Guide to Physical Diagnosis*, 7th ed. Chatswood, NSW: Elsevier Australia, 2014.

[79] Heart Foundation, "High blood pressure statistics," 2017. [Online]. Available: www.heartfoundation.org.au/about-us/what-we-do/heart-disease-in-australia/high-blood-pressure-statistics

[80] World Health Organization, "A Global Brief on Hypertension," WHO, Geneva, Switzerland, Tech. Rep., 2013. [Online]. Available: http://apps.who.int/iris/bitstream/10665/79059/1/WHO_DCO_WHD_2013.2_eng.pdf?ua=1

[81] F. Turnbull, "Effects of different regimens to lower blood pressure on major cardiovascular events in older and younger people: Meta-analysis of randomised trials," *BMJ*, vol. 336, no. 7653, pp. 1121–1123, May 2008. [Online]. Available: https://doi.org/10.1136/bmj.39548.738368.BE

[82] L. Bonsall, "Calculating the mean arterial pressure (MAP)," 2011. [Online]. Available: https://www.nursingcenter.com/ncblog/december-2011/calculating-the-map

[83] E. A. Wehrwein and M. J. Joyner, *Chapter 8 - Regulation of blood pressure by the arterial baroreflex and autonomic nervous system.* Elsevier, 2013, vol. 117. [Online]. Available: https://doi.org/10.1016/B978-0-444-53491-0.00008-0

[84] A. Berger, "Oscillatory Blood Pressure Monitoring Devices," *BMJl*, vol. 323, no. 7318, p. 919, Oct. 2001. [Online]. Available: https://doi.org/10.1136/bmj.323.7318.919

[85] Y. Bar Ziv *et al.*, "The Sphygmomanometer Pain Test: A Simple Method for Identifying Patients at Risk of Excessive Pain after Total Knee Arthroplasty," *The Journal of Arthroplasty*, vol. 31, no. 4, pp. 798–801, Oct. 2016. [Online]. Available: https://doi.org/10.1016/j.arth.2015.10.027

[86] S. Butler *et al.*, "Evaluation of Using the Sphygmomanometer Test to Assess Pain Sensitivity in Chronic Pain Patients vs Normal Controls," *Pain Medicine*, vol. 21, no. 11, Jul. 2020. [Online]. Available: https://doi.org/10.1093/pm/pnaa191

[87] E. M. Frese *et al.*, "Blood Pressure Measurement Guidelines for Physical Therapists," *Cardiopulmonary Physical Therapy Journal*, vol. 22, no. 2, pp. 5–12, Jun. 2011. [Online]. Available: http://doi.org/10.1097/01823246-201122020-00002

[88] Y. Cemal *et al.*, "Preventative measures for lymphedema: Separating fact from fiction," *Journal of the American College of Surgeons*, vol. 213, no. 4, pp. 543–551, Oct. 2011. [Online]. Available: http://doi.org/10.1016/j.jamcollsurg.2011.07.001

[89] S. S. Thomas *et al.*, "BioWatch: A Noninvasive Wrist-Based Blood Pressure Monitor That Incorporates Training Techniques for Posture and Subject Variability," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 5, pp. 1291–1300, Sep. 2016. [Online]. Available: https://doi.org/10.1109/JBHI.2015.2458779

[90] H. Lin *et al.*, "Noninvasive and Continuous Blood Pressure Monitoring Using Wearable Body Sensor Networks," *IEEE Intelligent Systems*, vol. 30, no. 6, pp. 38–48, Nov. 2015. [Online]. Available: https://doi.org/10.1109/MIS.2015.72

[91] Y. L. Zheng *et al.*, "An Armband Wearable Device for Overnight and Cuff-less Blood Pressure Measurement," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 7, pp. 2179–2186, Jul. 2014. [Online]. Available: https://doi.org/10.1109/TBME.2014.2318779

[92] D. Griggs *et al.*, "Design and development of continuous cuff-less blood pressure monitoring devices," *Proc. 2016 IEEE SENSORS*, pp. 1–3, Nov. 2016. [Online]. Available: https://doi.org/10.1109/ICSENS.2016.7808908

[93] T. H. Wu *et al.*, "Predicting Systolic Blood Pressure Using Machine Learning," in *Proc. 7th International Conference on Information and Automation for Sustainability*, Colombo, Sri Lanka, Mar. 2015, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICIAFS.2014.7069529

[94] S. Lee and J.-H. Chang, "Oscillometric Blood Pressure Estimation Based on Deep Learning," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 2, Apr. 2017. [Online]. Available: https://doi.org/10.1109/TII.2016.2612640

[95] S. Lee and J. H. Chang, "Deep Belief Networks Ensemble for Blood Pressure Estimation," *IEEE Access*, vol. 5, pp. 9962–9972, May 2017. [Online]. Available: https://doi.org/10.1109/ACCESS.2017.2701800

[96] S. Lee and J. H. Chang, "Deep Boltzmann Regression With Mimic Features for Oscillometric Blood Pressure Estimation," *IEEE Sensors*

Journal, vol. 17, no. 18, pp. 5982–5993, Sep. 2017. [Online]. Available: https://doi.org/10.1109/JSEN.2017.2734104

[97] Y. Yoon *et al.*, "Cuff-Less Blood Pressure Estimation Using Pulse Waveform Analysis and Pulse Arrival Time," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 4, pp. 1068–1074, Jul. 2018. [Online]. Available: https://doi.org/10.1109/JBHI.2017.2714674

[98] F. Miao *et al.*, "Multi-Sensor Fusion Approach for Cuff-Less Blood Pressure Measurement," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 79–91, Jan. 2020. [Online]. Available: https://doi.org/10.1109/JBHI.2019.2901724

[99] P. Su *et al.*, "Long-term blood pressure prediction with deep recurrent neural networks," in *Proc. 2018 IEEE EMBS International Conference on Biomedical and Health Informatics*, vol. 2018, Las Vegas, NV, USA, Jan. 2018, pp. 323–328. [Online]. Available: https://doi.org/10.1109/BHI.2018.8333434

[100] F. P. W. Lo *et al.*, "Continuous systolic and diastolic blood pressure estimation utilizing long short-term memory network," in *Proc. 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Jul. 2017, pp. 1853–1856. [Online]. Available: https://doi.org/10.1109/EMBC.2017.8037207

[101] A. Goldberger *et al.*, "The MIMIC-III Waveform Database," 2016. [Online]. Available: https://archive.physionet.org/physiobank/database/mimic3wdb/

[102] Association for the Advancement of Medical Instrumentation, "American National Standard: Electronic or Automated Sphygmomanometers," Arlington, VA, 1993.

[103] E. O'Brien *et al.*, "The British Hypertension Society protocol for the evaluation of blood pressure measuring devices," *Journal of Hypertension*, vol. 8, no. 7, pp. 607–619, Jul. 1990. [Online]. Available: https://doi.org/10.1097/00004872-199007000-00004

[104] K. Mochizuki *et al.*, "Importance of respiratory rate for the prediction of clinical deterioration after emergency department discharge: a single-center, case-control study," *Acute Medicine & Surgery*, vol. 4, no. 2, pp. 172–178, Nov. 2017. [Online]. Available: https://doi.org/10.1002/ams2.252

[105] J. F. Fieselmann *et al.*, "Respiratory rate predicts cardiopulmonary arrest for internal medicine inpatients," *Journal of General Internal Medicine*, vol. 8, no. 7, pp. 354–360, Jul. 1993. [Online]. Available: https://doi.org/10.1007/BF02600071

[106] D. R. Goldhill *et al.*, "A physiologically-based early warning score for ward patients: the association between score and outcome," *Anaesthesia*, vol. 60, no. 6, pp. 547–553, Jun. 2005. [Online]. Available: https://doi.org/10.1111/j.1365-2044.2005.04186.x

[107] C. P. Subbe *et al.*, "Effect of introducing the Modified Early Warning score on clinical outcomes, cardio-pulmonary arrests and intensive care utilisation in acute medical admissions," *Anaesthesia*, vol. 58, no. 8, pp. 797–802, Aug. 2003. [Online]. Available: https://doi.org/10.1046/j.1365-2044.2003.03258.x

[108] J. McBride *et al.*, "Long-term effect of introducing an early warning score on respiratory rate charting on general wards," *Resuscitation*, vol. 65, no. 1, pp. 41–44, Apr. 2005. [Online]. Available: https://doi.org/10.1016/j.resuscitation.2004.10.015

[109] T. J. Hodgetts *et al.*, "The identification of risk factors for cardiac arrest and formulation of activation criteria to alert a medical emergency team," *Resuscitation*, vol. 54, no. 2, pp. 125–131, Aug. 2002. [Online]. Available: https://doi.org/10.1016/S0300-9572(02)00100-4

[110] H. Ansell *et al.*, "Why don't nurses consistently take patient respiratory rates," *British Journal of Nursing*, vol. 23, no. 8, pp. 414–418, May 2014. [Online]. Available: https://doi.org/10.12968/bjon.2014.23.8.414

[111] S. Milici *et al.*, "Wireless Breathing Sensor Based on Wearable Modulated Frequency Selective Surface," *IEEE Sensors Journal*, vol. 17,

no. 5, pp. 1285 – 1292, Mar. 2017. [Online]. Available: https://doi.org/10.1109/JSEN.2016.2645766

[112] D. Oletic and V. Bilas, "Energy-Efficient Respiratory Sounds Sensing for Personal Mobile Asthma Monitoring," *IEEE Sensors Journal*, vol. 16, no. 23, pp. 8295–8303, Dec. 2016. [Online]. Available: https://doi.org/10.1109/JSEN.2016.2585039

[113] X. Yang *et al.*, "Textile Fiber Optic Microbend Sensor Used for Heartbeat and Respiration Monitoring," *IEEE Sensors Journal*, vol. 15, no. 2, pp. 757–761, Feb. 2015. [Online]. Available: https://doi.org/10.1109/JSEN.2014.2353640

[114] I. Mahbub *et al.*, "A Low-Power Wireless Piezoelectric Sensor-Based Respiration Monitoring System Realized in CMOS Process," *IEEE Sensors Journal*, vol. 17, no. 6, pp. 1858–1864, Mar. 2017. [Online]. Available: https://doi.org/10.1109/JSEN.2017.2651073

[115] O. Atalay *et al.*, "Weft-Knitted Strain Sensor for Monitoring Respiratory Rate and Its Electro-Mechanical Modeling," *IEEE Sensors Journal*, vol. 15, no. 1, pp. 110–122, Jan. 2015. [Online]. Available: https://doi.org/10.1109/JSEN.2014.2339739

[116] P. H. Charlton *et al.*, "An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram," *Physiological Measurement*, vol. 37, no. 4, pp. 610–626, Mar. 2016. [Online]. Available: http://doi.org/10.1088/0967-3334/37/4/610

[117] D. A. Birrenkott *et al.*, "Robust estimation of respiratory rate via ECG- and PPG-derived respiratory quality indices," in *Proc. of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Oct. 2016, pp. 676–679. [Online]. Available: https://doi.org/10.1109/EMBC.2016.7590792

[118] J. E. Zimmerman *et al.*, "Changes in hospital mortality for United States intensive care unit admissions from 1988 to

2012," *Critical Care*, vol. 17, no. 2, pp. R81–R81, 4 2013. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23622086https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4057290/

[119] W. A. Knaus *et al.*, "APACHE II: A severity of disease classification system," *Critical Care Medicine*, vol. 13, no. 10, pp. 818–829, Oct. 1985. [Online]. Available: https://doi.org/10.1097/00003246-198510000-00009

[120] J.-R. Le Gall *et al.*, "A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study," *The Journal of the American Medical Association*, vol. 270, no. 24, pp. 2957–2963, Dec. 1993. [Online]. Available: https://doi.org/10.1001/jama.1993.03510240069035

[121] J. L. Vincent *et al.*, "The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure," *Intensive Care Medicine*, vol. 22, no. 7, pp. 707–710, Jul. 1996. [Online]. Available: https://doi.org/10.1007/BF01709751

[122] A. P. Nassar *et al.*, "Evaluation of simplified acute physiology score 3 performance: A systematic review of external validation studies," *Critical Care*, vol. 18, no. 3, p. R117, Jun. 2014. [Online]. Available: https://doi.org/10.1186/cc13911

[123] Y. Sakr *et al.*, "Comparison of the performance of SAPS II, SAPS 3, APACHE II, and their customized prognostic models in a surgical intensive care unit," *British Journal of Anaesthesia*, vol. 101, no. 6, pp. 798–803, Dec. 2008. [Online]. Available: https://doi.org/10.1093/bja/aen291

[124] A. L. E. Falcão *et al.*, "The prognostic accuracy evaluation of SAPS 3, SOFA and APACHE II scores for mortality prediction in the surgical ICU: an external validation study and decision-making analysis," *Annals of Intensive Care*, vol. 9, no. 1, p. 18, 2019. [Online]. Available: https://doi.org/10.1186/s13613-019-0488-9

[125] C. C. H. Lew *et al.*, "Performance of the Acute Physiology and Chronic Health Evaluation II (APACHE II) in the prediction of

hospital mortality in a mixed ICU in Singapore," *Proc. Singapore Healthcare*, vol. 28, no. 3, pp. 147–152, Nov. 2018. [Online]. Available: https://doi.org/10.1177/2010105818812896

[126] H. M. Giannini *et al.*, "A Machine Learning Algorithm to Predict Severe Sepsis and Septic Shock: Development, Implementation, and Impact on Clinical Practice," *Critical Care Medicine*, vol. 47, no. 11, pp. 1485–1492, Nov. 2019. [Online]. Available: https://doi.org/10.1097/CCM.0000000000003891

[127] John Hopkins Medicine, "Vital Signs," 2019. [Online]. Available: https://www.hopkinsmedicine.org/health/conditions-and-diseases/vitalsigns-body-temperature-pulse-rate-respiration-rate-blood-pressure

[128] L. S. S. Wong and J. D. Young, "A comparison of ICU mortality prediction using the APACHE II scoring system and artificial neural networks," *Anaesthesia*, vol. 54, no. 11, pp. 1048–1054, Nov. 1999. [Online]. Available: https://doi.org/10.1046/j.1365-2044.1999.01104.x

[129] G. Clermont *et al.*, "Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models," *Critical Care Medicine*, vol. 29, no. 2, Feb. 2001. [Online]. Available: https://doi.org/10.1097/00003246-200102000-00012

[130] A. Nimgaonkar *et al.*, "Prediction of mortality in an Indian intensive care unit," *Intensive Care Medicine*, vol. 30, no. 2, pp. 248–253, Feb. 2004. [Online]. Available: https://doi.org/10.1007/s00134-003-2105-4

[131] H.-C. Thorsen-Meyer *et al.*, "Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records," *Lancet Digital Health*, vol. 2, no. 4, pp. e179–e191, Apr. 2020. [Online]. Available: https://doi.org/10.1016/S2589-7500(20)30018-2

[132] L. Liu *et al.*, "Global, regional, and national causes of under-5 mortality in 2000-2015: an updated systematic analysis with

implications for the Sustainable Development Goals," *The Lancet*, vol. 388, no. 10063, pp. 3027–3035, Dec. 2016. [Online]. Available: https://doi.org/10.1016/S0140-6736(16)31593-8

[133] March of Dimes *et al.*, "Born Too Soon: The Global Action Report on Preterm Birth," World Health Organization, Geneva, Switzerland, Tech. Rep., 2012. [Online]. Available: https://www.who.int/maternal_child_adolescent/documents/born_too_soon/en/

[134] G. Parry *et al.*, "CRIB II: an update of the clinical risk index for babies score." *Lancet (London, England)*, vol. 361, no. 9371, pp. 1789–1791, May 2003. [Online]. Available: https:/doi.org/10.1016/S0140-6736(03)13397-1

[135] W. Tarnow-Mordi *et al.*, "Predicting death from initial disease severity in very low birthweight infants: a method for comparing the performance of neonatal units." *BMJ*, vol. 300, no. 6740, pp. 1611–1614, Jun. 1990. [Online]. Available: https://doi.org/10.1136/bmj.300.6740.1611

[136] D. K. Richardson *et al.*, "Birth weight and illness severity: independent predictors of neonatal mortality." *Pediatrics*, vol. 91, no. 5, pp. 969–975, May 1993.

[137] D. K. Richardson *et al.*, "Score for Neonatal Acute Physiology: a physiologic severity index for neonatal intensive care." *Pediatrics*, vol. 91, no. 3, pp. 617–623, Mar. 1993.

[138] R. F. Maier *et al.*, "Comparison of mortality risk: a score for very low birthweight infants," *Archives of Disease in Childhood - Fetal and Neonatal Edition*, vol. 76, no. 3, pp. F146–F151, May 1997. [Online]. Available: http://doi.org/10.1136/fn.76.3.F146

[139] H. García *et al.*, "Validation of a prognostic index in the critically ill newborn." *Revista de Investigacion Clinica*, vol. 52, no. 4, pp. 406–414, 2000.

[140] S. S. Harsha and B. R. Archana, "SNAPPE-II (Score for Neonatal Acute Physiology with Perinatal Extension-II) in Predicting Mortality and Morbidity in NICU," *Journal of Clinical and Diagnostic Research*,

vol. 9, no. 10, pp. SC10–SC12, Oct. 2015. [Online]. Available: https://doi.org/10.7860/JCDR/2015/14848.6677

[141] T. A. J. Houweling *et al.*, "A prediction model for neonatal mortality in low- and middle-income countries: an analysis of data from population surveillance sites in India, Nepal and Bangladesh," *International Journal of Epidemiology*, vol. 48, no. 1, pp. 186–198, Oct. 2018. [Online]. Available: https://doi.org/10.1093/ije/dyy194

[142] M. M. Medvedev *et al.*, "Development and validation of a simplified score to predict neonatal mortality risk among neonates weighing 2000 g or less (NMR-2000): an analysis using data from the UK and The Gambia," *The Lancet Child & Adolescent Health*, vol. 4, no. 4, pp. 299–311, Apr. 2020. [Online]. Available: https://doi.org/10.1016/S2352-4642(20)30021-3

[143] J. Jaskari *et al.*, "Machine Learning Methods for Neonatal Mortality and Morbidity Classification," *IEEE Access*, p. 1, Jul. 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.3006710

[144] D. Wang *et al.*, "An Optimal Pulse System Design by Multichannel Sensors Fusion," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 2, pp. 450–459, Mar. 2016. [Online]. Available: https://doi.org/10.1109/JBHI.2015.2392132

[145] Y. Zhang *et al.*, "LSTM for septic shock: Adding unreliable labels to reliable predictions," in *2017 IEEE International Conference on Big Data (Big Data)*, 2017, pp. 1233–1242.

[146] A. Ahmed *et al.*, "A wearable sensor based multi-criteria-decision-system for real-time seizure detection," in *Proc. 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Seogwipo, South Korea, Sep. 2017, pp. 2377–2380. [Online]. Available: https://doi.org/10.1109/EMBC.2017.8037334

[147] R. Amirkhan *et al.*, "Using recurrent neural networks to predict colorectal cancer among patients," in *Proc. 2017 IEEE Symposium Series*

on Computational Intelligence, Honolulu, HI, USA, Feb. 2018. [Online]. Available: https://doi.org/10.1109/SSCI.2017.8280826

[148] S. Chauhan and L. Vig, "Anomaly detection in ECG time signals via deep long short-term memory networks," in *Proc. 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Paris, France, Dec. 2015. [Online]. Available: https://doi.org/10.1109/DSAA.2015.7344872

[149] S. Baker *et al.*, "A Computationally Efficient CNN-LSTM Network for Estimation of Blood Pressure," *Planned submission to PLOS One*, 2020.

[150] Y. Shahriari *et al.*, "Electrocardiogram Signal Quality Assessment Based on Structural Image Similarity Metric," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 4, pp. 748–753, 2018.

[151] A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, May 2016. [Online]. Available: http://www.nature.com/articles/sdata201635

[152] J. Li and H. W. Lewis, "Fuzzy Clustering Algorithms - Review of the Applications," *Proc. 2016 IEEE International Conference on Smart Cloud*, pp. 282–288, Dec. 2016. [Online]. Available: https://doi.org/10.1109/SmartCloud.2016.14

[153] Y. Zhang *et al.*, "A signal quality assessment method for mobile ECG using multiple features and fuzzy support vector machine," in *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, ICNC-FSKD 2016*, 2016, pp. 966–971.

[154] C. Orphanidou *et al.*, "Signal-Quality Indices for the Electrocardiogram and Photoplethysmogram: Derivation and Applications to Wireless Monitoring," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 3, pp. 832–838, May 2015. [Online]. Available: https://doi.org/10.1109/JBHI.2014.2338351

[155] J. Seladi-Schulman and C. Stephens, "Pulse Pressure Calculation Explained," 2017. [Online]. Available: https://www.healthline.com/health/pulse-pressure

[156] S. G. Sheps, "Pulse pressure: An indicator of heart health? - Mayo Clinic," 2016. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/expert-answers/pulse-pressure/faq-20058189

[157] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, Dec. 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[158] M. Kachuee *et al.*, "Cuff-less high-accuracy calibration-free blood pressure estimation using pulse transit time," in *Proc. IEEE International Symposium on Circuits and Systems*, Jul. 2015, pp. 1006–1009. [Online]. Available: https://doi.org/10.1109/ISCAS.2015.7168806

[159] Bureau of Labor Statistics, "Occupational employment and wages, May 2014: registered nurses," Bureau of Labor Statistics, Washington D.C., Tech. Rep., 2014. [Online]. Available: http://www.bls.gov/oes/current/oes291141.htm

[160] P. Charlton *et al.*, "Breathing Rate Estimation from the Electrocardiogram and Photoplethysmogram: A Review," *IEEE Reviews in Biomedical Engineering*, vol. 11, pp. 2–20, Oct. 2017. [Online]. Available: https://doi.org/10.1109/RBME.2017.2763681

[161] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals." *Circulation*, vol. 101, no. 23, pp. 215–220, Jun. 2000. [Online]. Available: http://doi.org/10.1161/01.cir.101.23.e215

[162] I. Silva and G. B. Moody, "An open-source toolbox for analysing and processing PhysioNet databases in MATLAB and Octave," *Journal of Open Research Software*, vol. 2, no. 1, p. 27, Sep. 2014. [Online]. Available: https://doi.org/10.5334/jors.bi

[163] T. Coffey, *First Aid Manual*, 9th ed. Australia: Healthcorp Pty Limited, 2016.

[164] W. Karlen *et al.*, "Capnobase: Signal database and tools to collect, share and annotate respiratory signals," in *Proc. Annual Meeting of the Society for Technology in Anesthesia*, Jan 2010.

[165] M. Saeed *et al.*, "Multiparameter Intelligent Monitoring in Intensive Care II (Mimic-II): A Public-Access Intensive Care Unit Database," *Critical care medicine*, vol. 39, pp. 952–60, 05 2011. [Online]. Available: https://doi.org/10.1097/CCM.0b013e31820a92c6

[166] Society of Critical Care Medicine, "Critical Care Statistics," Society of Critical Care Medicine, Tech. Rep., 2019. [Online]. Available: https://www.sccm.org/Communications/Critical-Care-Statistics

[167] M. Wu *et al.*, "Understanding vasopressor intervention and weaning: risk prediction in a public heterogeneous clinical time series database," *Journal of the American Medical Informatics Association*, vol. 24, no. 3, pp. 488–495, May 2017. [Online]. Available: https://doi.org/10.1093/jamia/ocw138

[168] M. Ghassemi *et al.*, "Predicting intervention onset in the ICU with switching state space models," *Proc. AMIA Joint Summits on Translational Science*, vol. 2017, pp. 82–91, Jul. 2017.

[169] G. Batchuluun *et al.*, "Gait-Based Human Identification by Combining Shallow Convolutional Neural Network-Stacked Long Short-Term Memory and Deep Convolutional Neural Network," *IEEE Access*, vol. 6, pp. 63 164–63 186, Oct. 2018. [Online]. Available: https://doi.org/10.1109/ACCESS.2018.2876890

[170] M. Wu *et al.*, "Optimizing for Interpretability in Deep Neural Networks with Tree Regularization," Aug. 2019. [Online]. Available: https://arxiv.org/abs/1908.05254

[171] A. Vellido, "The Importance of Interpretability and Visualization in Machine Learning for Applications in Medicine and Health Care," *Neural*

Compututing and Applications, p. 18069–18083, Feb. 2019. [Online]. Available: https://doi.org/10.1007/s00521-019-04051-w

[172] M. Reyes *et al.*, "On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities," *Radiology: Artificial Intelligence*, vol. 2, no. 3, p. e190043, May 2020. [Online]. Available: https://doi.org/10.1148/ryai.2020190043

[173] L. H. Gilpin *et al.*, "Explaining Explanations: An Overview of Interpretability of Machine Learning," in *Proc. 2018 IEEE 5th International Conference of Data Science and Advanced Analytics*, Turin, Italy, Feb. 2019, pp. 80–89. [Online]. Available: https://doi.org/10.1109/DSAA.2018.00018

[174] M. Churpek *et al.*, "The value of vital sign trends for detecting clinical deterioration on the wards," *Resuscitation*, vol. 102, pp. 1–5, Feb. 2016. [Online]. Available: https://doi.org/10.1016/j.resuscitation.2016.02.005

[175] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, Nov. 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[176] J. E. Lawn *et al.*, "Global report on preterm birth and stillbirth (1 of 7): definitions, description of the burden and opportunities to improve data," *BMC Pregnancy and Childbirth*, p. S1, Feb. 2010. [Online]. Available: https://doi.org/10.1186/1471-2393-10-S1-S1

[177] W. Harrison and D. Goodman, "Epidemiologic Trends in Neonatal Intensive Care, 2007-2012," *JAMA Pediatrics*, vol. 169, no. 9, pp. 855–862, Sep. 2015. [Online]. Available: https://doi.org/10.1001/jamapediatrics.2015.1305

[178] J. H. Park *et al.*, "Predicting mortality in extremely low birth weight infants: Comparison between gestational age, birth weight, Apgar score, CRIB II score, initial and lowest serum albumin levels," *PloS One*, vol. 13, no. 2, pp. e0 192 232–e0 192 232, Feb. 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0192232

[179] D. K. Richardson *et al.*, "SNAP-II and SNAPPE-II: Simplified newborn illness severity and mortality risk scores." *The Journal of Pediatrics*, vol. 138, no. 1, pp. 92–100, Jan. 2001. [Online]. Available: https://doi.org/10.1067/mpd.2001.109608

[180] E. Nagy *et al.*, "Gender-Related Heart Rate Differences in Human Neonates," *Pediatric Research*, vol. 47, no. 6, pp. 778–780, Jun. 2000. [Online]. Available: https://doi.org/10.1203/00006450-200006000-00016

[181] N. Kumar *et al.*, "Continuous vital sign analysis for predicting and preventing neonatal diseases in the twenty-first century: big data to the forefront." *Pediatric Research*, vol. 87, no. 2, pp. 210–220, Jan. 2020. [Online]. Available: https://doi.org/10.1038/s41390-019-0527-0

[182] S. Saria *et al.*, "Integration of early physiological responses predicts later illness severity in preterm infants," *Science Translational Medicine*, vol. 2, no. 48, pp. 65–48, Sep. 2010. [Online]. Available: https://doi.org/10.1126/scitranslmed.3001304

[183] T. Tanner, "The Problem of Alarm Fatigue," *Nursing for Women's Health*, vol. 17, no. 2, pp. 153–157, Dec. 2013. [Online]. Available: https://doi.org/10.1111/1751-486X.12025

[184] S. Sendelbach and M. Funk, "Alarm Fatigue: A Patient Safety Concern," *AACN Advanced Critical Care*, vol. 24, no. 4, pp. 378–386, Oct. 2013. [Online]. Available: https://doi.org/10.4037/NCI.0b013e3182a903f9

[185] C. F. Poets, "Reducing alarms in the NICU," *Archives of Disease in Childhood - Fetal and Neonatal Edition*, vol. 103, no. 4, pp. F297–F298, Nov. 2017. [Online]. Available: http://doi.org/10.1136/archdischild-2017-314259