Machine Learning for the Detection of Clinical Deterioration on Hospital Wards



Farah Shamout

Department of Engineering

University of Oxford

Supervised by:

Professor David A. Clifton

Dr. Tingting Zhu

This thesis is submitted to the Department of Engineering Science, University of Oxford, in fulfilment of the requirements for the degree of *Doctor of Philosophy*

Balliol College Trinity Term 2019

This thesis is dedicated to,

My mother Hanady and father Emad,
for instilling in me the love of learning;

My brothers Khaled and Abdulrahman,
for inspiring me with their persistence and hardwork;

My partner Hazem,
for encouraging me to believe in myself.

Thank you for your love and support.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are entirely my own work.

Farah Shamout

Trinity Term 2019

Acknowledgements

"الحيرة هي بداية المعرفة."
— جبران خليل جبران

"Perplexity is the beginning of knowledge."
— Gibran Khalil Gibran

Around four years ago, I became curious about the role of data and artificial intelligence in health care. I then emailed Professor David Clifton to enquire about the graduate programs in "intelligent healthcare systems" at my dream school - the University of Oxford. Never had I known that I would pursue my doctoral studies there a couple of months later. Thank you Professor Clifton for agreeing to speak with me in the first place, and for your supervision as I deep dived with my research into the field of medical informatics. I appreciate your guidance and support throughout the journey.

I would also like to thank my co-supervisor Dr. Tingting Zhu. Thank you for sharing your time and expertise. I am grateful that you have always been there for me and that you have kept me motivated to stay on track. I am also lucky to have a role model to look up to as I pursue a career in academia.

Thank you to our clinical collaborator Dr. Peter Watkinson, for taking the time to explain medical concepts to a computer engineer, for providing me with access to a dataset that is core to my research, and for ensuring that I always feel welcomed at the Kadoorie Center at the John Radcliffe Hospital. For that, I am also thankful to other members of the HAVEN team at the Kadoorie Centre, especially Verity Westgate, Julie Darbyshire, Robert Hatch, and Dr. Duncan Young. Thank you to Dr. Marco Pimentel and Dr. Mauro Santos for your support and for all of our informative discussions about early warning score systems.

Thank you to the old, new, and honorary members of the *Time Series Terrace*, who kept me energised during our lunch and coffee breaks (and of course, the occasional trips to *Heavenly Desserts*). I would also like to thank all of the senior researchers in the Computational Health Informatics Laboratory for always providing valuable feedback during our group meetings. Thank you to Caroline Brown and the wider administrative team at the Institute of Biomedical Engineering for their support.

I am also thankful to the Rhodes Trust for funding my doctoral studies and for all their efforts in making the Oxford experience even more special, especially through the forums, scholar-led talks, and first- and second- year retreats. I would also like to thank my colleagues in the Kellogg Artificial Intelligence Programme. Our discussions during the course trips broadened my understanding of AI.

Thank you to my friends from the Rhodes Community and Balliol College who made Oxford feel like a home to me. I enjoyed our walks, afternoon teas, and lovely dinners. I am also grateful to others who lived miles away from Oxford, but took an extra effort to pay me a vist, chat over Skype, or even send me a motivational text message.

I was showered with love throughout the process by my lifelong mentors Emad Shammout and Hanadi Al Tammam; the entirety of this thesis is dedicated to you. Thank you to my brothers Khaled and Abdulrahman for your understanding, love, and endless support.

This thesis would have not been possible without the support of my loving and caring partner, and best friend, Hazem Hashem. Thank you for your patience.

Finally, thank you to the city of dreaming spires for a magical three years.

Machine Learning for the Detection of Clinical Deterioration on Hospital Wards
Farah Shamout
Balliol College
DPhil in Engineering Science
Trinity Term 2019

Abstract

Assessment of physiological instability preceding adverse events on hospital wards has been previously investigated through clinical early warning score (EWS) systems. Such systems often rely on routinely collected vital-sign data, such as the National Early Warning Score system that is currently deployed across the United Kingdom. In this thesis, we propose improved and generalisable risk assessment algorithms, that can continuously alert for deterioration preceding the composite outcome of unplanned intensive care unit admission, cardiac arrest, and mortality. We develop and validate three statistical and machine learning models using large-scale datasets from two independent centres: Oxford University Hospitals and Portsmouth NHS Hospitals within the 'HAVEN' database. The retrieved datasets include patient demographics, vital signs, laboratory tests, and data of the occurrence of any adverse events. The first model is the *Age- and Sex- specific Early Warning Score* (ASEWS), which was derived from statistical distributions of vital signs. The work suggests that accounting for age-related vital-sign changes can more accurately detect deterioration in younger patients (16-45 years old). We also propose the *Deep Early Warning Score* (DEWS), which consists of an end-to-end attention-based deep learning architecture that processes vital-sign time-series data. The vital-sign data is initially modelled using Gaussian Process Regression. The results suggest that deep learning improves the detection of clinical outcomes by recognising complex patterns in the data. The final model is the *information Fusion in a multi-modal Early Warning System* (iFEWS), which incorporates additional information about the patient, such as results of laboratory tests or the first diagnosis assigned at admission. In this framework, representation learning and continued learning within a multi-modal system improved the performance of detecting deterioration. All of our proposed models achieved a better performance than the state-of-the-art clinical EWS systems across two independent testing sets. Given their high performance, clinical utility, and illustrated interpretability, our models can be easily deployed in clinical settings to supplement existing EWS systems since they use the same data streams.

List of Publications

Thesis Publications

Shamout, F., Zhu, T., Sharma, P., Watkinson, P., & Clifton, D. A. (2019) Deep Interpretable Early Warning System. IEEE Journal of Biomedical Health Informatics.

Shamout, F., Zhu, T., Clifton, L., Prytherch, D., Briggs, J., Tarassenko, L., Watkinson, P., Clifton, D. A. (2019) Early warning score adjusted for age to predict the composite outcome of mortality, cardiac arrest or unplanned intensive care unit admission using observational vital-sign data: a multicentre development and validation. BMJ Open.

Related Publications

Wong, D., Gerry, S., **Shamout, F.**, Pimentel, M., Clifton, D. A., Watkinson, P. (2020) Cross-Sectional Centiles of Blood Pressure by Age and Sex: a four-hospital database retrospective observational analysis. BMJ Open.

Sharma, P., **Shamout, F.**, Abrol, V., Clifton, D. A. Neural Processes for Modelling Personalised Vital-Sign Time-Series Data. (*Under Review*)

Other Publications

Yang, Y., Yangyang, E., Walker, T., Walker, S., Wilson, D., Peto, T., Crook, D., **Shamout, F.**, Zhu, T., Clifton, D. A. (2019) DeepAMR for predicting resistance co-occurrence of Mycobacterium tuberculosis. Bioinformatics.

Sharma, P., **Shamout, F.**, Clifton, D. A. (2019) Preserving Patient Privacy while Training a Predictive Model of In-hospital Mortality. Neurips 2019 AI for Social Good Workshop.

Table of contents

List of figures **xiii**

List of tables **xv**

1 Introduction **1**

1.1 Clinical Deterioration on Hospital Wards 2

1.2 Early Warning Score Systems 5

1.2.1 Strengths and Limitations 5

1.2.2 Clinical Baselines 6

1.3 Thesis Summary 7

2 Review of Machine Learning for Early Warning Models **9**

2.1 Feature Extraction 11

2.1.1 Hand-crafted Features 12

2.1.2 Data Standardisation 12

2.1.3 Representation Learning 14

2.2 Predictive Inference 18

2.2.1 Outcome Classification Framework 18

2.2.2 Clustering for Abnormality Detection 21

2.3 Evaluation of Early Warning Models 21

2.3.1 Performance Metrics 22

2.3.2 Interpretability 24

3 Summary of Datasets 26

3.1 Introduction 26

3.2 Electronic Health Records 26

3.2.1 Inclusion Criteria 28

3.2.2 Oxford University Hospitals Datasets 29

3.2.3 Portsmouth Hospitals NHS Trust Dataset 31

3.3 Extracted Variables 32

3.3.1 Demographics 32

3.3.2 Outcomes 33

3.3.3 Vital Signs 36

3.3.4 Laboratory Tests 40

3.3.5 Diagnosis Codes 44

3.4 Detection Framework 46

4 Age- and Sex- based Early Warning Score 48

4.1 Introduction 48

4.2 Alerting Model 50

4.3 Model Evaluation 53

4.4 Results 54

4.4.1 Fine-tuned Model Parameters 54

4.4.2 Performance Assessment 58

4.4.3 Clinical Utility 60

4.5 Discussion 61

5 Deep Interpretable Early Warning System 64

5.1 Introduction 64

5.2 State-of-the-art . . . 66
5.3 DEWS: an end-to-end learning framework . . . 67
5.3.1 Patient-specific Feature Transformation . . . 69
5.3.2 Classifier Architecture . . . 72
5.4 Empirical Study . . . 73
5.4.1 Evaluation Metrics . . . 74
5.4.2 Data Modelling . . . 75
5.4.3 Model Variants as Baselines . . . 78
5.4.4 Deep Learning Experiments . . . 79
5.4.5 Results . . . 80
5.5 Discussion . . . 86

6 Information Fusion in a Multi-modal Early Warning System 91
6.1 Introduction . . . 91
6.2 State-of-the-art . . . 92
6.3 iFEWS: a multi-modal alerting framework . . . 94
6.3.1 Learning from vital signs . . . 95
6.3.2 Learning from Laboratory Tests . . . 97
6.3.3 Learning from the first diagnosis . . . 99
6.3.4 Label Fusion . . . 99
6.3.5 Continued Training . . . 100
6.4 Empirical Study . . . 102
6.4.1 Model Variants as Baselines . . . 102
6.4.2 Evaluation Metrics . . . 103
6.4.3 Deep Learning Experiments . . . 103
6.4.4 Results . . . 105
6.5 Discussion . . . 114

7 Conclusion **117**

7.1 Thesis Summary 117

7.1.1 Limitations 118

7.2 Recommendations for Early Warning Systems 120

7.3 Pathway to Translation 122

7.4 Future Work 124

7.4.1 Learning Optimal Treatment Strategies 124

7.4.2 Interpretability, Fairness, and Privacy 125

References **127**

Appendix A **143**

Appendix B **148**

Appendix C **150**

List of figures

1.1 Schematic of factors that may contribute to clinical deterioration during a hospital admission. 4

2.1 Sizes of datasets reported in related works. 10
2.2 Schematic of the general machine learning pipeline in studies related to predicting clinical deterioration. 11

3.1 Consort diagram visualising cohort extraction from the datasets. 30
3.2 Distribution of ICD-10 codes assigned within the experimental datasets. 45
3.3 Example of a patient admission with a subsequent adverse outcome. . . 46

4.1 Prevalence of abnormal observation sets in the training set with varying age. 51
4.2 Cumulative distribution function for heart rate and systolic blood pressure for two different age groups. 52
4.3 Cumulative distribution function of heart rate where the centiles are used to obtain the alerting thresholds. 53
4.4 Visualisation of the alerting thresholds of AEWS, NEWS, and MCEWS. 56
4.5 Visualisation of the alerting thresholds for AEWS and ASEWS across females and males. 57
4.6 Distributions of bootstrapped samples to compute performance metrics. 58

4.7 Efficiency curve for AEWS and NEWS across the two testing sets. . . . 61
4.8 Case studies to compare the AEWS and NEWS scores. 62

5.1 Visualisation of the feature extraction techniques: Carrying Last Value Forward, Linear Interpolation, and Gaussian Process Regression. 69
5.2 Schematic of the deep learning architecture DEWS. 71
5.3 Examples of physiological variables modelled using GPR. 77
5.4 Efficiency curves comparing NEWS with DEWS. 84
5.5 Visualisation of the attention weights assigned to the sequences using DEWS. 86
5.6 Mean probability of DEWS and the normalised NEWS score at admission and prior to the adverse event. 87
5.7 Performance in terms of AUROC at every two hours prior to an adverse event. 88

6.1 Architecture of the multi-channel attention-based autoencoder. 96
6.2 Schematic of iFEWS model that learns from multi-modal data. 100
6.3 Reconstructed respiratory rate sequence using the autoencoder architectures. 106
6.4 Weights of logistic regression model using laboratory tests. 111
6.5 Visualisation of coefficients assigned to auxiliary outputs. 112
6.6 Visualisation of embeddings learned from diagnosis groups in iFEWS. . 113
6.7 Efficiency curves comparing iFEWS and LDTEWS:NEWS. 115

A.1 Probability density function of the continuous vital signs. 145
A.2 Probability density function of the laboratory tests. 146

List of tables

1.1 Primary antecedent instability prior to adverse events as reported in the literature. 3

2.1 Feature representation techniques used in previous studies. 17
2.2 Summary of classifiers used for outcome prediction in the literature. . . 20

3.1 Characteristics of the experimental datasets. 33
3.2 Distribution of outcomes across the experimental datasets. 35
3.3 Characteristics of the vital-sign variables across the experimental datasets. 39
3.4 Number of laboratory-test measurements in the experimental datasets. 43
3.5 Median and interquartile range (IQR) of laboratory-test measurements across the experimental datasets. 44
3.6 ICD-10 codes assigned to the admissions in the experimental datasets. . 45
3.7 Summary of the prevalence of 'abnormal' observation sets measured within 24 hours of the composite outcome. 47

4.1 Model performance summary of ASEWS, AEWS, NEWS, and MCEWS for the different age groups in the $\mathcal{D}_{O,2}$ testing set. 59
4.2 Model performance summary of ASEWS, AEWS, NEWS, and MCEWS for the different age groups in the $\mathcal{D}_P$ testing set. 60

5.1 Characteristics and demographics of the new balanced training set $\mathcal{D}_{O,1B}$ used for training and validation. 74
5.2 Distributions of the vital signs in terms of median and interquartile range in the new balanced training set $\mathcal{D}_{O,1B}$. 76
5.3 Mean and standard deviation of the root mean squared error for all the modelled vital-sign sequences in the training set. 78
5.4 Model performance summary of DEWS and NEWS in $\mathcal{D}_{O,2}$. 81
5.5 Model performance summary of DEWS and NEWS in $\mathcal{D}_{O,2}$ across different patient sub-groups. 82
5.6 Model performance summary of DEWS and NEWS in $\mathcal{D}_P$ across different patient subgroups. 83

6.1 Mean and standard deviation of the mean squared error for reconstructing vital-sign sequences using autoencoder architectures. 105
6.2 Performance of iFEWS on the validation set using different training schemes. 106
6.3 Predictive power of laboratory tests on the validation set. 107
6.4 Performance evaluation of iFEWS and the baseline classifiers on $\mathcal{D}_{O,2}$. 108
6.5 Performance evaluation of iFEWS in comparison to LDTEWS:NEWS across different patient subgroups in $\mathcal{D}_{O,2}$. 109
6.6 Performance evaluation of iFEWS in comparison to LDTEWS:NEWS across different patient subgroups in $\mathcal{D}_P$. 110

A.1 Plausible values of the vital signs and laboratory tests. 144
A.2 Reference of the ICD-10 group numbers and their description. 147

B.1 Alerting thresholds of NEWS. 148
B.2 Alerting thresholds of MCEWS. 148

B.3 Scores assigned to thresholds based on centile cut-offs in MCEWS. . . . 149

C.1 Model complexities in terms of number of parameters in Chapter 5. . . 150

C.2 Model complexities in terms of number of parameters in Chapter 6. . . 151

Chapter 1

Introduction

Despite efforts to improve outcomes, hospitalised patients still suffer from unexpected complications that are life-threatening due to the delayed detection of deterioration [1, 2]. In fact, avoidable deaths, or deaths that should not occur in the presence of timely and effective health care, constitute 23% of total deaths in the United Kingdom [3]. The delayed recognition of deterioration has been shown to be associated with human related monitoring failures [4].

Machine learning (ML), a sub-field of artificial intelligence (AI), is the scientific study of optimising mathematical models using large-scale data [5]. With the recent surge of labelled data in healthcare, ML has gained popularity for building predictive models with the potential of improving patient care [6]. Designing robust ML pipelines involves evaluating several engineering choices, ranging from problem formulation and data pre-processing to algorithm optimisation. Domain expertise is also essential for guiding the process as it ensures that the modelling assumptions are valid. The **first objective** of this thesis is to develop and validate end-to-end ML frameworks for automated patient monitoring while addressing realistic constraints within clinical settings.

Predictive ML algorithms have been previously developed to assess the future occurrence of a specific diagnosis or outcome using high-dimensional and multi-modal data [7]. Despite their high performance, those algorithms have not been yet been deployed in clinical settings, especially with increasing AI regulation that require the individual's 'right to an explanation' [8]. Therefore, the **second objective** of this thesis is to develop state-of-the-art models that could achieve high predictive performance and are inherently *interpretable* - a first step towards explainable AI [9]. We specifically focus on designing models that can be compared to existing decision support tools, namely 'early warning systems', which are presented in more detail in Section 1.2.

Since the definition of clinical deterioration varies across the literature, we offer a brief introduction to clinical deterioration on hospital wards and early warning systems; which will be helpful for framing the objectives of this thesis going forward. We also summarise the thesis contributions and structure of chapters.

1.1 Clinical Deterioration on Hospital Wards

Clinical deterioration refers to the worsening of a patient's condition on hospital wards. However, the definition of what exactly constitutes clinical deterioration has evolved over time based on the data collection and processing techniques. Early attempts to define clinical deterioration focused on medical neglect and its end result of clinical complications [10]. Subsequent studies focused on more discrete clinical events, such as severe sepsis, unexpected cardiac arrest, ICU admission or death [11, 12], and tend to select one or more end-point measures of clinical deterioration. Similarly in our work, we seek to define a composite outcome consisting of more than one discrete clinical event.

Table 1.1: Primary antecedent instability within n hours to adverse events as reported in the literature.

Reference	Clinical Conditions	n	Adverse Event
[15]	Respiratory and metabolic derangement	8	Cardiac arrest
[2]	Hypo-tension, tachypnoea, & tachycardia	8	Death not preceded by cardiac arrest or ICU admission
[16]	Hypo-tension, tachycardia, tachypnoea, and sudden change in level of consciousness	8	ICU admission
[17]	Low systolic blood pressure, level of consciousness, heart rate & respiratory rate	24	Death, cardiac arrest or unplanned ICU admission

An adverse event is defined as unintended injury or complication, which results in disability, death or prolongation of hospital stay. It is caused by health care management rather than the patient's underlying disease [13]. Such events incur high costs of prolonged hospital stays, litigation, staff time, impact on patients and staff, and broader economic consequences [14]. One major contributory factor is failure to identify at-risk patients and to undertake preventative measures [14]. This motivated the development of early detection models to ensure deterioration is managed in a timely manner.

Previous studies that assessed data retrospectively have also highlighted antecedent instability, or serious physiological abnormalities, prior to an event [15, 2, 17, 11]. Table 1.1 provides a brief overview of primary antecedent instability as reported in the literature. We note that a standardisation of findings is not possible due to variations in patient cohorts. However, the nature of the observed derangement prior to outcomes seems to vary across studies. For example, [2] suggests that tachypnoea (i.e., rapid breathing) and tachycardia (i.e., high heart rate) precede ICU admission, while [17] suggests that low heart rate and breathing rate may also precede ICU admission. The general consensus seems to be that derangement develops over time prior to an event.

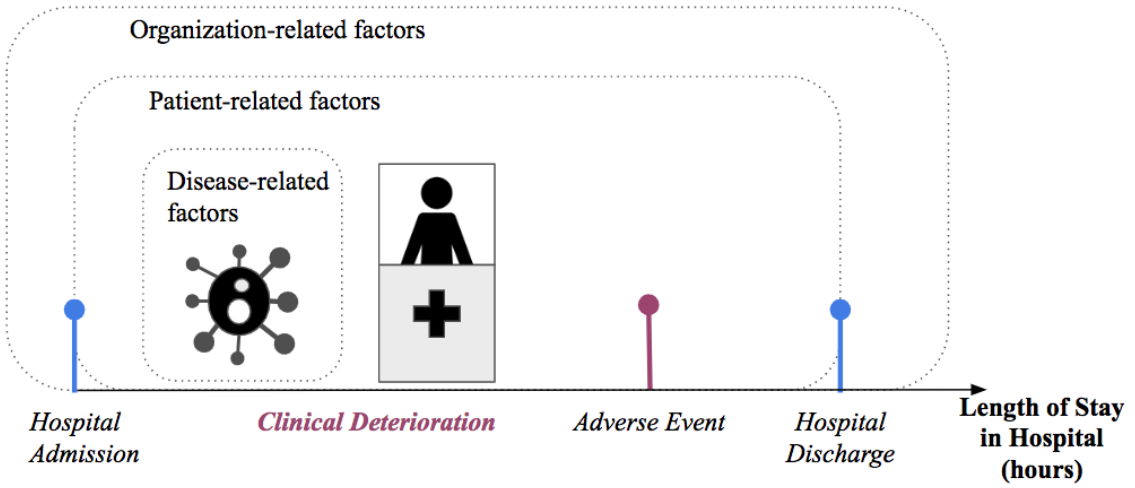


Fig. 1.1: Schematic of factors that may contribute to clinical deterioration during a patient's hospital admission.

There has also been interest in assessing changes in laboratory variables [18]. However, currently deployed systems mainly use vital signs as they are most frequently collected on hospital wards.

Overall, the clinical deterioration framework must account for patient-, disease-, and organisational-based factors [11], as illustrated in Fig. 1.1. On an individual-level, patient-related factors may include demographics, lifestyle factors, and coexisting medical conditions. Disease-related factors may include degree of severity, medications and therapy, rate of progression, interventions, surgeries, and procedures. Organisational-factors may include type of hospital, time of the day, staff ratio, or staff training. We thus expect the coherent model to maximise information retrieval on the patient-level by developing organisation-agnostic models; i.e., models that can generalise across different patient populations and across different hospitals.

1.2 Early Warning Score Systems

Early Warning Score (EWS) systems, also known as 'track-and-trigger' systems, assess patients for clinical deterioration by assigning scores to simple measured physiological variables, such as heart rate, respiratory rate, and oxygen saturation. The first physiological EWS system was introduced in 1997 [19]. Subsequent modifications were developed and adopted across different hospitals. The scores (i.e. 0-3 points) are assigned based on predetermined normality ranges for the variables. The aggregate EWS score, which is usually a sum of all individual scores, indicates whether a patient will deteriorate within the next N hours, usually 24 hours [20]. EWS systems perform continuous assessment during a patient admission, whenever a set of vital signs is recorded.

1.2.1 Strengths and Limitations

EWS systems have been shown to be good predictors of patient mortality, cardiac arrest, and unplanned ICU admission [21, 12]. Most of them are simple to use and can be easily calculated with pen and paper. They can also be considered as highly interpretable, as a clinician can easily infer which variable is most relevant to a patient's health status based on the scores assigned.

However, the alerting thresholds of EWS systems are usually developed through heuristic means based on clinical expertise [22]. The score also only depends on the last set of recorded observations and does not consider any additional historical or patient-specific information. While some EWS are even recommended to be standardised across different hospitals and patient sub-populations [23], they have not been extensively validated across different patient cohorts. When evaluated in low-resource settings, it has been shown that disease and population differences may influence the performance of EWS systems [24].

1.2.2 Clinical Baselines

We benchmark our proposed models to EWS systems that are considered to be state-of-the-art:

1. **National Early Warning Score:** The Royal College of Physicians (RCP) launched the National Early Warning Score (NEWS) in 2012 to standardise the assessment of acutely ill adult patients across the UK, which is a modified version of the VitalPAC Early Warning Score (ViEWS) [25]. NEWS assigns scores to seven physiological parameters and aggregates their scores to identify patients who are at risk of deterioration. NEWS has shown to have superior performance to other early warning scores based on vital-sign variables [26] and has been widely adopted across the UK National Health Service.

2. **Manual Centile-based Early Warning Score:** Cumulative distribution functions (cdfs) determined the normality ranges of vital signs in the Centile-based Early Warning Score (CEWS) and thereafter the Manual Centile-based Early Warning Score (MCEWS) systems [27, 28]. MCEWS achieved a similar performance to NEWS and is considered as one of the first data-driven scores.

3. **Laboratory Decision Tree Early Warning Score:** Decision trees were built independently for female and male patients to obtain the alerting thresholds of the Laboratory Decision Tree Early Warning Score (LDTEWS), using the first set of measurements of seven laboratory tests [29].

4. **LDTEWS:NEWS index:** The LDTEWS:NEWS index assesses routinely collected laboratory measurements with vital signs to calculate the risk of unplanned ICU admission or in-hospital mortality within 24 hours [30]. It combines NEWS with LDTEWS [29].

Overall, the current practice for patient monitoring generally involves (i) continuous alerting using routinely collected variables, such as vital signs or laboratory tests, (ii) assessment of a pre-defined outcome within a future window of length N hours, and (iii) a score that can be displayed on hospital systems to indicate the patient's risk of deterioration. With larger datasets and increased computational resources, we seek to build data-driven systems that achieve state-of-the-art performance and can supplement the decision-making process using current practices.

1.3 Thesis Summary

The focus of this thesis is to design algorithms for patient monitoring by processing multi-modal time series data. Our work seeks to support the decision-making process in the hospital by capturing complex relationships between the input and output variables. The thesis is structured as follows:

- In Chapter 2, we perform a literature review of (i) state-of-the-art in ML applications in health care with a particular focus on patient monitoring applications and (ii) the interpretability of existing ML models in health care.
- Chapter 3 presents qualitative and quantitative summaries of the datasets and variables used throughout the thesis for model training and validation. The datasets were extracted from electronic medical records collected in the UK.
- We present a patient-specific statistical early warning score in Chapter 4 using routinely collected physiological variables. Our results suggest that incorporating age improves the performance for young patients (16-25 years old) in comparison to the clinical benchmark.

- In Chapter 5, we propose an end-to-end ML framework consisting of patient-specific data modelling and a deep learning early warning system with an attention mechanism for interpretability that learns complex patterns in the data.
- Chapter 6 presents a multi-modal model that uses vital signs, laboratory tests, and the first diagnosis code assigned to the patient at admission. We use representation learning and continued learning to fine-tune the final model.
- Finally, Chapter 7 presents concluding remarks and directions for future research.

Chapter 2

Review of Machine Learning for Early Warning Models

Machine learning models learned from the analysis of large quantities of medical data can be used to assist clinical practice, ranging from natural language processing to computer vision applications [31, 32]. In the last few years, many machine learning models were proposed to alert for adverse events or to provide early diagnosis [33–36]. Recent surveys focus on deep learning for broad health informatics applications [37] or for EHR analysis [38]. In this chapter, we present a high-level overview of machine learning models that were specifically developed for early warning purposes.

Designing an end-to-end machine learning model involves a multidisciplinary effort that extends from feature engineering to model development and evaluation. We consider a machine learning model as a mapping of an input to an output:

$$f : \mathbf{X} \rightarrow y \tag{2.1}$$

where $f(.)$ is a function consisting of parameters Θ, $\mathbf{X}$ is the *input* and y is the *output*. For example, $\mathbf{X}$ can consist of vital signs and results of laboratory-test measurements of

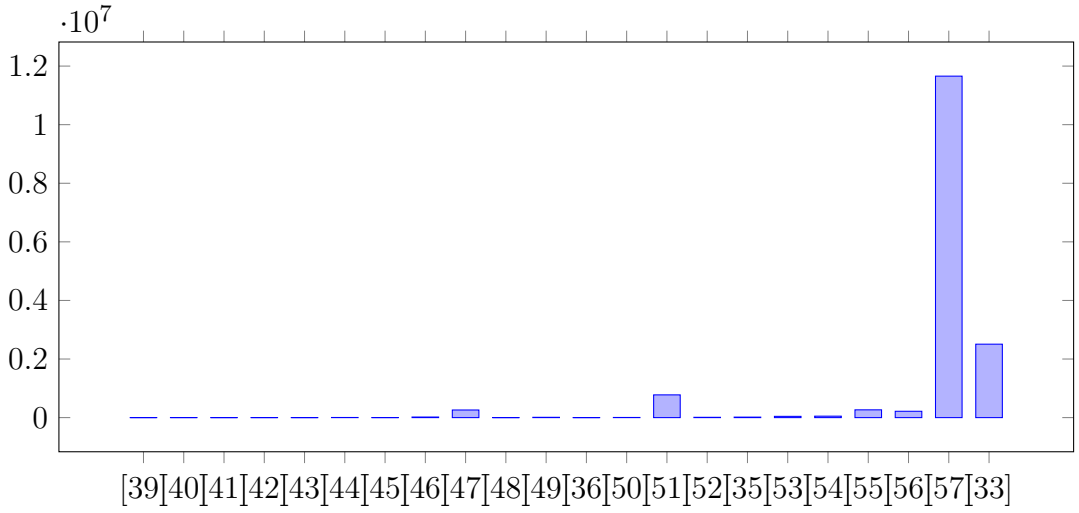


Fig. 2.1: Dataset sizes reported in the literature in ascending order from left (2008) to right (2019). The vertical axis represents the dataset size, in terms of the number of patient admissions, and the horizontal axis represents the reference number.

a patient and y can represent a binary label indicating the occurrence of ICU transfer or cardiac arrest during hospitalisation, as in [36].

The parameters Θ are typically optimised using a *training set* and evaluated on a held-out *test set*.[1] Fig. 2.1 shows the overall dataset sizes reported in studies published in the last 10 years (arranged in chronological order from left to right). There is an increase of six orders of magnitude between 2008 and 2019, which highlights the accessibility of EHR datasets for research purposes in recent years.

An objective function $\mathcal{L}(y, \hat{y}|\Theta)$, also known as the loss function, measures the difference between the true labels y and the values predicted by the approximated model $\hat{y}$. Minimising $\mathcal{L}(y, \hat{y}|\Theta)$ involves an iterative optimisation algorithm such as gradient descent [58]. As such, every machine learning algorithm involves approximating a model $f(.)$, a loss function $\mathcal{L}(y, \hat{y}|\Theta)$, and an optimisation algorithm, in order to map the input to the output.

[1] In clinical studies, the test set is usually termed the validation set, not to be confused with the portion of the training set used for ML-oriented tasks, such as hyperparameter selection.

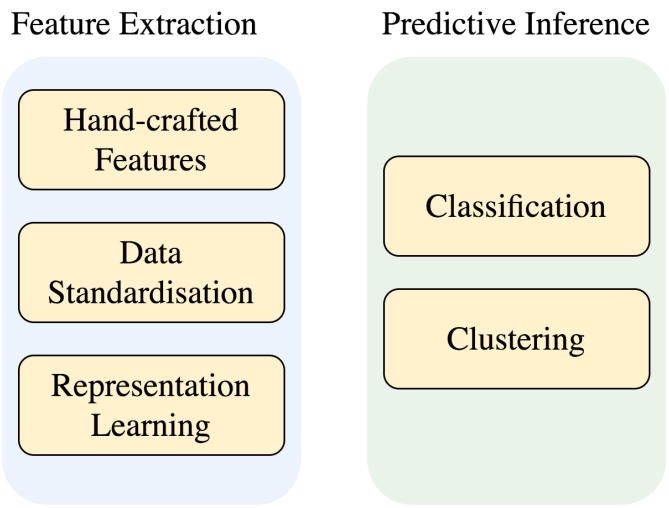


Fig. 2.2: General steps taken in a machine learning pipeline pertaining to clinical deterioration. The two steps are (i) extraction of intermediary features after data cleaning and (ii) risk prediction through a classification or clustering framework.

Fig. 2.2 depicts the typical pipeline of a machine learning application, starting from the input $\mathbf{X}$ to the corresponding output represented by y. The first task learns to extract intermediary features (Section 2.1) while the second task learns patterns in the data to produce the output (Section 2.2). Such models are usually evaluated based on clinical utility and interpretability (Section 2.3). While we do not delve into the mathematical details of the various techniques, we provide intuitive explanations for context of use.

2.1 Feature Extraction

The performance of early warning models relies on the feature representation as for any other machine learning model [59]. As reported in related works, feature extraction generally involves at least one of domain-expertise for hand-crafted features (Section 2.1.1), data standardisation (Section 2.1.2), or representation learning (Section 2.1.3).

2.1.1 Hand-crafted Features

Domain expertise is commonly used to provide guidance on the design of the data pre-processing pipeline. This involves (i) feature selection from the input space, (ii) designing hand-crafted features, and (iii) incorporating prior knowledge of the structure of the data in the model design.

Examples of hand-crafted features in related works are pulse pressure [52, 40], shock index [39, 48, 52], mean arterial pressure (MAP) [41, 52], oxygen delivery index [48], absolute successive difference of heart rate, estimated cardiac output, slope of fitted regression lines, or slope projections [39]. Statistical measures can be obtained from the distributions of the raw data, such as minimum and maximum extremes, moments (mean, standard deviation, and skewness), percentiles or the difference between two percentiles [39, 60, 46].

While waveform data was not used in this thesis, it has been previously used to obtain time series features [42, 61, 40, 34]. Those features may be categorised into four types: data adaptive, non-data adaptive, model-based and data-dictated approaches [62]. Overall, deriving hand-crafted features is a powerful tool in the design of machine learning models and has been used extensively over the years. However, it is a time-consuming and labour-intensive process, requires expert knowledge, and does not scale well to new problems.

2.1.2 Data Standardisation

Machine learning algorithms require further data preparation steps to ensure stability of learning. Within the scope of this thesis, related works typically deal with the noise, sparsity and irregularity of the clinical data and align the scales of the various predictor variables. Issues pertaining to data quality will be further elaborated in Section 3.2.

Time-series Modelling

Time-series modelling is widely used in studies pertaining to early warning models [43, 54]. It is often used either (i) to infer a pattern of the physiological trajectory or (ii) as an interpolation technique to overcome the sparsity and irregularity of physiological data.

Linear dynamic systems have been previously used to model physiological variables for ICU monitoring [63] and detection of sepsis [64]. Hidden Markov Models (HMMs) were also used to model health trajectories of patients [45, 65]. However, such models cannot easily adapt to irregularly sampled vital-sign data. Additionally, each hidden state in an HMM only depends on the previous state [66]. Another approach for modelling similar data is the kernel-based support vector regression [43].

One of the most popular techniques for time series modelling within the clinical domain is Gaussian Process Regression (GPR). GPR is based on a non-parametric stochastic process that offers a probabilistic approach for time-series modelling by providing confidence intervals for estimated values at unobserved time instances. We introduce GPR in Section 5.3.1, and a more comprehensive introduction can be found in [67]. Previous studies illustrate the robustness of the single-task GPR [43, 68, 69] in modelling a single physiological time-series variable. Others focus on multi-task GPR [70, 54, 49], which learns similarities across several time-series data data and models them simultaneously. The use of GPR relies heavily on the choice of the kernel that encodes prior knowledge of any nonlinear time-series dynamics that might be hypothesised to exist in the data.

Most recently, Neural Processes (NP), a class of neural latent variable models, were also introduced as a probabilistic regression approach [71], which generalises GPR through the use of generative models from deep learning.

Modelling the physiological trajectory of patients has become increasingly popular for further use in classification [54] or clustering early warning applications [45, 68].

Feature Scaling

Empirical studies show that the performance of predictive models relies on the statistical normalisation of the input space [72]. Z-score normalisation with zero mean and unit standard deviation is a widely used tool in feature scaling of numeric clinical variables [73, 74, 56, 35, 75]. Min-max normalisation performs a scaling of the feature values to lie within a range, such as [0,1] in [33]. A rigorous comparison of the different normalisation techniques in the context of clinical deterioration does not exist. The current practice is to choose the normalisation technique based on its effect on the performance of the respective classifier. This presents an opportunity for future research.

2.1.3 Representation Learning

Learning a suitable lower-dimensional *embedding* or *representation* of a high-dimensional input space is a fundamental component of machine learning research [59]. The embedding can represent a medical concept [76] or summarise a patient's hospital visit [77]. It often performs better than the raw input for learning subsequent tasks [78–80]. We now provide an overview of the state-of-the-art techniques for obtaining embeddings in related medical applications: (i) standard dimensionality reduction techniques, (ii) distributed representations used in language modelling, (iii) using embedding layers as part of a larger model, or (iv) through the latent space of autoencoders and their variants. Such compact representations are then further used as inputs for classification or clustering purposes (covered in Section 2.2).

Standard Dimensionality Reduction Techniques

One of the most popular statistical dimensionality reduction techniques is principal component analysis (PCA) [81]. PCA transforms a set of possibly correlated variables to a set of linearly uncorrelated components. It has been used to extract features for various clinical applications [7, 61, 82], such as for the detection of hypotensive episodes [40], mortality prediction across stroke patients [83], or prediction of hospital readmission [84]. The main limitation of PCA is that it extracts linear features that may not well represent non-linear relationships present in complex clinical data [59]. Another popular technique is independent component analysis (ICA) [85, 51], which transforms the variables to set of independent components.

Distributed Concept Representations

Patient records may contain discrete categorical codes, such as diagnosis, medication, or treatment codes. Several studies [55, 53, 86] propose learning from such variables using embedding techniques derived from the distributional hypothesis in semantic modelling. The distributional hypothesis states that words that appear in similar contexts in large samples of language data are semantically similar [87]. The skip-gram algorithm learns the co-occurrence of information inside a context window of a fixed size [88]. It has been used to convert medical codes to dense representations in [47, 76, 55]. Similar to skip-gram, the Global Vectors (GloVe) algorithm was also used to learn the global co-occurrence matrix of medical codes [89].

Embedding Layers

Embedding layers can also be integrated as part of a larger model to transform high-dimensional features into a lower-dimensional space. The embedding can consist of a simple linear transformation [90, 91] or as a fully-connected (deep) network [33, 90, 86].

One study projected the input into a higher-dimensional space using a convolutional layer [53].

Autoencoders and their variants

An autoencoder is a neural network architecture that is often used for dimensionality reduction or feature extraction [92]. It first transforms the input space to a (typically noise-free) lower-dimensional representation using an encoder, and then reconstructs the input from this compact representation. The sparse autoencoder (SAE) enforces a sparsity constraint on the learned representation, and it has been used to learn latent representations of clinical data [44, 77]. The denoising autoencoder (DAE) reconstructs the input from a partially corrupted version of the input. The stacked DAE, which consists of several autoencoders that are initially pre-trained independently then connected in one network, has also been used for clinical applications [93, 51, 73, 94]. Another popular variant of autoencoders is the variational autoencoder [95], which is a generative model that learns a probabilistic latent space, unlike the previously mentioned discriminative autoencoders.

In Table 2.1, we summarise the feature extraction techniques in related studies. In terms of variable selection, we observe that free clinical text is the least-used input. That may be due to the availability of limited datasets. We also note that representation learning has become more popular from approximately 2013 and on wards, and we expect it to continue to be an active area of future research. The consistent use of hand-crafted features over the years indicates its effectiveness in training machine learning models. Additionally, time-series modelling of the variables is not widespread as it requires hyperparameter tuning and high computational resources.

Table 2.1: Overview of feature representation techniques adopted in related works using a variety of predictor variables: vital signs (VS), laboratory tests (LT), demographic information (DI), diagnostic codes (DC), interventions (INT) such as procedures and medications, and free text (TEX).

		Predictor Variables						Feature Representation		
Ref	Year	VS	LT	DI	DC	INT	TEX	Hand-crafted Features	Time-series Modelling	Representation Learning
[39]	2008	✓						✓		
[40]	2010	✓		✓		✓		✓		
[41]	2012	✓		✓				✓		
[42]	2012	✓		✓				✓		
[65]	2012	✓							✓	
[43]	2013	✓							✓	
[44]	2013		✓						✓	✓
[46]	2014			✓			✓	✓		
[48]	2015	✓	✓					✓		
[49]	2015	✓					✓	✓	✓	
[47]	2015				✓	✓				✓
[36]	2016	✓	✓	✓				✓		
[51]	2016		✓	✓	✓	✓	✓			✓
[77]	2016			✓	✓	✓				✓
[76]	2016		✓		✓	✓				✓
[35]	2017	✓	✓	✓	✓	✓		✓		
[54]	2017	✓	✓	✓	✓	✓			✓	
[55]	2017			✓	✓	✓				✓
[96]	2017				✓	✓		✓		
[52]	2018	✓		✓				✓		
[56]	2018	✓	✓	✓	✓	✓	✓			✓
[33]	2019	✓	✓	✓	✓	✓	✓	✓		✓

2.2 Predictive Inference

The extracted features are then used to train the model to perform a particular task related to clinical deterioration. The task is posed either as a classification (Section 2.2.1) or clustering (Section 2.2.2) problem.

2.2.1 Outcome Classification Framework

Machine learning models for classification are trained to detect a clinical outcome or a diagnosis. Table 2.2 provides an overview of the presented models and the clinical outcomes defined in recent papers. While we only list the base model considered in each study, current practice involves comparing the performance of the base model to the performance of traditional machine learning models. Popular baselines include logistic regression or the multi-layer perceptron [56, 91]. We also observe that all outcomes are defined within a particular time-frame, such as the next 48 hours for real-time scores [33] or 6 months for long-term prognostic models [55]. The lack of consistency in outcome definition makes it challenging to compare methods directly. Additionally, some studies tend to focus on specific patient subgroups, such as paediatrics [52], due to the datasets that are available for analysis.

Regression Models

Logistic regression is one of the simplest linear classifiers [97] and is often considered as a standard benchmark for sophisticated clinical models [98]. Previous studies used logistic regression to predict hemodynamic instability [39], imminent mortality [99], or the composite outcome of cardiac arrest, unplanned ICU admission, and mortality [12]. However, logistic regression cannot learn non-linear relationships and assumes independence across the input variables.

Decision tree learning involves the stratification of the feature space based on a criterion defined by information theory, such as entropy. One study developed an early warning score based on decision trees, using seven routinely-collected laboratory tests [29], while another constructed an ensemble model with gradient tree boosting and adaptive boosting to predict the likelihood of transfer to paediatric ICU [52]. Despite the high interpretability of the aforementioned studies, they heavily rely on task-specific hand-engineered features and do not learn complex patterns in the data.

Kernel Methods

Kernel methods rely on a user-defined kernel function that estimates the 'similarity' between pairs of data [100]. The support vector machine is a popular example of kernel methods. It projects data into a higher-dimensional space and finds the optimal discriminatory hyper-planes between classes [101]. The use of support vector machines heavily relies on the choice of the kernel and regularisation, and they have shown strong performance in recent clinical applications [42, 102, 103, 48]. Computing the kernel matrix for all pairs of data may be computationally expensive for large clinical datasets especially when a non-linear kernel is used. Further work must investigate approximation techniques for applications involving medical data.

Deep Learning

Deep learning models are also becoming increasingly popular for outcome prediction tasks [38, 36, 34, 41, 54]. The simplest form of neural networks is the multi-layer perceptron (MLP), which consists of fully-connected perceptrons. The main limitation of the MLP is its inability to account for temporal dependencies. Recurrent neural networks and their variants seek to model temporal behaviour through feedback connections. Both *Long Short Term Memory* (LSTM) networks [104, 105, 54] and

Table 2.2: Overview of classifiers used for outcome prediction in related works.

Model	Outcome	References
Novelty detection	ICU readmission	[43]
Logistic regression	Hemodynamic instability 2 hours in advance	[39]
	Gout vs. acute leukaemia	[44]
	Mortality on the same or next day	[99]
Support vector machine	Cardiac arrest within 72 hours	[42, 48]
	Mortality within 72 hours	[42, 49, 45]
Random forest classifier	Diseases within one-year interval	[51]
Support ensemble boosting	Paediatric transfer to ICU	[52]
Gaussian process classier	Cardiac arrhythmia	[61]
Multi-layer perceptron	ICU transfer and cardiac arrest	[36]
	Hypotensive episodes	[40, 41]
	Ventricular tachycardia	[34]
	In-hospital mortality	[57]
Convolutional neural network	Congestive heart failure after 6 months	[50]
	COPD after 6 months	[50]
	Hospital readmission after discharge	[96]
Recurrent neural networks	Mortality	[35]
	Acute kidney injury	[33]
	Diagnosis & medication codes for next visit	[47]
Gated recurrent units	Heart failure	[55]
	Multi-label diagnoses	[47]
Long short term memory networks	Sepsis at least 4 hours in advance	[54]

Gated Recurrent Units (GRU) [90, 55] were constructed to predict (and alert in advance of) clinical outcomes. There is also a growing interest in developing 'end-to-end' architectures that can jointly extract features and perform classification [91, 96, 106]. Although deep learning techniques are typically characterised by strong performance, their decision-making process lacks interpretability. This is an area of active research and we will present current solutions in Section 2.3.2.

2.2.2 Clustering for Abnormality Detection

Clustering algorithms are unsupervised learning techniques that group data based on similarity measures. Within the context of clinical deterioration, this can involve creating a 'dictionary' or cluster of healthy patients and computing a similarity metric for a new patient [60, 68, 107]. Popular similarity metrics are the Kullback-Leibler (KL) divergence [108] and the Mahalanobis distance [60]. Clustering analysis has also been useful for patient phenotyping [44].

The concept of creating patient dictionaries is a subset of novelty detection. An example of such approaches is 'one-class classification' [109, 63].

2.3 Evaluation of Early Warning Models

The performance of supervised early warning models on the *testing set* is evaluated using various statistical methods. Those statistical methods mainly assess the performance of the model in terms of accuracy metrics. In recent years, model interpretability has also become an area of interest as it directly reflects how we translate technologies into clinical practice [110].

2.3.1 Performance Metrics

Model discrimination refers to the model's ability in separating classes of interest. In the context of clinical deterioration, we will here refer to deteriorating patients as the *positive class* and healthy patients as the *negative class*. Many machine learning models are trained to return the probability of the positive class, especially for deep learning models with a sigmoid function as the last layer. It is common practice to convert the regression values to binary values by fixing a decision threshold. The predictions are then compared to the true labels and classified into one of four categories:

- **True Positives (TP):** model correctly predicts the positive class.
- **True Negatives (TN):** model correctly predicts the negative class.
- **False Positives (FP):** model incorrectly predicts the positive class.
- **False Negatives (FN):** model incorrectly predicts the negative class.

Accuracy, which summarises the proportion of correctly classified samples amongst all samples, is highly biased when using highly imbalanced datasets. Therefore, other metrics are usually considered. **Sensitivity**, or the True Positive Rate (TPR), assesses the model's ability to correctly predict the positive class.

$$TPR = \frac{TP}{TP + FN} \tag{2.2}$$

Specificity, also known as the True Negative Rate (TNR), assesses the model's ability to correctly predict the negative class.

$$TNR = \frac{TN}{TN + FP} \tag{2.3}$$

The receiving operator characteristic (ROC) curve plots the TPR on the vertical axis and (1-TNR), also known as the False Positive Rate (FPR), on the horizontal axis. The integral under the curve is the **Area Under the Receiving Operator Characteristic Curve (AUROC)** [111].[2] The AUROC assesses the model's overall diagnostic ability as the decision threshold is varied, where an AUROC of 0.5 means that the model is making predictions at random in a two-class setting. The authors of state-of-the-art EWS systems [112] claim that, for track-and-trigger systems, an AUROC higher than 0.8 implies that the model has good diagnostic ability while an AUROC higher than 0.9 means that the model has excellent diagnostic ability [112].

Precision, also known as the Positive Predictive Value (PPV), assesses the proportion of correctly predicted positive class across all of the true positive class.

$$PPV = \frac{TP}{TP + FP} \tag{2.4}$$

The Precision-Recall curve, where recall is essentially sensitivity, plots the TPR on the horizontal axis and the Precision on the vertical axis and integrates the area under the curve. The integral under the curve is the **Area under Precision-Recall Curve (AUPRC)**. Unlike the AUROC, the AUPRC and PPV are highly sensitive to class imbalance. EWS systems are in general characterised with low AUPRC and PPV values [28]. Due to low PPV values, such systems should be considered as risk stratifiers rather than predictors [40].

There are other commonly assessed metrics, such as the F1-score [113, 38] and the likelihood ratio [114]. Some studies also report the false positives to true positives ratio [33] and the inverse of the PPV known as the work-up-to-detection ratio [115, 56]. The efficiency curve [25, 29] is a qualitative summary that plots the number of positives generated at different decision thresholds against the sensitivity of the model. This

[2]Some papers refer to the AUROC as the 'concordance-statistic' (C-statistic).

tool is essential to evaluate the trade-off between the total number of alerts and the false alerts, and is typically used in the evaluation of early warning systems.

2.3.2 Interpretability

Despite the good performance of recently introduced machine learning models, interpretability remains to be a challenge for their clinical utility [110]. There are various definitions of interpretability in existing literature and they refer to several distinct ideas [116, 117]. Most of these ideas pertaining to the clinical domain revolve around trustworthiness of the results and transparency of the model. In the context of this review, we summarise the efforts of outcome prediction models that considered interpretability as a key component of model assessment.

Mimic learning assumes that shallow models, such as linear models, are interpretable. It aims to identify the features that are potentially relevant to the prediction. It involves first training a deep learning model for a specific clinical task. It then trains a shallow model, such as gradient boosting trees, to mimic the behaviour of the deep learning model [94, 118]. The local interpretable model-agnostic explanation (LIME) [119] generates a local explanation of the model behaviour using a shallow model. It has been even used to explain machine learning models for the prediction of in-hospital mortality [120]. However, it has also been argued that linear models, rule-based models, and decision trees are not intrinsically interpretable [116]. Other post-hoc interpretability techniques such as *saliency maps* rely on qualitative visual interpretations commonly used in computer vision applications.

It is often argued that deep learning models compromise interpretability for high accuracy [121]. Thus, there have been recent breakthroughs in developing inherently interpretable deep learning models instead of performing post-hoc interpretation [9]. For instance, attention mechanisms are incorporated within deep learning models

and assign normalised weights to a set of features. The weights indicate the feature importance for the prediction of a future diagnosis [106, 53, 89] or high risk vascular diseases [113]. Other works impose non-negativity [77] or sparsity [44] constraints on the learned embedding space of medical data.

Chapter 3

Summary of Datasets

3.1 Introduction

In this chapter, we provide a brief summary of the opportunities and challenges of using data from EHRs. We provide an overview of the characteristics of the datasets used to develop and validate our models, where the data were extracted from the Oxford University Hospitals (OUH) and Portsmouth Hospitals NHS Trust (PH), within the 'HAVEN' project. We empirically assess the datasets in terms of completeness and correctness; we then present methods of pre-processing for to the different variable types available. Local research ethics committee approvals were obtained for accessing the datasets.

3.2 Electronic Health Records

EHRs are used in hospitals to store longitudinal information of patients collected in a care delivery setting. Such information includes patient demographics, vital signs, medications, laboratory data, and description of any adverse events that may have occurred to the patient during hospitalisation. These can be categorised as one of four

data types: categorical data, time series data, textual data, and visual data. In this thesis, we focus on developing techniques that use categorical and time series data due to their availability in our datasets.

The heterogeneous nature of EHRs and the fact that they may reflect the recording process present in the hospital rather than being a direct reflection of patient physiology [122] imposes challenges on using EHRs for research. First, EHRs are complex as they may include structured and unstructured data; an example of the latter is textual information which could require natural language processing techniques to process [123]. Additionally, categorical data, such as diagnostic coding (described later), may adopt different coding systems across different institutions.

Another important dimension is data completeness, which may be defined as "the proportion of observations that are actually recorded in the system" [124]. Incompleteness of EHRs can be a result of health service fragmentation due to inefficient communication following patient transfer among institutions; the recording of data taking place only during healthcare episodes that correspond to illness, or the increased personalisation of attributes per patient [122, 125]. Completeness may also vary across institutions based on adopted protocols.

The third challenge is the accuracy of the data, or "the proportion of recorded observations in the system that are correct" [124]. Errors can occur while clinical staff observe a patient or record data, and their occurrence may be influenced by random and systematic errors such as billing requirements or avoidance of liability [122]. The accuracy of EHRs can be assessed by checking agreement between different elements within the EHR (such as assigned diagnosis and supplied medications), or by verifying whether values are within expected ranges [126].

Finally, it is important to verify whether the data was recorded within a reasonable period of time [126]. For example, the recorded collection time of vital signs may

precede the time of admission. Although this aspect of data quality is highly dependent on the efficiency of the clinical staff, it also depends on the work flow protocols adopted at different institutions. Timeliness of data must be assessed to evaluate the chronology of data elements in relation to admission or discharge decisions, for example laboratory results prior to admission may be considered as part of subsequent admission, or death within 24 hours of discharge can be considered as in-hospital mortality.

We here provide an overview of key terminology used throughout the thesis. *Subject* refers to an individual patient while *admission* or *episode* refers to an in-hospital stay, such that a subject could have multiple hospital admissions. A hospital admission can be an emergency or elective admission. An elective admission is one that has been "arranged in advance", while an emergency admission is one that is "unpredictable and at short notice due to clinical need" [127]. A *vital-sign set* or *observations set* is a group of routinely-collected vital signs that are measured at the same time. A *unique laboratory test* is the measurement of a test that must be ordered, such as by collecting a blood specimen sample from the patient to measure the amount of red blood cells.

3.2.1 Inclusion Criteria

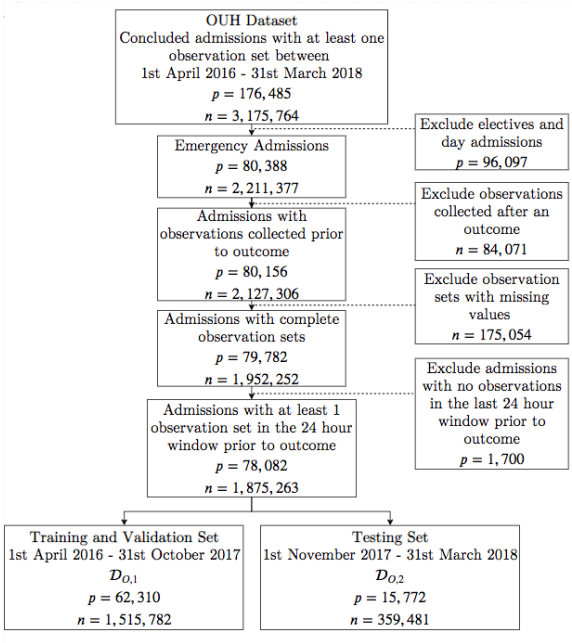
In our work, we are interested in developing models for acutely-ill patients (emergency admissions), who represent the cohort that will more likely experience adverse events in comparison to patients who are admitted for less acute conditions (elective admissions). We would also like to develop our alerting systems using a comparable cohort to the one used in the development of the scoring system that is currently used in practice (i.e., NEWS [25]). We therefore exclude elective and maternity admissions, and data from patients who were deemed sufficiently well to be discharged alive before midnight on the day of admission. It is important to note that maternity admissions have their own early warning systems. Some patients may experience multiple outcomes during

their hospital stay. We therefore exclude observations recorded after a patient had experienced their first outcome, because deterioration would have already occurred and we are interested in alerting for deterioration before it occurs.

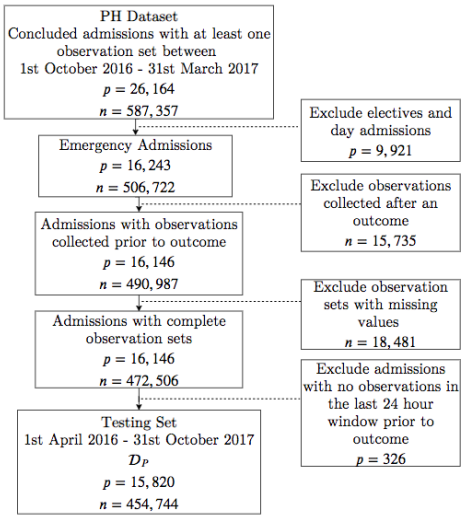
We also exclude admissions with no observation sets recorded within the last 24 hours prior to an event, as such patients were likely to be on terminal care pathways. Patients on terminal care pathways are usually assigned 'Do Not Resuscitate' (DNR) orders, which instruct clinical staff to not provide any emergency medical care. We further excluded any admission with missing age or sex, and observation sets with missing or implausible values, to avoid making any assumptions that may incur bias while developing our models. The plausible ranges for vital signs and laboratory tests are shown in Table A.1. The plausible ranges were chosen based on clinical expertise, existing published clinical reference ranges [128, 29], and by assessing the frequency of extreme values of the considered variables. This approach aligns with the cohorts that were used to develop existing EWS systems [30, 25]. Our criteria were applied during the extraction of the datasets described below.

3.2.2 Oxford University Hospitals Datasets

We extracted a dataset from routinely collected observations of adult patients admitted to the Oxford University Hospitals NHS Trust, retrieved from the "Hospital Alerting Via Electronic Noticeboard" (HAVEN) project (REC reference: 16/SC/0264 and Confidential Advisory Group reference 08/02/1394). The data was originally collected using the 'System for Electronic Notification and Documentation' (SEND) [129] at four hospitals in Oxfordshire, United Kingdom. Within the OUH, the John Radcliffe Hospital and Horton General Hospital provide acute medical and surgical services; Churchill Hospital specialises in cancer care; and the Nuffield Orthopaedic Hospital provides services for musculoskeletal problems. Since all four hospitals belong to the



(a)



(b)

Fig. 3.1: Consort diagram visualising the extraction of the final patient cohorts from the (a) Oxford University Hospitals (OUH) dataset and (b) Portsmouth Hospitals NHS Trust (PH) dataset, where p is the number of patient admissions and n is the number of corresponding observation sets.

OUH NHS Trust, we assumed that they represent a single centre. The extracted dataset contained vital signs, laboratory tests, demographics and data of the occurrence of any adverse events. In keeping with the medical informatics literature, we then split the dataset based on time as recommended in TRIPOD guidelines [130] into $\mathcal{D}_{O,1}$ (1 April, 2016 - 31 October, 2017), which was used as a training and validation set, and $\mathcal{D}_{O,2}$ (1 November 2017 - 31 March 2018), which was used as a testing set. The characteristics of these datasets will be presented later.

3.2.3 Portsmouth Hospitals NHS Trust Dataset

The second dataset, denoted as $\mathcal{D}_P$, was extracted from routinely-collected observations of adult patients admitted to the Queen Alexandria Hospital, also retrieved from the HAVEN database. Queen Alexandria Hospital, the single hospital serving the Portsmouth Hospitals (PH) NHS trust, is a large district general hospital offering acute and specialist services. The PH dataset was collected using the VitalPAC system [22]. In our work, $\mathcal{D}_P$ (1 October, 2016 - 31 March, 2017) was used as an external and independent testing set. It contained the same variables as the OUH datasets, with extensive curation from researchers within the HAVEN consortium to ensure that this was the case. The data extraction flow charts of all datasets are shown in Fig. 3.1.

3.3 Extracted Variables

In this section, we describe the main variables retrieved from the datasets and compare their distributions across the three datasets.[1]

3.3.1 Demographics

We extracted demographic information such as age at admission, sex and the Charlson Commorbidity Index, which estimates the risk of ten-year morbidity from comorbid conditions [131].

The characteristics of the retrieved datasets are summarised in Table 3.1. The mean age is highest in $\mathcal{D}_P$ (73.4 years), compared to 68.6 years in $\mathcal{D}_{O,1}$ and 69.0 in $\mathcal{D}_{O,2}$. The proportion of the eldest population ($>$ 60 years old) is the highest in all datasets (58.9% in $\mathcal{D}_{O,1}$, 60.5% in $\mathcal{D}_{O,2}$, and 66.5% in $\mathcal{D}_P$) across all age groups. The average number of vital-sign measurements per admission was approximately 24, 23, and 29 across $\mathcal{D}_{O,1}$, $\mathcal{D}_{O,2}$, and $\mathcal{D}_P$, respectively. The average number of unique laboratory tests per admission was 52, 49, 59 across $\mathcal{D}_{O,1}$, $\mathcal{D}_{O,2}$, and $\mathcal{D}_P$, respectively.

The results of the statistical testing for demographics variables (i.e., age, sex, length of stay, and CCI) showed that the training set ($\mathcal{D}_{O,1}$) and the testing sets ($\mathcal{D}_{O,1}$ and $\mathcal{D}_{O,2}$) have different distributions. This topic is further studied in the context of dataset shift as a phenomenon also known as dataset shift [132]. This will be important to consider throughout the thesis while evaluating the performance of the models developed using the training set on the testing sets.

[1] We conducted a Chi-squared test for categorical variables to test for equal proportion of categorical variables and a non-parametric two-sided Mann-Whitney U test for the continuous variables with the null hypothesis that the datasets have equivalent expected values. As for the one-sided distributions, we used the Wilcoxon rank sum to test for significance. The p-values were adjusted for multiple comparisons using the Bonferroni correction (i.e., dividing the p-value by the total number of comparisons), where the null hypothesis was rejected when $p > 0.002$. Although the distributions of most variables looked broadly similar by eye, most statistical testing rejected the null hypothesis that the distributions of the datasets are identical, which is assumed throughout unless otherwise stated.

Table 3.1: Summary of the patient demographics across the $\mathcal{D}_{O,1}$, $\mathcal{D}_{O,2}$, and $\mathcal{D}_P$ datasets.

Characteristic	$\mathcal{D}_{O,1}$	$\mathcal{D}_{O,2}$	$\mathcal{D}_P$
Subjects, n	42,716	12,859	12,773
Admissions, n	62,310	15,772	15,820
Vital Signs Sets, n	1,515,782	359,481	454,744
Unique Laboratory Results, n	3,220,097	775,089	940,192
Age, mean (SD)	68.6 (18.9)	69.0 (19.1)	73.4 (17.3)
Age groups, n (%) (per admissions)			
$16-45$ years old	14,629 (23.5)	3,568 (22.6)	2,772 (17.5)
$46-60$ years old	11,005 (17.7)	2,664 (16.9)	2,528 (16.0)
>60 years old	36,676 (58.9)	9,540 (60.5)	10,520 (66.5)
Females, n (%) (per admissions)	32,064 (51.5)	8,276 (52.5)	8,268 (52.3)
Length of stay (hr), mean (SD)	159.7 (258.9)	141.7 (204.6)	197.1 (296.1)
Charlson Commorbidity Index, median (IQR)	3 (0-11)	3 (0-12)	4 (0-13)

3.3.2 Outcomes

The literature reporting models that detect clinical deterioration varies in terms of outcome definition [11]. Previously investigated discrete clinical outcomes include sepsis, ICU admission, post-surgery relapse, and mortality [20, 28, 54]. To build a model that is comparable to EWS systems that are currently implemented in clinical practice, we defined our primary outcome as the composite outcome of unplanned ICU admission, cardiac arrest, or mortality within 24h of a ward observations set. In the case of multiple occurrences of adverse events, we only took account of the timing of the first occurring event.

Unplanned ICU admission occurs when the patient deteriorates to the point that they require intensive care, which would have been unexpected at the point of admission. Cardiac arrest (CA) occurs when the palpable pulse is lost as the heart suddenly stops pumping blood. The main limitations of adopting CA as an outcome is that 'do not attempt resuscitation' (DNAR) information is not available in our datasets. However, a previous study showed that the majority of treated in-hospital CAs are potentially

avoidable, which highlights the importance of detecting for CA [1]. The occurrence of ICU admission, CA, and mortality in terms of date and time were present in all datasets. We extracted this data and the prevalence of the outcomes across all datasets is presented in Table 3.2.

We note that in-hospital mortality constituted the highest proportion of outcomes in all the datasets (55.4% in $\mathcal{D}_{O,1}$, 68.3% in $\mathcal{D}_{O,2}$, and 67.1% in $\mathcal{D}_P$), while CA constituted the lowest proportion of outcomes (6.7% in $\mathcal{D}_{O,1}$, 8.4% in $\mathcal{D}_{O,2}$, and 10.7% in $\mathcal{D}_P$). The prevalence of a composite outcome was highest in $\mathcal{D}_P$ (5.1%), compared to 3.9% and 3.7% in $\mathcal{D}_{O,1}$ and $\mathcal{D}_{O,2}$, respectively. This may be due to the fact that Portsmouth Hospitals contained the most elderly population.

Across the younger age groups, the highest number of outcomes was unplanned ICU admission in all datasets (215 in $\mathcal{D}_{O,1}$, 31 in $\mathcal{D}_{O,2}$, and 22 in $\mathcal{D}_P$ the 16-45 years old group and 218 in $\mathcal{D}_{O,1}$, 35 in $\mathcal{D}_{O,2}$, and 48 in $\mathcal{D}_P$ in the 46-60 years old group). Across all datasets, the eldest age group (>60 years old) constituted the highest proportion of the mortality outcome (94.2% in $\mathcal{D}_{O,1}$, 94.8% in $\mathcal{D}_{O,2}$, and 95.7% in $\mathcal{D}_P$).

Table 3.2: Number of outcomes experienced based on patient admissions across the $\mathcal{D}_{O,1}$, $\mathcal{D}_{O,2}$, and $\mathcal{D}_P$ datasets. Percentages represent the population proportion within each outcome unless stated otherwise.

Outcome	$\mathcal{D}_{O,1}$	$\mathcal{D}_{O,2}$	$\mathcal{D}_P$
Unplanned ICU Admission			
Total, n (% per admissions)	915 (1.5)	137 (0.9)	185 (1.2)
16-45 years old, n (%)	215 (23.5)	31 (22.6)	22 (11.9)
46-60 years old, n (%)	218 (23.8)	35 (25.5)	48 (25.9)
$>$ 60 years old, n (%)	482 (52.7)	71 (51.8)	115 (62.2)
Cardiac Arrest			
Total, n (% per admissions)	161 (0.3)	49 (0.3)	88 (0.6)
16-45 years old, n (%)	8 (5.0)	0 (0)	1 (1.1)
46-60 years old, n (%)	17 (10.6)	5 (10.2)	8 (9.1)
$>$ 60 years old, n (%)	136 (84.5)	44 (89.8)	79 (89.8)
In-hospital Mortality			
Total, n (% per admissions)	1,338 (2.1)	500 (2.5)	556 (3.5)
16-45 years old, n (%)	16 (1.2)	4 (0.8)	3 (0.5)
46-60 years old, n (%)	61 (4.6)	22 (4.4)	21 (3.8)
$>$ 60 years old, n (%)	1,261 (94.2)	374 (94.8)	532 (95.7)
Composite outcome			
Total, n (% per admissions)	2,414 (3.9)	586 (3.7)	803 (5.1)
16-45 years old, n (%)	239 (9.9)	35 (6.0)	24 (3.0)
46-60 years old, n (%)	296 (12.3)	62 (10.6)	74 (9.2)
$>$ 60 years old, n (%)	1,879 (77.8)	489 (83.4)	705 (87.8)
Unplanned ICU Admission, n (%)	915 (37.9)	137 (23.4)	178 (22.2)
Cardiac Arrest, n (%)	161 (6.7)	49 (8.4)	86 (10.7)
Mortality, n (%)	1,338 (55.4)	400 (68.3)	539 (67.1)

3.3.3 Vital Signs

Physiological parameters, also referred to as vital-sign variables, are powerful predictors of clinical deterioration [2, 22]. Previous studies have mainly used vital signs in EHRs to develop and validate EWS systems [25, 28] or to understand characteristics of different patient sub-groups, such as differences between elderly and non-elderly patients prior to cardiac arrest [133]. The most commonly collected vital signs that constitute the basis of most EWS systems are heart rate, systolic blood pressure, respiratory rate, oxygen saturation, level of consciousness, temperature, and the provision of supplemental oxygen [23]. Given the empirically-proven performance of physiology-based EWS systems, which form the majority of those in use in clinical practice, we consider the vital signs as key predictor variables in all of our models and summarise their relevance below.

- **Heart Rate** (HR) ($beats/minute$) is a measurable characteristic of the pulse, which is produced by the contraction and relaxation of the heart [134]. It is affected by various factors such as age, medications, or existing medical conditions. A low or high heart rate raises the risk of health complications, such as heart failure or blood clots.

- **Respiratory Rate** (RR) ($breaths/minute$) is a sensitive indicator of critical illness, including acidosis, respiratory distress and hypoxaemia [134]. Previous studies have also found RR is an important predictor of cardiac arrest [135], mortality [136], and ICU admission [1].

- **Systolic Blood Pressure** (SBP) ($mmHg$) is a reflection of blood flow when the heart is contracting. High blood pressure, or hypertension, is a major risk factor of cardiovascular disease, stroke, and mortality [137].

- **Temperature** (TEMP) ($°C$) is a measure of the body's thermoregulation that is influenced by underlying pathophysiology or age [134]. TEMP is used to characterise the diagnosis or to monitor the course of a particular disease. However, it can be recorded at different anatomical sites since no single measurement site is recommended as best practice.

- **Oxygen Saturation** (SPO_2) is an estimate of the amount of oxygen in the arterial blood. It is measured using a pulse oximeter and the measurements may be biased by patient movement or underlying disease. It is recommended to assess SPO_2 with knowledge of other vital signs, such as RR or SBP [134].

- **Level of Consciousness** is recorded using the Alert, Voice, Pain and Unresponsive (AVPU) Score to assess the patient's neurological function. The limitation of the AVPU score is that it does not factor in confusion of the patient [23]. For example, a dementia patient may score 'A' despite not being fully oriented. When the patient's consciousness level was assessed only using the Glasgow Coma Scale (GCS), we converted the GCS score to an AVPU score [25]. If the GCS value was also missing, we assumed that the patient was 'Alert'.

- **Supplemental Oxygen** is a binary variable that indicates whether or not supplemental oxygen is being provided to a patient. The performance of EWS systems has been shown to increase when this variable is included [28]. When the provision or absence of supplemental oxygen was missing, we assumed that supplemental oxygen was not provided.

The characteristics of the vital signs across all datasets are summarised in Table 3.3, in terms of median, interquartile range, and the number of incorrect values. Incorrect values refer to measurements that do not fall within plausible physiological ranges (the ranges are summarised in Table A.1 in Appendix A for reference). We observe that the

proportion of incorrect values is lower than 1% across all datasets, which highlights the feasibility of the measurements. The highest proportions of incorrect values were in $\mathcal{D}_P$. For example, there were 1,529 HR measurements less than 10 beats/min and 1,946 TEMP measurements less than 15 °C. We considered those observations as outliers due to systematic or random errors.

As shown in Table 3.3, the median and interquartile ranges of the vital-sign variables were comparable. However, the distributions of all variables significantly differed across all datasets based on statistical testing. The lowest HR median was 79 in $\mathcal{D}_P$, compared to 80 in $\mathcal{D}_{O,1}$ and 81 in $\mathcal{D}_{O,2}$. The lowest SBP was 124 in $\mathcal{D}_P$, compared to 126 in $\mathcal{D}_{O,1}$ and $\mathcal{D}_{O,2}$. Although the median (IQR) for RR is identical in all datasets, i.e., 17 (16-18), the distributions (shown in Fig. A.1) differ significantly since RR is manually collected by nurses and as such is subject to measurement changes due to human errors. We compared the distributions of all of the the vital signs visually (refer to Fig. A.1 in Appendix A). The significant differences in the distributions of the vital signs across the three datasets relates to the concept of domain shift as discussed earlier. For example, it may be due to differences in measurement systems used in the different datasets.

Table 3.3: Characteristics of the vital signs across the $\mathcal{D}_{O,1}$, $\mathcal{D}_{O,2}$, and $\mathcal{D}_P$ datasets, in terms of the median and interquartile range of the distributions after accounting for missing and physiologically implausible values. n represents the number of measurements.

Vital Sign, unit	$\mathcal{D}_{O,1}$	$\mathcal{D}_{O,2}$	$\mathcal{D}_P$
Heart rate, beats/min			
median (IQR)	80 (70-91)	81 (70-92)	79 (69-89)
incorrect, n (%)	0 (0)	0 (0)	1,529 (0.3)
Systolic blood pressure, mm Hg			
median (IQR)	126 (111-142)	126 (112-142)	124 (110-140)
incorrect, n (%)	0 (0)	0 (0)	2,269 (0.5)
Temperature, °C			
median (IQR)	36.4 (36.0-36.8)	36.4 (36.0-36.8)	36.7 (36.4-37.0)
incorrect, n (%)	0 (0)	0 (0)	1,955 (0.4)
Respiratory Rate, breaths/min			
median (IQR)	17 (16-18)	17 (16-18)	17 (16-18)
incorrect, n (%)	116 (0.008)	18 (0.005)	924 (0.2)
Oxygen Saturation, (%)			
median (IQR)	96 (95-98)	96 (95-98)	96 (95-97)
incorrect, n (%)	0 (0)	0 (0)	1,973 (0.4)
Level of Consciousness[a]			
Alert, n (%)	1,482,983 (97.8)	352,367 (98.0)	451,460 (99.3)
Voice, n (%)	27,033 (1.8)	5,895 (1.6)	2,025 (0.4)
Pain, n (%)	4,248 (0.3)	940 (0.3)	799 (0.2)
Unresponsive, n (%)	1,518 (0.1)	279 (0.08)	460 (0.1)
Oxygen Support[b]			
Provision of Oxygen, n (%)	236,602 (15.6)	60,037 (16.7)	64,862 (14.3)

[a] computed from available GCS scores when missing.

[b] computed from type of oxygen mask provided when missing.

3.3.4 Laboratory Tests

Recently introduced EWS systems also highlight the power of laboratory tests in alerting for adverse outcomes [29, 30]. Laboratory tests involve the collection of blood or urine specimens as ordered by a clinician to rule in or rule out a diagnosis, monitor the effect of intervention, screen for particular conditions, or for research purposes [138]. In our work, we investigate the routinely collected laboratory tests described below.

- **Haemoglobin** (HGB) is the number of red blood cells that transport oxygen to the body organs and carry back carbon dioxide to the lungs, measured by a blood test. Low HGB, also known as anaemia, is associated with higher risks of mortality, acute stroke, and poor outcomes [139]. Increased risk of mortality has also been shown to be associated with high extreme values of HGB.

- **Hematocrit** (HCT) measures the proportion of red blood cells in the total blood count. Low levels of HCT indicate anaemia, while sustained increases in HCT levels are associated with reduced patient morbidity [140].

- **White Blood Cells** (WBC), or leukocytes, are counted in blood tests to help detect infection that the immune system is trying to fight. Low WBC count may be caused by various risk factors, including viral infections, diseases that damage the bone marrow such as cancer, or medication. High WBC count has also been shown to be significantly associated with cardiovascular mortality amongst elderly patients [141].

- **Sodium** (Na) test is a blood test that measures the amount of sodium in the blood, an electrolyte that regulates the amount of water surrounding the cells and maintains blood pressure. It can be influenced by hormonal changes, dehydration, heart diseases, or age. Severely low Na levels, or hyponatremia, usually involves

dehydration and may be associated with mortality based on the underlying illness [142].

- **Potassium** (K) is also an electrolyte that is vital for regulating fluid volumes in cells and blood pH. High and low K levels were shown to be associated with increased risk of cardiovascular disease, and even mortality when evaluated across two independent cohorts with racial differences [143].

- **Albumin** (ALB) is a protein made by the liver that prevents fluid in the bloodstream from leaking. ALB levels are usually measured to assess liver and kidney function and test for heart disease [144], which ultimately affects clinical outcomes.

- **Urea** (UR), measured by urine or blood tests, is the metabolic waste product of protein breakdown, which is important for ammonia detoxification and water conservation [145]. Reduced or increased UR levels can indicate kidney disease, heart failure, or dehydration.

- **Creatinine** (CR) is a waste product generated by the breakdown of muscle tissue that specifically indicates kidney function. It is influenced by age, race, gender, and weight. Small increases in CR levels have also been shown to be significantly associated with mortality and length of stay in hospital [146].

- **Bilirubin** (BIL) is a yellow pigment in the blood that is produced by the breakdown of red blood cells. It is used as an indicator of anaemia, jaundice or liver disease. Increased BIL levels have been found to be as strong predictors of cardiovascular death, heart failure hospitalisation and mortality [147].

- **Troponin** (TROP) are proteins in the blood that measure contractions in the heart muscle and is an indicator of heart muscle damage [148].

- **C-Reactive Protein** (CRP) is a acute-phase protein released by the liver after tissue injury, such as sepsis or strokes, that indicates degree of infection or inflammation. CRP levels are predictors of mortality in patients with chronic respiratory failure [149] and of cardiovascular risk in patients with chronic obstructive pulmonary disease [150].

We summarised the laboratory-test features in terms of median and interquartile range and correctness. The results are shown in Table 3.5. In terms of correctness, we notice that incorrect values represent less than 1% of values in $\mathcal{D}_{O,1}$ and $\mathcal{D}_{O,2}$. In $\mathcal{D}_P$, the proportion of incorrect values reaches up to 5.6% for Potassium. By inspecting the raw data, we notice that there were 5,470 Potassium measurements equal to 0, representing the 5.6% incorrect values. This may be related to errors in the data collection or data curation process. In general, the majority of laboratory test measurements are feasible since they fall within plausible physiological ranges.

The distributions of the laboratory test variables are visualised in Fig. A.2 in Appendix A. Although the distributions look similar visually, the distributions were significantly different except for WBC and CRP across $\mathcal{D}_{O,1}$ and $\mathcal{D}_{O,2}$ and HGB across $\mathcal{D}_{O,2}$ and $\mathcal{D}_P$.

Table 3.4 shows the total number of laboratory tests measured in all datasets. We notice that HGB, Na, WBC, K, CR, HCT, and UR are the most frequently measured tests, constituting $> 10\%$ of all laboratory-test measurements in all three datasets. This is followed by ALB, BIL, and CRP, which constitute approximately 8% of total measurements. Finally, TROP is the least measured test ($< 1\%$) across the three datasets. Generally, laboratory tests are collected less frequently than vital-sign measurements because they are usually ordered for specific diagnostic purposes or based on the clinician's judgement of the patient status. For example, TROP is a protein that is released to the bloodstream when a heart attack occurs and is usually

only ordered by the clinician when the patient has chest pain or other symptoms of a heart attack.

Table 3.4: Total number of laboratory tests measured in the $\mathcal{D}_{O,1}$, $\mathcal{D}_{O,2}$, and $\mathcal{D}_P$ datasets, and the number of measurements with implausible values.

Laboratory test, n (%)	$\mathcal{D}_{O,1}$	$\mathcal{D}_{O,2}$	$\mathcal{D}_P$
Unique Lab Results, n	3,220,097	775,089	940,192
Haemoglobin			
total	352,025 (10.9)	84,940 (11.0)	96893 (10.3)
incorrect	0 (0)	0 (0)	0 (0)
Sodium			
total	347,399 (10.8)	82,914 (10.7)	97,146 (10.3)
incorrect	33 (0.009)	3 (0.004)	21 (0.02)
White Blood Count			
total	352,016 (10.9)	84,940 (11.0)	96,888 (10.3)
incorrect	0 (0)	0 (0)	0 (0)
Potassium			
total	347,593 (10.8)	61,653 (8.0)	83,879 (8.9)
incorrect	4 (0.001)	0 (0)	5,470 (5.6)
Albumin			
total	253,032 (7.9)	61,653 (8.0)	83,879 (8.9)
incorrect	339 (0.1)	53 (0.09)	209 (0.2)
Creatinine			
total	348,588 (10.8)	83,206 (10.7)	97,086 (10.3)
incorrect	7 (0.2)	0 (0)	30 (3.1)
Urea			
total	347,283 (10.8)	83,206 (10.7)	97,086 (10.3)
incorrect	2 (0.0006)	0 (0)	651 (0.7)
Bilirubin			
total	236,420 (7.3)	58,568 (7.6)	81,542 (8.7)
incorrect	0 (0)	0 (0)	0 (0)
Hematocrit			
total	351,944 (10.9)	84,925 (11.0)	96,802 (10.3)
incorrect	0 (0)	0 (0)	0 (0)
Troponin			
total	23,810 (0.7)	4,795 (0.6)	7,191 (0.8)
incorrect	0 (0)	0 (0)	0 (0)
C-reactive Protein			
total	259,987 (8.0)	64,402 (8.3)	88,646 (9.4)
incorrect	0 (0)	0 (0)	0 (0)

Table 3.5: The median and interquartile range (IQR) of the laboratory tests in the $\mathcal{D}_{O,1}$, $\mathcal{D}_{O,2}$, and $\mathcal{D}_P$ datasets after accounting for incorrect values. *Note that we abbreviate 'litre' as 'L' in the units to prevent it from looking like the digit 1.*

Laboratory test, unit	$\mathcal{D}_{O,1}$	$\mathcal{D}_{O,2}$	$\mathcal{D}_P$
Hemoglobin, g/L	11.3 (9.5, 13.0)	11.5 (9.7, 13.2)	11.5 (10.0, 13.0)
Sodium, mmol/L	138 (135, 140)	138 (135, 140)	136 (134, 139)
White Blood Count, 10^9 cells/L	9.3 (6.8, 12.5)	9.3 (6.8,12.7)	9.0 (6.7, 12.2)
Potassium, mmol/L	4.0 (3.7,4.3)	4.0 (3.7,4.3)	4.1 (3.8,4.5)
Albumin, g/L	28 (23, 34)	28 (23, 33)	30 (25,35)
Creatinine, μmol/L	74 (58, 103)	72 (56, 101)	78 (59, 116)
Urea, mmol/L	5.6 (3.9, 8.9)	5.8 (4.0, 9.1)	6.5 (4.5,10.4)
Bilirubin, μmol/L	9 (6, 14)	9 (6, 14)	9 (6, 13)
Hematocrit, L/L (%)	0.3 (0.3, 0.4)	0.4 (0.3, 0.4)	0.3 (0.3, 0.4)
Troponin, mg/L	0.04 (0.04, 0.1)	0.04 (0.04, 0.1)	0.04 (0.04, 0.2)
C-reactive Protein, mg/L	42.1 (10.2, 112.4)	43.6 (10.5, 114.2)	39 (10, 107)

3.3.5 Diagnosis Codes

In both datasets, diagnosis codes assigned to the patient during hospital admission were classified according to the hierarchical International Classification of Diseases (ICD-10). The highest level of ICD-10 groups are shown in Table A.2. A patient usually has multiple diagnosis codes assigned during an episode.

The *primary diagnosis* is the main condition, symptom, or problem that a patient was treated for during admission. This variable is not available in the OUH and PH datasets. We therefore assume that the diagnosis code assigned at admission is the primary diagnosis. The frequency of assigned codes is summarised in Fig. 3.2. The results are similar for all datasets.

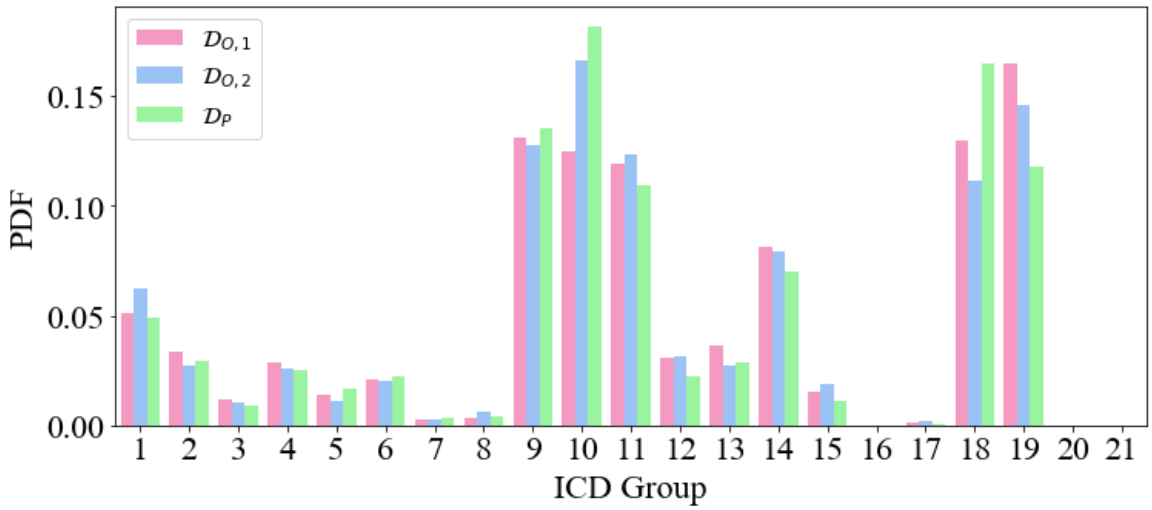


Fig. 3.2: Distribution of ICD-10 codes assigned to patients admitted to the Oxford University Hospitals (OUH) and Portsmouth Hospitals (PH).

We further provide a summary in Table 3.6 demonstrating the three most frequent diagnosis codes across all datasets.

Table 3.6: Summary of the ICD-10 codes assigned to hospital admissions, based on ICD-10 groups, across the $\mathcal{D}_{O,1}$, $\mathcal{D}_{O,2}$, and $\mathcal{D}_P$ datasets.

Characteristic	$\mathcal{D}_{O,1}$	$\mathcal{D}_{O,2}$	$\mathcal{D}_P$
Admissions with missing codes, n (%)	94	82	3
Most frequent ICD-10 group (%)	19 (16.5)	10 (16.5)	10 (18.1)
Second most frequent ICD-10 group (%)	9 (13.0)	19 (14.5)	18 (16.5)
Third most frequent ICD-10 group (%)	18 (12.9)	9 (12.6)	9 (13.5)
Strokes, n (%)	1,611 (2.6)	389 (2.5)	461 (2.9)
Diabetes, n (%)	514 (0.8)	121 (0.8)	155 (1.0)
Cancer, n (%)	2,078 (3.3)	431 (2.7)	462 (2.9)

3.4 Detection Framework

After accounting for missingness and incorrectness of the different data types as described earlier, we subsequently construct our models to detect clinical deterioration by following the general pipeline presented in Chapter 2. For the sake of clarifying the functionality of EWS systems, Fig. 3.3 visualises the vital signs of a patient who eventually experiences an outcome. Since at t=140 hours from admission (solid green line), an observation set is recorded that is within the 24 hour window of a composite outcome; thus, we would for this example expect the model to produce an alert.

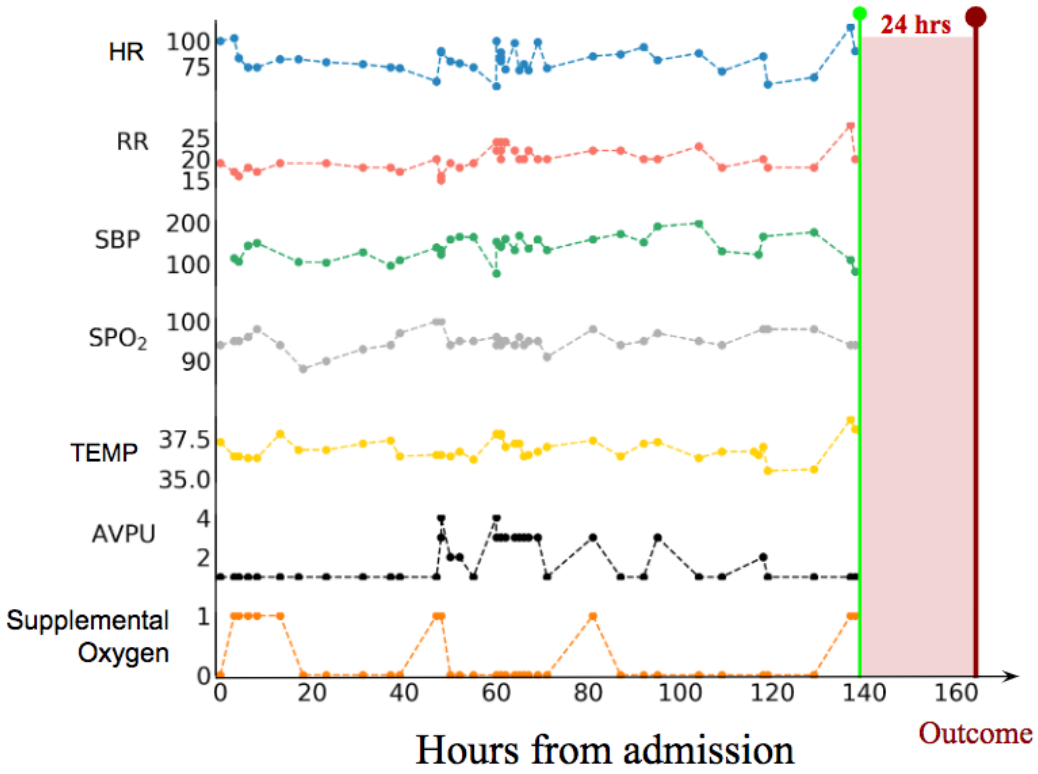


Fig. 3.3: Example of a patient admission that eventually experiences an outcome. At the solid green line, the model calculates the score that the just collected observations set is within 24 hours of an outcome. In this case, we expect the model to assign a high score and produce an alert.

One of the main assumptions of this thesis is that the proposed models are real-time, i.e., the models produce an alert every time the vital signs are measured as in

EWS systems currently used in practice. We then convert our problem into a binary classification task by defining two classes. The positive class is an observations set that is within 24 hours of a composite outcome, while the negative class is an observations set that is not within 24 hours of a composite outcome. We treat observation sets as independent data points. Overall, this is a commonly-used approach in terms of developing EWS systems [25, 28]. We summarise the distribution of the positive class in Table 3.7.

Table 3.7: Summary of the positive class prevalence in the $\mathcal{D}_{O,1}$, $\mathcal{D}_{O,2}$, and $\mathcal{D}_P$ datasets, in terms of number of observation sets, except where stated otherwise. The positive class represents observation sets that are within 24 hours of a composite outcome.

Characteristic	$\mathcal{D}_{O,1}$	$\mathcal{D}_{O,2}$	$\mathcal{D}_P$
No. of admissions, n (%)	2,414 (3.9)	586 (3.7)	829 (5.2)
Observations Sets, n (%)	13,309 (0.9)	2,983 (0.8)	4,851 (1.1)
16-45 years old, n (%)	1,378 (10.4)	164 (5.5)	187 (3.9)
46-60 years old, n (%)	1,613 (12.1)	374 (12.5)	481 (9.9)
$>$ 60 years old, n (%)	10,318 (77.5)	2,445 (82.0)	4,183 (86.2)

We note that there is a large class imbalance across all datasets, with the positive class representing less than 6% of each dataset. This is important to consider during model development in order to assess for model discrimination and calibration, with the metrics described in Chapter 2.

Chapter 4

Age- and Sex- based Early Warning Score

4.1 Introduction

Many existing EWS systems adopt a 'one size fits all' approach, such that the same alerting thresholds of vital signs are used for all patients, despite age differences for example. They also do account for variations in vital signs due to ageing changes or sex differences. For example, NEWS does not include age or sex in its scoring system [20]. Additionally, the median ages in the development sets for both NEWS and MCEWS were greater than 60 years, meaning that both scores were optimised for use with in an older age group [25, 28]. We hypothesise that the incorporation of individual variations in an EWS system may improve performance over current systems used in clinical practice. In this chapter, we present an age- and sex- specific EWS system developed using a statistical framework.

Age is considered to be a major risk factor of cardiovascular disease [151]. Recent studies show that the pathological and physiological changes that occur with age differences may alter vital signs, specifically for heart rate, blood pressure, temperature,

and respiratory rate [152]. At rest in the seated position, during routine activities of daily living, and during exertion, heart rate decreases with age due to cardiac and vascular ageing [151]. While systolic blood pressure shows a gradual increase between the ages of 30 and 84 years, diastolic blood pressure increases until the fifth decade of age and then slowly decreases to at least 84 years of age [153]. Females have a higher incidence of hypertension with ageing due to menopause and estrogen deficiency [154]. Pre-menopausal females, however, have lower blood pressure than age-matched males due to less stiff arteries [155, 156].

Respiratory rate increases with ageing due to a decline in the lung's ability to recoil after inhalation which subsequently leads to a decrease in tidal volume and an increased work of breathing [157]. Respiratory muscle strength decreases much more in men than in women with ageing [158].

With increased age, thermoregulatory responses and immune systems are also altered, leading to lower core body temperature among older adults. An increase in body temperature is an important reflection of changes in the immune system to combat microbial infections. Thus, early detection of subtle temperature variations in older patients could avoid life-threatening outcomes [152]. As for oxygen saturation, no significant differences were observed amongst different age groups [159].

Most EWS systems do not include age as it increases complexity, conveys little benefit, and raises ethical concerns [25, 160]. The various existing EWS systems fall into three broad categories: (a) objective EWS systems applicable for all patients, regardless of age and sex, such as ViEWS and MCEWS [25, 28]; (b) patient's age and/or sex are considered as parameters of the system and are assigned a score in a heuristic manner [25, 161, 162, 160]; or (c) the EWS is applicable for a specific patient population [163], whereas this thesis considers systems that are applicable to general patient cohorts.

Having identified the complexity of physiology between sexes and with increasing age, we develop and validate a new Age- and Sex- based Early Warning Score (ASEWS) to obtain alerting thresholds for the common vital signs across our clinical baselines: HR, RR, TEMP, SPO_2, SBP, O_2S, and AVPU score. To evaluate the added benefit of incorporating knowledge of the sex of the patient, we also develop an Age-based Early Warning Score (AEWS) for benchmarking purposes.

The rest of the chapter is organised as follows. Section 4.2 describes the methodology to develop the alerting thresholds of ASEWS while Section 4.3 presents how the model was evaluated. Section 4.4 presents the results using the training and testing sets and Section 4.5 discusses the findings.

4.2 Alerting Model

Our ASEWS system was developed in a three-fold process. Fig. 4.1 shows the distribution of the training set $\mathcal{D}_{O,1}$ in terms of age. We note that our dataset includes more observations for elder patients in both (a) and (b). Additionally, $N < 100$ for observations within 24 hours of an outcome for patients younger than 40 years old, as shown in Fig. 4.1 b. As shown in Table 3.2, the abnormal observations for 16-45 years old and 45-60 years old patients constitute 10.4% and 12.1% of the total number of abnormal observations, respectively. The majority of the abnormal observations (77.5%) are collected from > 60 years old patients.

To overcome this uneven distribution of observations, we first created training subsets S_a from the training set $\mathcal{D}_{O,1}$ corresponding to each age a for each sex s (females and males), such that

$$Age(S_a) = \{a \in \mathcal{N} | a - \epsilon \leq a \leq a + \epsilon\} \tag{4.1}$$

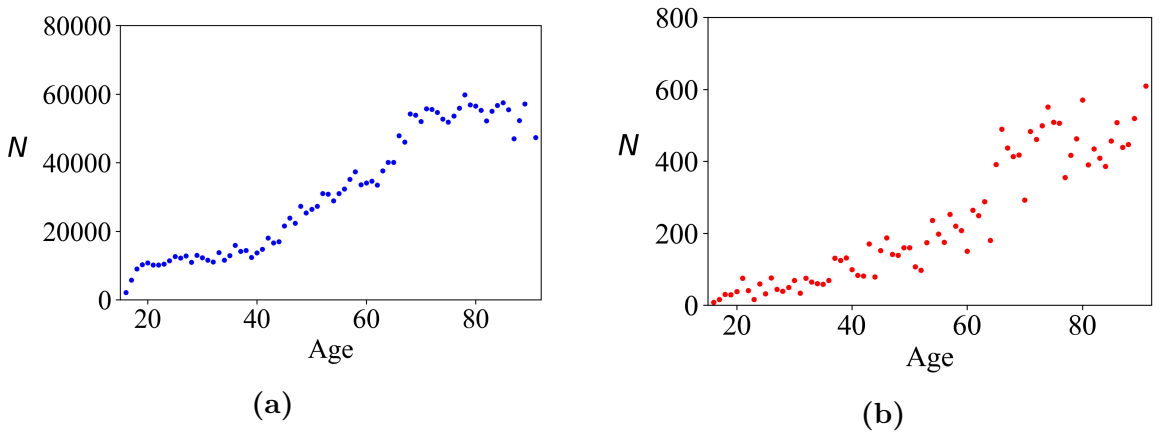
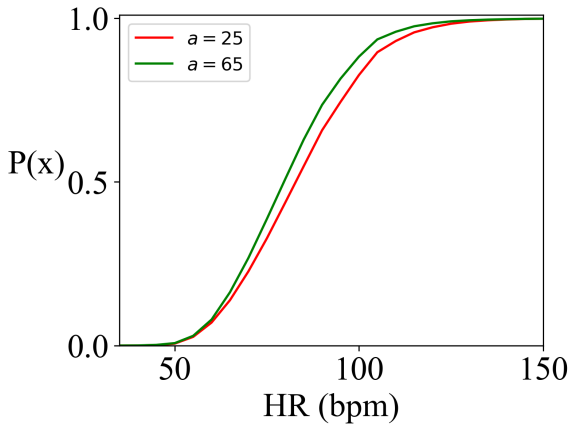


Fig. 4.1: Distribution of vital-sign measurements that are (a) not within 24 hours of a composite outcome and (b) within 24 hours of a composite outcome for each age in the training set $\mathcal{D}_{O,1}$. We note that there are fewer measurements collected for younger patients both in (a) and (b).

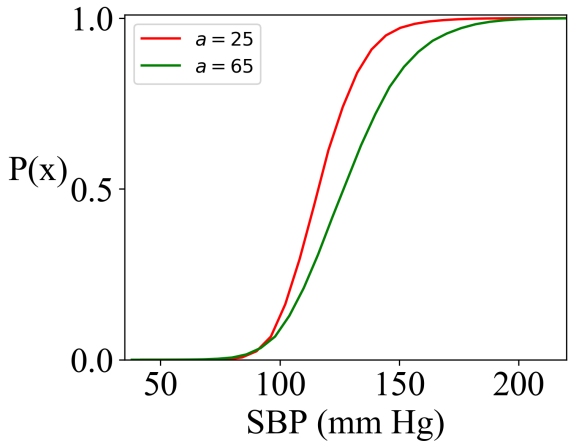
where ϵ is a user-defined number. Thus, $S_{a,s}$ includes observation sets of patients aged between $a-\epsilon$ and $a+\epsilon$ for sex s. For example, when $\epsilon = 5$, the training subset $S_{30,f}$ for 30 years old females includes vital signs of female patients between 25 and 35 years old. We investigated different values of ϵ ranging between 1 to 10 years.

Second, we computed the cumulative distribution function (cdf) $P(x)$ of each continuous vital-sign variable x for each subset $S_{a,s}$. The cdf is the integral of the probability density function $p(x)$ for the random variable x. Fig. 4.2 compares the cdf for HR and SBP for $a = 25$ and $a = 65$ (since the median age in MCEWS and NEWS is above 60) with $\epsilon = 5$. We observe that the distributions are different for the two ages. For example, as shown in Fig.4.2b, the mean SBP is 117 mm Hg $a = 25$ and 127 mm Hg for $a = 65$.

Finally, we used six centiles of the cdf to produce the cut-off thresholds of the warning system at each year of age for each sex. Fig. 4.3 shows the centiles (red lines) of a sample HR cdf. In [27, 28], the alerting thresholds corresponded to fixed centiles, namely 1%, 5%, 10%, 90%, 95%, & 99% for symmetrical distributions and 2%, 10% and 20% for skewed distributions, namely SPO_2 (Fig. A.1 e). The weights assigned



(a)



(b)

Fig. 4.2: Cumulative distribution function P(x) for (a) HR and (b) SBP for two age subsets in the training set ($a = 25$ and $a = 65$) with $\epsilon = 5$.

to measured vital signs that fall within the centile ranges in MCEWS are shown in Appendix A Table B.3. We adopted the same weightings of MCEWS and NEWS (0, 1, 2, and, 3) to avoid increasing the complexity of the system [25].

However, we explored a range of centiles to obtain the alerting thresholds. The centiles varied between 0.1% and 99.9% for symmetrical distributions and between 75% and 99% for skewed distributions. The optimal centiles and ϵ in Eq.4.1 were first selected through a grid-based search approach. We then further optimised the thresholds heuristically based on our existing knowledge to improve the performance

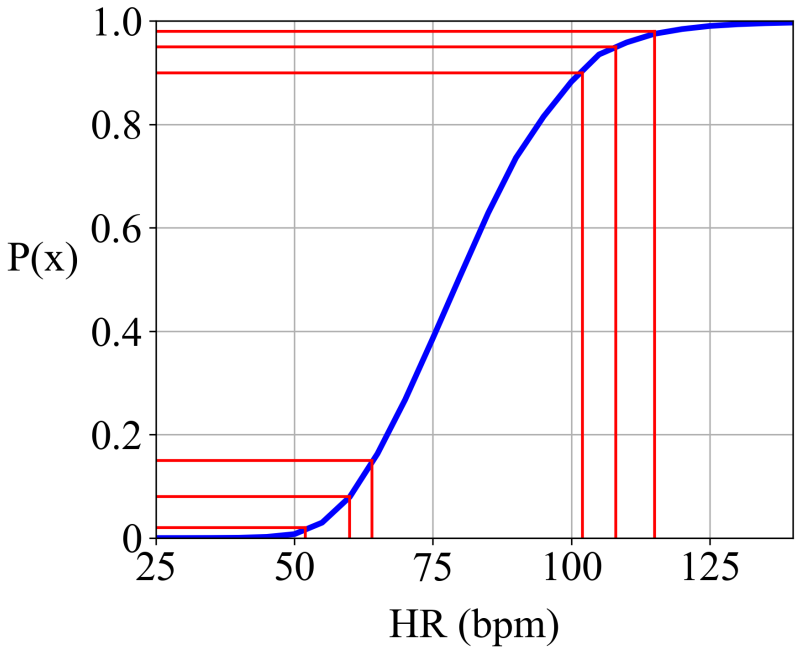


Fig. 4.3: Cumulative distribution function of HR where the centiles used to obtain the alerting thresholds are shown in red.

on the training set. The normal ranges for AVPU and supplemental oxygen were adopted from NEWS, a score of 0 for Alert and a score of 3 otherwise (Voice, Pain or Unresponsive), and a score of 2 when supplemental oxygen was provided [28]. In addition to modelling ASEWS, we excluded sex and modelled an age-specific EWS (AEWS) for benchmarking purposes.

4.3 Model Evaluation

To assess the performance of ASEWS and AEWS developed using the training set $\mathcal{D}_{O,1}$, we evaluated the models using the test sets $\mathcal{D}_{O,2}$ and $\mathcal{D}_P$ and computed performance metrics commonly used in binary classification problems and EWS systems. These metrics were AUROC, sensitivity, specificity, and PPV, as defined earlier in Chapter 2. All metrics were computed through a bootstrapping technique and the binomial 95% confidence interval was computed as described in [164]. We chose the number

of bootstraps (nb) by inspecting the goodness-of-fit of bootstrapped samples using $nb = 50$, 300, and $1,000$.

We also plotted the 'efficiency curve' of each EWS system, which is a qualitative assessment that plots the sensitivity on the horizontal axis and the percentage of observations with a total EWS score greater than or equal to a given total EWS threshold, also known as positives or triggers [165].

We separated the test populations into patients below and above the age of 45 to give sufficient separation from the median age of above 60 in previous score derivations (67.7 and 63 in ViEWS and MCEWS respectively [25, 28]).

4.4 Results

We now summarise the main results in terms of fine-tuning the parameters, the performance of the model, and clinical utility.

4.4.1 Fine-tuned Model Parameters

The optimal value of ϵ that achieved the highest AUROC across all years of age was $\epsilon = 5$ years (the results are not shown here for the sake of brevity). We now present the new alerting thresholds of AEWS and ASEWS.

Age-based Early Warning Score (AEWS)

The alerting thresholds across age are visualised for AEWS, NEWS, and MCEWS in Fig. 4.4. We observe that the alerting thresholds of AEWS change with increasing age. For example, a HR of 110 bpm is assigned a score of 2 for a 16 years old patients and a score of 3 for a 60 years old patient. SBP of 200 mm Hg is assigned a score of 2 for a 76 years old patient but a score of 3 for a 46 years old patient. For TEMP, we assign

a score of 2 to the highest alerting threshold (indicated in orange) as in NEWS. For SPO_2, the only one-sided distribution, we notice that the alerting thresholds of AEWS fall in between the thresholds of NEWS and MCEWS.

We computed the Chi-squared test for a vector of average thresholds across age for AEWS compared to NEWS and MCEWS, respectively. Note that in this case NEWS and MCEWS had the same average alerting threshold across age. However, we found no significant statistical differences using this test between the average alerting threshold across age in AEWS and either of NEWS and MCEWS.

Age- and sex- based Early Warning Score (ASEWS)

The alerting thresholds of ASEWS (females) and ASEWS (males) are shown in Fig. 4.5. We do not observe any significant differences between the HR and TEMP thresholds across females and males. We notice a difference between the RR alerting thresholds across females and males, where younger males (<36 years old) are assigned higher scores for high RR in comparison to their female counterparts. For example, a RR of 25 (breaths/min) of a 30-years old patient is assigned a score of 3 if the patient is a male and a score of 2 if the patient is a female. We also observe that the changes in SBP in females due to menopause is reflected in the alerting thresholds. For example, a SBP of 170 would be assigned a score of 1 for a 56 years-old female but a score of 3 for a male counterpart.

We computed the Chi-squared test for a vector of average thresholds across age for females and males in ASEWS compared to AEWS. We also found no significant statistical differences using this test between the average alerting threshold across age in AEWS and ASEWS.

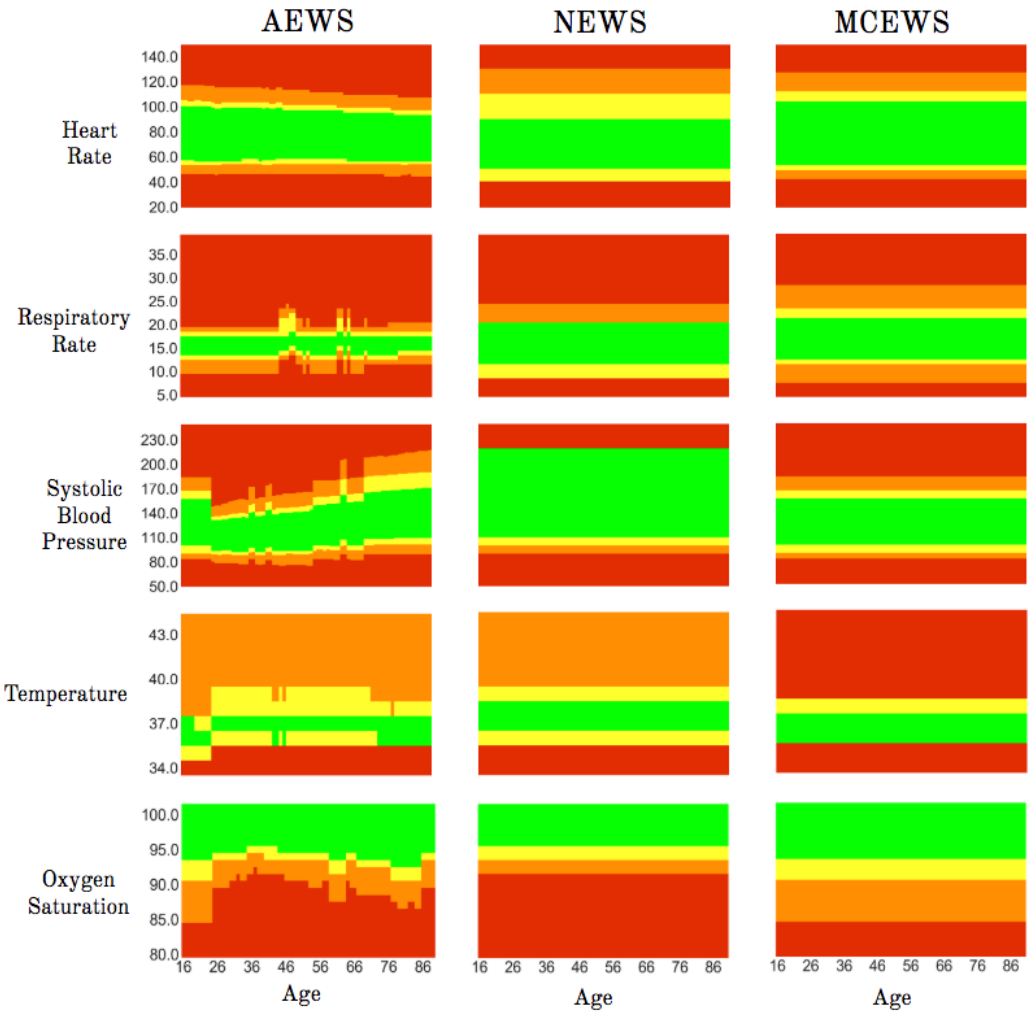


Fig. 4.4: Visualisation of alerting thresholds of AEWS, NEWS, and MCEWS for heart rate, systolic blood pressure, temperature, respiratory rate, and oxygen saturation. Red indicates an alerting score of 3, orange a score of 2, yellow a score of 1, and green a score of 0.

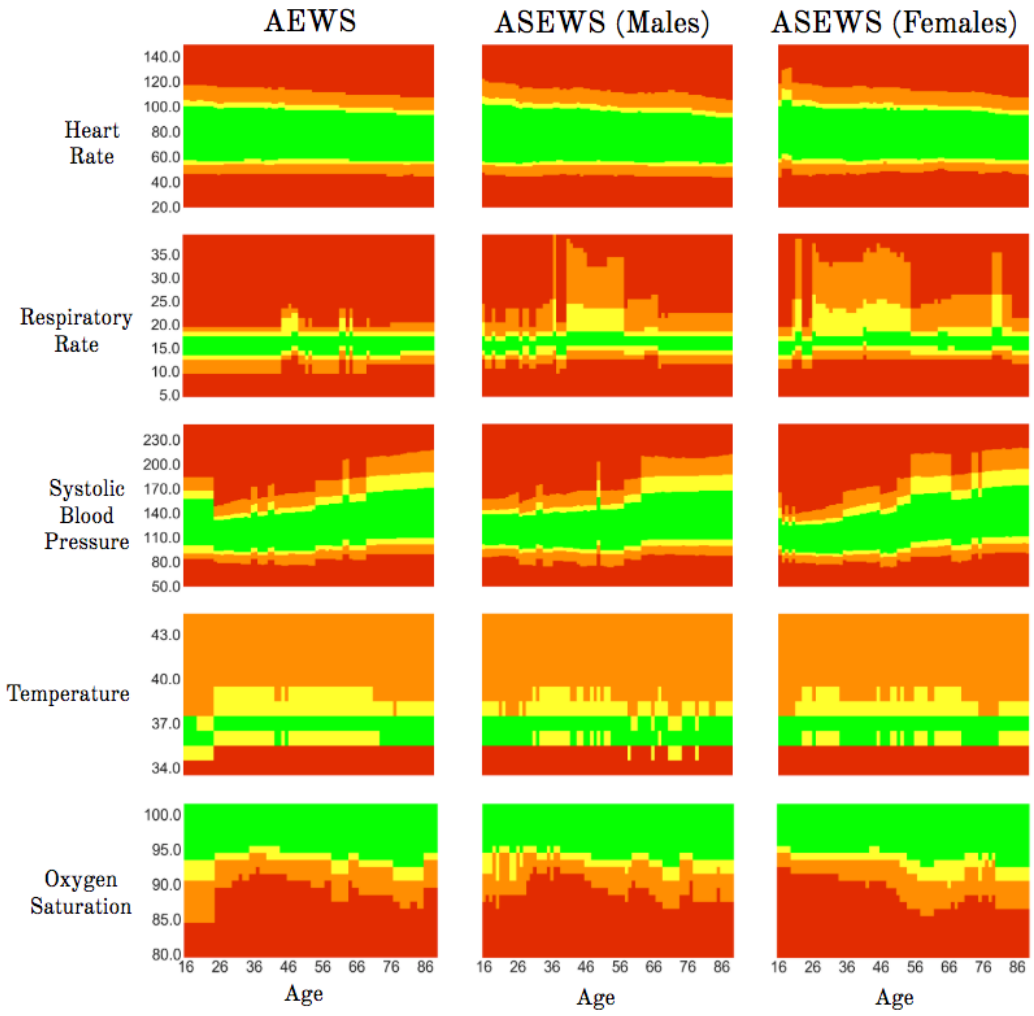


Fig. 4.5: Visualisation of alerting thresholds for AEWS and ASEWS across females and males, for heart rate, systolic blood pressure, temperature, respiratory rate, and oxygen saturation. Red indicates an alerting score of 3, orange a score of 2, yellow a score of 1, and green a score of 0.

4.4.2 Performance Assessment

The bootstrapping histograms for measuring AUROC and AUPRC in the $\mathcal{D}_{O,1}$ dataset when using NEWS are shown in Fig. 4.6. We observe that the most normal distribution is obtained with $nb = 1,000$, when compared to $nb = 50$ and $nb = 300$. We therefore consecutively perform all bootstrapping computations with $nb = 1,000$.

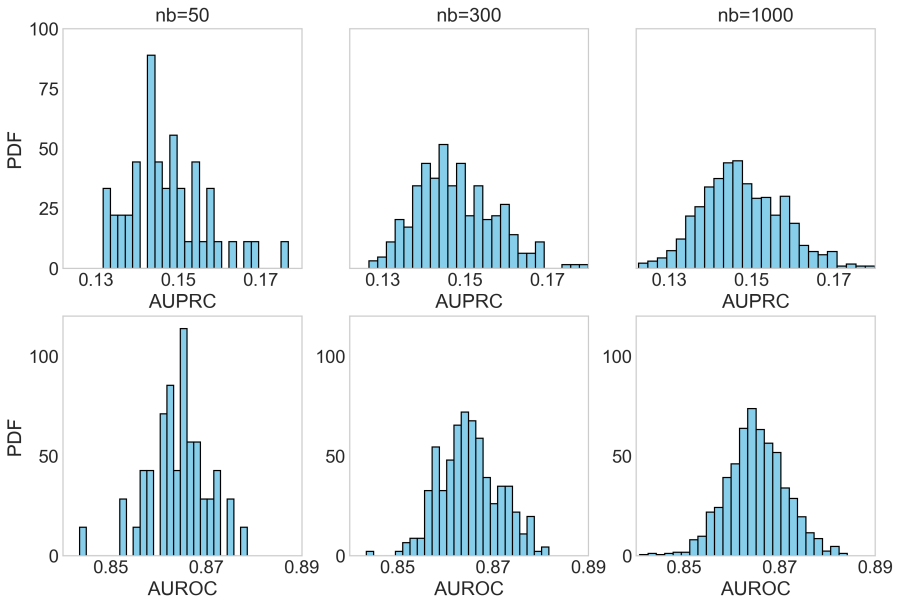


Fig. 4.6: To choose the number of bootstraps nb during evaluation, we compared the distributions of bootstrapped samples. The first row shows the bootstrap samples of AUPRC and the second row of AUROC, when applying NEWS on the $\mathcal{D}_{O,1}$ dataset.

The performance of the EWS systems in predicting the composite outcome within 24 hours of an observation set measurement using bootstrapping in $\mathcal{D}_{O,2}$ is summarised in Table 4.1. When evaluated for all patients in $\mathcal{D}_{O,2}$, AEWS achieved 87.3% AUROC [95% CI 87.3-87.4], which was higher than the AUROC of all other scores. In the younger age group (16-45 years old), AEWS significantly outperformed all other EWS systems, especially NEWS (83.2 vs. 75.8). Across the elder age group (>45 years old), AEWS and ASEWS had the highest performance and they performed similarly.

The results obtained when analysing $\mathcal{D}_P$, an independent and external testing set, are shown in Table 4.2. AEWS and NEWS achieved the highest performance in comparison to ASEWS and MCEWS. Amongst the younger population (16-45 years old), AEWS has a superior performance in comparison to all other EWS 89.0% AUROC [95% CI 88.8-89.1]. In the elder population (>45 years old), AEWS and NEWS performed similarly.

Overall, AEWS achieved the highest discriminatory performance across the 16-45 years old. Both existing EWS (MCEWS and NEWS) perform worse on the younger population than the elderly population. AEWS maintains a consistent performance across both datasets.

Table 4.1: Performance of ASEWS and AEWS in $\mathcal{D}_{O,2}$, in comparison to existing EWS systems, namely NEWS and MCEWS, in terms of the AUROC, Sensitivity, Specificity, and PPV. The alerting threshold of NEWS is adjusted at 5, while the alerting thresholds of other scores are adjusted to be at least as specific as NEWS. Mean and confidence intervals were evaluated using bootstrapping ($nb = 1,000$).

EWS	AUROC	Sensitivity	Specificity	PPV
Overall				
AEWS	**87.3 (87.3-87.4)**	69.9 (69.8-70.0)	88.7 (88.7-88.7)	4.9 (4.9-4.9)
ASEWS	**87.3 (87.3-87.3)**	67.4 (67.3-67.5)	90.2 (90.2-90.3)	**5.5 (5.5-5.5)**
NEWS	86.5 (86.5-86.6)	**70.1 (70.1-70.2)**	88.8 (88.8-88.8)	5.0 (5.0-5.0)
MCEWS	86.1 (86.1-86.1)	66.3 (66.2-66.4)	**90.4 (90.4-90.4)**	5.5 (5.5-5.5)
16-45 years old				
AEWS	**83.2 (83.0-83.3)**	51.9 (51.6-52.3)	92.6 (92.6-92.6)	2.3 (2.3-2.3)
ASEWS	82.1 (81.9-82.2)	42.9 (42.5-43.3)	93.1 (93.1-93.1)	2.1 (2.1-2.1)
NEWS	75.7 (75.4-75.9)	51.7 (51.3-52.1)	93.4 (93.4-93.5)	**3.0 (3.0-3.1)**
MCEWS	82.4 (82.2-82.6)	**52.7 (52.4-53.1)**	**94.3 (94.3-94.3)**	2.6 (2.6-2.6)
> 45 years old				
AEWS	**87.4 (87.3-87.4)**	70.9 (70.8-71.0)	88.1 (88.1-88.1)	5.2 (5.1-5.2)
ASEWS	**87.4 (87.4-87.5)**	68.7 (68.7-68.8)	89.8 (89.8-89.8)	**5.8 (5.8-5.8)**
NEWS	86.9 (86.9-86.9)	67.0 (66.9-67.0)	88.0 (88.0-88.0)	5.2(5.2-5.2)
MCEWS	86.1 (86.0-86.1)	**71.2 (71.1-71.3)**	**89.8 (89.8-89.8)**	5.7 (5.7-5.7)

Table 4.2: Performance of ASEWS and AEWS in $\mathcal{D}_P$, in comparison to existing EWS systems, namely NEWS and MCEWS, in terms of the AUROC, Sensitivity, Specificity, and PPV. The alerting threshold of NEWS is adjusted at 5, while the alerting thresholds of other scores are adjusted to be at least as specific as NEWS. Mean and confidence intervals were evaluated using bootstrapping ($nb = 1,000$).

EWS	AUROC	Sensitivity	Specificity	PPV
Overall				
AEWS	**86.9 (86.8-86.9)**	63.1 (63.1-63.2)	91.1 (91.1-91.1)	7.1 (7.1-7.1)
ASEWS	86.5 (86.5-86.6)	60.1 (60.0-60.2)	92.1 (92.1-92.1)	**7.6 (7.6-7.6)**
NEWS	86.8 (86.8-86.9)	59.6 (59.5-59.6)	**92.2 (92.2-92.2)**	7.4 (7.4-7.5)
MCEWS	85.2 (85.2-85.2)	**63.3 (63.2-63.4)**	91.5 (91.5-91.5)	**7.6 (7.6-7.7)**
16-45 years old				
AEWS	**89.0 (88.8-89.1)**	**66.5 (66.2-66.8)**	93.8 (93.8-93.8)	5.0 (5.0-5.0)
ASEWS	86.9 (86.8-87.1)	61.6 (61.3-62.0)	94.3 (94.3-94.3)	5.1 (5.0-5.1)
NEWS	87.3 (87.1-87.4)	60.6 (60.2-60.9)	**95.5 (95.5-95.5)**	**6.3 (6.2-6.3)**
MCEWS	86.7 (86.6-86.9)	61.6 (61.3-62.0)	95.0 (95.0-95.0)	5.6 (5.6-5.7)
> 45 years old				
AEWS	86.6 (86.6-86.6)	63.0 (63.0-63.1)	90.8 (90.8-90.8)	7.2 (7.2-7.2)
ASEWS	86.4 (86.4-86.5)	60.1 (60.0-60.1)	**91.9 (91.9-91.9)**	**7.8 (7.8-7.8)**
NEWS	**86.7 (86.7-86.7)**	**63.4 (63.3-63.4)**	91.1 (91.1-91.1)	7.5 (7.5-7.5)
MCEWS	85.0 (85.0-85.0)	59.5 (59.4-59.6)	92.0 (92.0-92.0)	7.7 (7.7-7.7)

4.4.3 Clinical Utility

We compared the trigger rates for achieving a sensitivity rate of 80%; i.e., to correctly identify 80% of observations that are within 24 hours of unplanned ICU admission, cardiac arrest or mortality, as shown in Fig. 4.7. When observing (i) all patients and (ii) the elderly patients (>45 years old), AEWS and NEWS had similar trigger rates in $\mathcal{D}_{O,2}$ and $\mathcal{D}_P$. For example, for younger patients (16-45 years old), the trigger rate of AEWS was 28% in comparison to a higher trigger rate of $\approx$ 60% by NEWS in $\mathcal{D}_{O,2}$. In $\mathcal{D}_P$, AEWS decreases the trigger rate by around $\approx$3% to achieve a sensitivity of 80%.

We consider two case studies to evaluate scores produced by the best performing models, AEWS and NEWS. Fig. 4.8a shows a 34-years-old male, admitted due to alcoholic cirrhosis of the liver (within dataset $\mathcal{D}_P$) who experienced unplanned ICU admission as indicated with the vertical green line. Since NEWS alerts for a total score

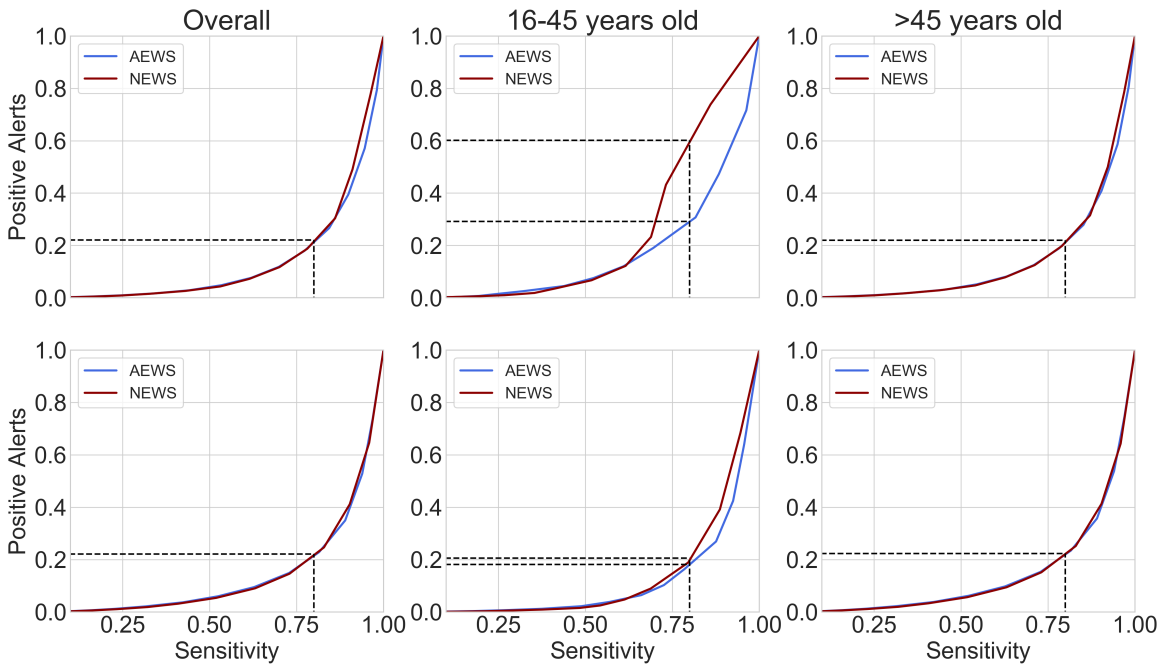


Fig. 4.7: EWS efficiency curves for AEWS (red) and NEWS (blue) when applied to $\mathcal{D}_{O,2}$ (first row) and $\mathcal{D}_P$ (second row). The black vertical dashed line indicates a sensitivity of 80%.

$\geq$5 (dashed horizontal line), we note that AEWS begins to produce alerts at least 28 hours in advance of the outcome, whereas NEWS begins to alert at around 8 hours prior to the outcome.

Fig. 4.8b shows a 23-years-old male, admitted with pneumonitis (within dataset $\mathcal{D}_P$) who is eventually discharged without experiencing an adverse outcome. We note that both AEWS and NEWS produce scores lower than 5 during the stay of the patient.

4.5 Discussion

Based on the performance across two independent testing sets, $\mathcal{D}_{O,2}$ and $\mathcal{D}_P$, we note that incorporating sex has no clear added benefit, as ASEWS only performs similar or worse than AEWS. This suggests that the proposed technique may not be capturing information in the vital signs pertaining to sex, if such information exists. However,

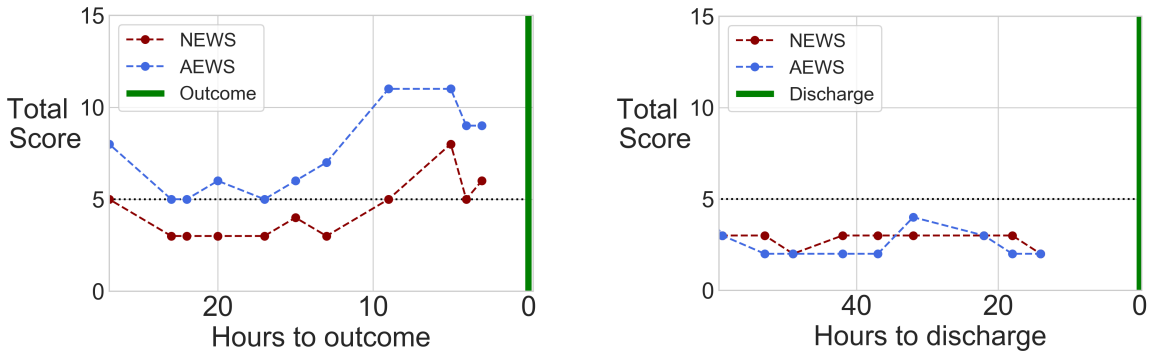


Fig. 4.8: (a) Case study of a 34 year old male who experiences unplanned ICU admission (green line). The horizontal dashed line at $y = 5$ is the alerting threshold of NEWS. (b) Case study of a 23 year old male who is discharged (green line) without any adverse outcomes. The horizontal dashed line at $y = 5$ is the alerting threshold of NEWS.

the new alerting thresholds of AEWS improved the discrimination performance in younger patients in the two independent testing sets, one being an external validation set from a completely different centre to that of $\mathcal{D}_2$. Since AEWS performs similar to NEWS in the elderly population (>45 years old) in both testing sets, NEWS could remain as a better tool for use with such groups due to its simplicity.

We also show that, to achieve a sensitivity of 80% using AEWS (which is deemed to be an acceptable rate by our clinicians) the medical staff would need to respond to only half of the number of triggers as would be generated by NEWS amongst younger patients, improving the efficiency of ward care. AEWS also produces slightly fewer alerts than NEWS in the $\mathcal{D}_P$ validation set, which demonstrates its robustness in performance and clinical utility for previously unseen data. Our findings may be used to maximise performance in more patient-specific, computer-calculated scores.

Our study has several limitations. We assume that the vital-sign observation sets of each patient are independent and identically distributed (i.i.d.) random sets and therefore it is assumed that there is no correlation between vital signs. This may not be the case in reality since variations of a vital sign across a time window

may be indicative of deterioration. However, the assumption of i.i.d. is the common methodology in studies of this type [25, 28], due to the clinical assumption that the extent of derangement is indicative of deterioration.

Additionally, the methodology seems to overfit to the training set. For example, we observe sudden changes in the alerting thresholds of AEWS and ASEWS due to overfitting to the training set $\mathcal{D}_{O,1}$, since the alerting thresholds were chosen using a grid-based search approach of centiles that maximise the performance of the EWS on $\mathcal{D}_{O,1}$. Overfitting is further emphasised in Fig. 4.7, since AEWS does leads to a very small decrease in the positive alert rate in $\mathcal{D}_P$ compared to a larger decrease in $\mathcal{D}_{O,2}$.

In general, the development and evaluation of EWS systems is also challenged by the low prevalence of adverse outcomes, which is typically even lower for younger patients than for elderly patients. Our solution to limited data is to group patients in age subsets. The severe class imbalance leads to low PPV values, which is a common limitation of existing EWS systems across various patient cohorts [28, 114]. Nevertheless, NEWS is still endorsed by the Royal College of Physicians and deployed in various clinical settings to assess acutely ill adult patients across the UK.

Chapter 5

Deep Interpretable Early Warning System

5.1 Introduction

Although traditional statistical EWS systems are easy to use and quick to compute on paper, their construction relies on extensive manual engineering based on clinical knowledge and heuristics. A pragmatic yet inflexible approach to their construction is to evaluate existing EWS systems and then develop one customised for a specific patient cohort, leading to a multiplicity of EWS systems. Additionally, EWS systems assign an independent score to each vital sign and assume that vital-sign measurements are i.i.d. random variables. Given this simplistic nature, traditional EWS systems do not learn any temporal information from the vital signs. We hypothesise that the use of deep learning may improve the accuracy of predicting clinical outcomes by recognising complex patterns in the data. In this chapter, we present a more flexible and scalable approach by using deep learning techniques.

Significant improvements over clinical scores and machine learning models were achieved using deep learning, such as to predict ICU mortality for paediatrics [35] or to

detect sepsis [54]. LSTM networks in particular have illustrated superior performance when considering various benchmarks [75, 74, 166]. Most of the relevant work, however, was primarily based in intensive care settings. We designed a flexible EWS framework that could generalise across a heterogeneous patient population in non-critical care wards, from pre-processing sparse vital-sign variables to predicting the probability of an outcome.

Additionally, the decision-making process of the previously-proposed deep learning models lacked interpretability, and as such they are viewed as 'black box' models by clinical staff since they do not provide any insight on the patterns learned from the data. We defined interpretability as the ability of the clinician to inspect 'trends' of vital signs that most contribute to the model's predicted probability. Inspired by natural language processing, our approach incorporated an 'attention' mechanism with recurrent architectures that highlights those parts of the input time-series that are most relevant to the output. This is useful since interpretability is considered to be a core component of clinical utility [110].

The physiological data recorded in an EHR is often sparse, noisy, and incomplete, especially when collected in non-critical care wards, which is challenging for recurrent deep learning architectures that require regularly-spaced data points. Regularly-sampled data can be interpolated using naïve methods, such as carrying the most recent value forward (CF) and linear interpolation (LI). Although such approaches are computationally inexpensive, they may impose bias and error [167] and do not account for the uncertainty of the imputed data. In a probabilistic approach, Gaussian Process Regression (GPR) was used to model irregularly-sampled physiological time-series data [44, 43, 49], to interpolate the posterior mean and variance at unseen time points. In our work described in this chapter, we evaluated the benefit of GPR modelling in comparison to CF and LI.

We therefore develop and validate an interpretable end-to-end Deep Early Warning System (DEWS) that alerts for clinical deterioration, by extracting information from sparse variables and computing the probability of an outcome. From a clinical perspective, we aim that the model identifies physiological trends indicative of clinical deterioration. Another main priority of the proposed model is to achieve generalisable state-of-the-art performance across unseen data.

The rest of the chapter is organised as follows. We begin by summarising the state-of-the-art related to detecting clinical deterioration during a patient admission using deep learning and other machine learning techniques in Section 5.2. Section 5.3 describes the methodology pipeline in terms of feature extraction and outcome prediction, Section 5.4 presents the experimental observations using the proposed models and datasets, and Section 5.5 discusses findings.

5.2 State-of-the-art

Attention-based deep learning models, which will be defined below, have been shown to improve interpretability; they can model extended long-term-dependencies; and they have been used numerously in computer vision and natural language processing [168, 169]. In traditional sequence-to-sequence modelling problems, attention enables the model to learn deferentially from more and less important parts of the input sequence; i.e., words in the case of sentence translation, or sentences for document classification [170]. Our goal was to learn different content from different time-series signals, and then fuse the information to predict the probability of an outcome. To the best of our knowledge, no existing attention-based deep learning model focuses on learning from a combination of vital-sign time-series data to indicate a patient's health status in real-time.

Within clinical settings, attention models have gained limited recognition and have been used for classifying atrial fibrillation in ECG data [73, 171], predicting a future diagnosis [106, 89], detecting deterioration preceding mortality using vital signs and laboratory tests [56], or predicting high risk vascular disease using diagnosis codes and medication data [113]. The limitation of using diagnosis codes is that they may not be readily available in a real-time setting, as in retrospective databases. We aimed to leverage information in routinely collected vital-sign data, as in currently implemented EWS systems, constituting multiple sequential inputs.

The primary contribution of this chapter is a novel deep learning architecture with high clinical utility and interpretability to predict clinical outcomes. The model, as will be described in detail, learns from regularly-sampled mean and variance features interpolated by modelling sparse vital-sign data GPR. We also evaluated the framework's ability in detecting deterioration prior to the composite outcome, as in previous studies [28, 25].

5.3 DEWS: an end-to-end learning framework

We framed the problem of detecting clinical deterioration as a binary classification task, such that the model assigns a binary label, based on a computed probability and a pre-defined alerting threshold, to each vital-sign measurement. For each vital-sign measurement, we would like to predict the probability of the composite outcome within the next N hours. An event window was defined as a set of vital-sign measurements within N hours of a composite outcome and its preceding w hours of observations. A non-event window was defined as a vital-sign measurement that was not within N hours of a composite outcome and its preceding w hours window. We set N=24 hours in our study, which is a common evaluation window in the development of EWS systems [28, 25], and we evaluated w at 24, 12, and 6 hours.

Algorithm 1 Proposed framework of DEWS that predicts whether a window $\mathcal{D}_W$ of vital-sign measurements is within N hours of an outcome.

Input: Unlabelled window $\mathcal{D}_W = [x_i, \mathbf{y}_i]_{i=1}^{n}$
and regularly sampled time instances $P = [x_t^*]_{t=1}^{T}$ per window.
Outputs: Label $l \in [0, 1]$.

Data Interpolation

For $j = 1$ to m:
    $GPR_j \leftarrow$ Fit a GPR$_j$ using $[x_i, y_i]_{i=1}^{n}$
    $\mathbf{y}_{\mu,j}$, $\mathbf{y}_{\sigma,j} \leftarrow GPR([x_t^*]_{t=1}^{T})$
$\mathbf{Y}_\mu = [\mathbf{y}_{\mu,1}, \ldots, \mathbf{y}_{\mu,m}]$
$\mathbf{Y}_\sigma = [\mathbf{y}_{\sigma,1}, \ldots, \mathbf{y}_{\sigma,m}]$

Classification

$\mathbf{c}_\mu \leftarrow Encoder(\mathbf{Y}_\mu)$
$\mathbf{c}_\sigma \leftarrow Encoder(\mathbf{Y}_\sigma)$
$l \leftarrow Decoder(\mathbf{c}_\mu + \mathbf{c}_\sigma)$

Assume $\mathcal{D}_W = [x_i, \mathbf{y}_i]_{i=1}^{n}$ is an unlabelled sample window, where x_i is the i^{th} time instance and $\mathbf{y}_i \in \mathbb{R}^m$ denotes the feature space consisting of m vital-sign sequences. The recurrent classification model required the vital-sign data to be measured at a fixed set of regularly-sampled time instances to compute the output escalation label $l \in (0, 1)$. However, each vital-sign sequence j was temporally irregular due to the nature of EHR data. Hence, we deployed a patient-specific GPR for each vital-sign sequence to interpolate the mean and variance at fixed regularly-spaced time instances $P = [x_t^*]_{t=1}^{T}$.

These posterior mean and variance estimates were concatenated for all the vital signs to obtain: $\mathbf{Y}_\mu = [\mathbf{y}_{\mu,j}]_{j=1}^{m}$ and $\mathbf{Y}_\sigma = [\mathbf{y}_{\sigma,j}]_{j=1}^{m}$, where $\mathbf{Y}_\mu, \mathbf{Y}_\sigma \in \mathbb{R}^{m \times T}$ and $\mathbf{y}_{\mu,j}$ and $\mathbf{y}_{\sigma,j}$ are the GPR mean and variance for the j^{th} vital sign, such that $j = 1, \ldots, m$. Attention-based *encoders* learned from each interpolated sequence in $\mathbf{Y}_\mu$ and $\mathbf{Y}_\sigma$ to obtain the sequence-level context vectors $[\mathbf{c}_{\mu,j}]_{j=1}^{m}$ and $[\mathbf{c}_{\sigma,j}]_{j=1}^{m}$, respectively. Finally, the summary context vectors $\mathbf{c}_\mu$ and $\mathbf{c}_\sigma$ summed up the sequence-level context vectors and were fed to decoding layers to compute the probability of an outcome. If the

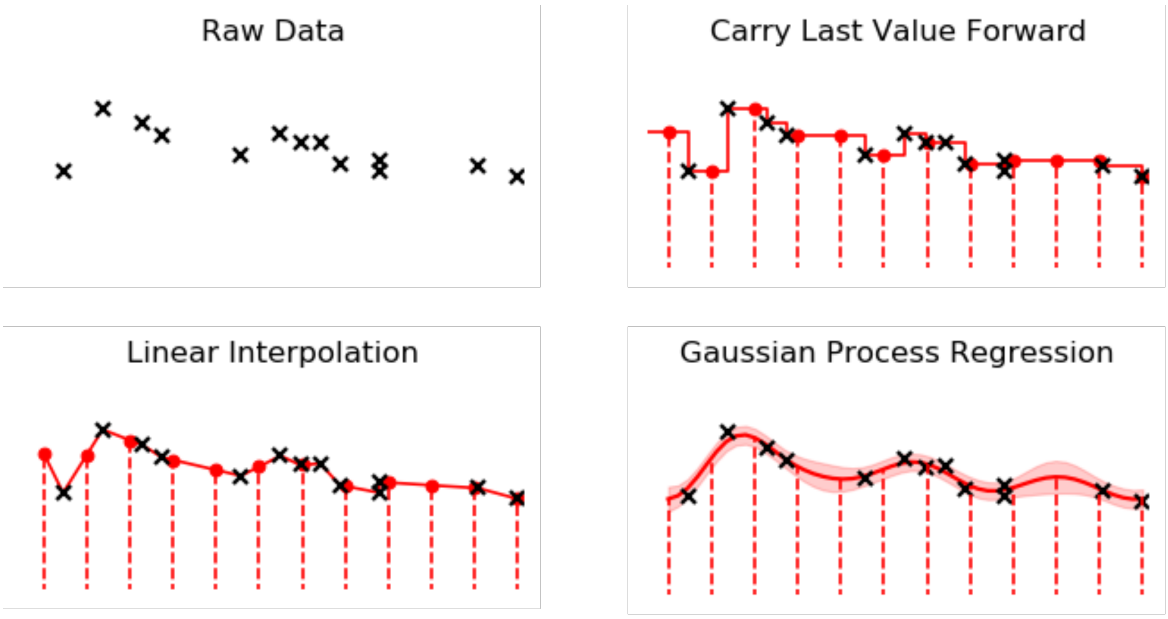


Fig. 5.1: Visualisation of the feature extraction interpolation techniques from the raw data within a w hour window: Carrying Last Value Forward, Linear Interpolation, and Gaussian Process Regression. The vertical axis represents the range of any vital sign and the horizontal axis represents time since admission, with the rightmost data point representing the time of prediction.

probability exceeded the pre-defined alerting threshold, then $\mathcal{D}_W = 1$. The proposed framework is described in Algorithm 1. We now describe each step in more detail.

5.3.1 Patient-specific Feature Transformation

Each window of length w hours was modelled using GPR. It was also modelled using CF and LI for benchmarking purposes, and an overview of the modelling techniques is shown in Fig. 5.1. GPR generalises multivariate Gaussian distributions to infinite dimensionality and offers a probabilistic and non-parametric approach to model a sparse vital-sign time-series $[x_i, y_i]_{i=1}^{n}$ as a function of time from admission, such that

$$\mathbf{y} \sim \mathcal{GP}(m(x), k(x, x')) \tag{5.1}$$

where $m(x)$ is the mean function and $k(x,x')$ is the covariance function that describes the relationship between the observed values. We adopted a radial basis function (RBF) with added white noise as our covariance function to map the similarity between pairs of data points x and x', such that

$$k(x,x') = \sigma_f^2 \exp\left(-\frac{(x-x')^2}{2l^2}\right) + \sigma_n^2 \delta(x,x') \tag{5.2}$$

$\delta(x,x')$ is the Kronecker delta function and $\Theta = \{l, \sigma_f, \sigma_w\}$ is the set of hyperparameters, where l is the lengthscale, σ_f is the variance of the RBF, and σ_w is the variance of the added white noise. Since the work involved patient-specific modelling for each sequence per window across a large-scale population, we adopted a Bayesian approach to increase the modelling efficiency. First, we defined the expected value, or the mean function, of each vital-sign GPR as a constant function equivalent to the population mean of patients with the same age and sex. Second, we used lognormal distributions as priors to constrain each hyperparameter to be clinically meaningful, where the starting points of the hyperparameters are sampled from pre-specified distributions rather than being randomly selected. The models were optimised by minimising the negative log marginal likelihood (LML) with respect to the hyperparameters Θ.

$$\log[p(\mathbf{y}|\mathbf{x},\Theta)] = -\frac{1}{2}(\mathbf{y}-\mathbf{m})^T K^{-1}(\mathbf{y}-\mathbf{m}) - \frac{1}{2}\log|K| - \frac{n}{2}\log(2\pi) \tag{5.3}$$

where K is the covariance function. After fitting the GPR to the training data $[x_i, y_i]_{i=1}^n$ of vital sign j, the GPR kernel is applied to interpolate missing values at equally-spaced time steps $[x_t^*]_{t=1}^T$ across the input window, such that the posterior mean is

$$\mathbf{y}_\mu = K_* K^{-1} \mathbf{y}, \tag{5.4}$$

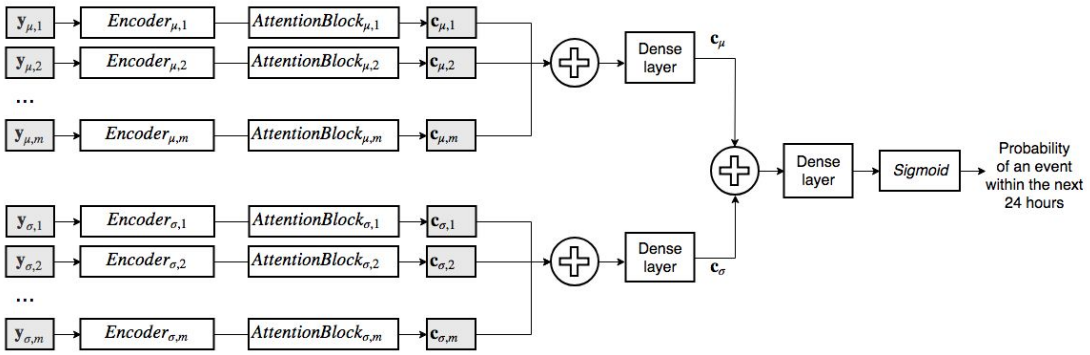


Fig. 5.2: Schematic diagram of the classification architecture DEWS which learns the attention weights from the mean and variance of each vital-sign variable and produces a binary output to indicate whether an observation set of a patient is within 24 hours of a composite outcome (cardiac arrest, mortality or unplanned ICU admission). Each regularly sampled input $\mathbf{y}$ $(T \times 1)$ is processed by an *encoder* unit consisting of a BiLSTM with e output units at each time step to produce the hidden states $(T \times e)$. The attention blocks then compute the context vectors $(e \times 1)$ that summarise their respective inputs. The context vectors of the mean and variance of the vital signs are finally aggregated using dense layers.

with variance,

$$\mathbf{y}_{\sigma} = K_{**} - K_{*}K^{-1}K_{*}^{T}, \tag{5.5}$$

where $\mathbf{y}$ is the training data $[y_i]_{i=1}^{n}$, K represents the similarity measure between all training values, K_* represents the similarity measure between all training and missing values, and K_{**} represents the similarity measure between all missing values. Finally, we obtained a set of equally spaced measurements as an input to the neural network.

For any point of prediction with historical data spanning less than w hours, we pre-padded the sequence with the population mean and maximum variance of the respective vital sign. Discrete variables were modelled by CF and LI, which only interpolated mean values. The extracted features, through GPR, CF, and LI, were then scaled for the training and testing of the classifiers.

5.3.2 Classifier Architecture

We here describe the architecture of the proposed DEWS method. The interpolated mean and variance features of each vital-sign input were first processed through a Bi-directional LSTM (BiLSTM) network [172], in order to maximise information retrieval in the forward and backward directions. An attention-based BiLSTM model previously performed well for classifying sequential healthcare data [73], and we extended upon it by customising the mechanism in the attention block and accounting for the uncertainty of the input.

The BiLSTM consisted of two layers which processed each mean and variance input in forward and reverse directions and yielded two hidden layer states $h_{t,f}$ and $h_{t,r}$. The average of $h_{t,f}$ and $h_{t,r}$, denoted as $\overline{h_t}$ served as the input of our attention mechanism. While the definition of attention varies across the literature, we adopted the definition in [170, 73], to learn the most important parts of each sequence. For each vital sign j, $e_{t,j}$ measured the importance of the information at each time step t:

$$e_{t,j} = U_j a(W_j \overline{h_{t,j}} + b_j), \tag{5.6}$$

where U_j is a trainable sequence-level context vector and a is the rectified linear unit. Next, $e_{t,j}$ was used to derive α_j, or the normalised weights assigned to the hidden states, as:

$$\alpha_j = \frac{\exp(e_{t,j})}{\sum_{t=1}^{T} \exp(e_{t,j})} \tag{5.7}$$

$\alpha_{t,j}$ was further employed to derive the final context vector $\mathbf{c}_j$:

$$\mathbf{c}_j = \sum_{t=1}^{T} \alpha_{t,j} \overline{h_{t,j}} \tag{5.8}$$

Equations 5.6-5.8 were applied to the BiLSTM outputs of each mean and variance input of each vital sign j. The sequence-level context vectors obtained from all vital signs were then aggregated by the trainable weights V_μ for the mean features:

$$\mathbf{c}_\mu = V_\mu \sum_j \mathbf{c}_{\mu,j} \tag{5.9}$$

and V_σ for the variance features:

$$\mathbf{c}_\sigma = V_\sigma \sum_j \mathbf{c}_{\sigma,j} \tag{5.10}$$

The aggregated context vectors were then summed and processed by two dense layers consisting of a rectified linear unit and a sigmoid function, respectively. Finally, the computed probability of an outcome within the next 24 hours was compared to a pre-defined alerting threshold and a binary label was assigned to the window. The model schematic is shown in Fig. 5.2.

During training, the weights U_k, W_k, b_k, V_μ and V_σ were optimised, where V_μ and V_σ accounted for the correlations across the vital signs since they combined their respective context vectors. We chose the other hyperparameters, such as number of output nodes e per encoder timestep shown in Fig. 5.2, and the alerting threshold through experimentation. The motivation for the final architecture is based on the performance of simpler model variants described in Section 5.4.3.

5.4 Empirical Study

In this section, we summarise the main findings of the experimental choices and evaluation results. We used $\mathcal{D}_{O,1}$ for training and validation, $\mathcal{D}_{O,2}$ for testing, and $\mathcal{D}_P$ as an external testing set. Due to the severe class imbalance as shown in Table 3.2,

Table 5.1: Characteristics and demographics of the new balanced training set $\mathcal{D}_{O,1B}$ used for training and validation.

Characteristic	$\mathcal{D}_{O,1B}$
Age, mean (SD)	68 (18)
Admissions, n	21,512
Females, n (%)	10.703 (49.8)
16-45 yrs old, n (%)	3,462 (16.1)
> 45 yrs old, n (%)	18,050 (83.9)
Vital-sign measurements, n	45,314
Non-event, n (%)	22,657 (50)
Event, n (%)	22,657 (50)
Unplanned ICU	10,942 (48.3)
Cardiac Arrest	1,314 (5.8)
Mortality	10,401 (45.9)

the deep learning models were initially unable to learn from the data. We therefore performed under-sampling of the non-event windows in $\mathcal{D}_{O,1}$ to match the maximum number of event windows, obtaining a balanced training set denoted as $\mathcal{D}_{O,1B}$. The main limitation of under-sampling is the possible loss of important data samples. The sampling was performed in a random manner, to avoid selection bias. Oversampling of the minority class, an alternative sampling strategy, often leads to overfitting to the training data. We chose random under-sampling to avoid overfitting and since it is the simplest solution compared to random oversampling, informed under-sampling, and data interpolation. Our results are in alignment with previous findings such that balanced training datasets improve learning and predictive performance [173]. $\mathcal{D}_{O,2}$ and $\mathcal{D}_P$ remained imbalanced to mimic real-life testing sets.

5.4.1 Evaluation Metrics

We assessed the modelling quality of GPR, CF, and LI, using $\mathcal{D}_{O,1B}$. For each vital sign per window, we randomly held out 20% of the data as test points and modelled the rest using the three interpolation techniques. We then calculated the root mean

squared error (RMSE) to compare the true values and the interpolated values at the held out test points.

We evaluated the performance of DEWS using the AUROC, sensitivity, specificity, and positive predictive value on the testing set. All metrics were performed using a bootstrapping technique without replacement [164] with a fixed number of bootstraps (nb). We compared the performance of the models across 16-45 years old patients and >45 years old patients, and across the three outcomes independently.

To assess the clinical utility of DEWS in comparison to the clinical benchmark, we plotted the percentage of generated 'triggers', or windows at or above a given EWS score, on the vertical axis against sensitivity on the horizontal axis [25].

We also assessed the proposed model's decision-making process by visualising the attention weights computed for a case study. Finally, we compared the average normalised NEWS score and the average DEWS probability for the first 120 hours from admission and the last 24 hours prior to an outcome.

5.4.2 Data Modelling

The demographics of the new balanced training set are shown in Table 5.1 and the distribution of its vital signs are shown in Table 5.2. The new balanced training set has a similar mean age and proportion of females as the testing sets. However, since it was balanced while the testing sets were imbalanced, we observe differences between the distributions of some variables.

Lognormal priors over the hyperparameters were selected to ensure that the modelled vital signs fell within the expected ranges in a clinical setting. The lognormal distributions chosen as priors for l were $(\mu = 1.0, \sigma = 0.1)$ for HR, RR, TEMP, and SPO_2, and $(\mu = 1.5, \sigma = 0.1)$ for SBP and DBP. The lognormal distributions chosen as priors for σ_f were $(\mu = 0.0, \sigma = 0.1)$ for HR, SBP, DBP, and SPO_2, $(\mu = 1.5, \sigma = 0.1)$

Table 5.2: Distributions of the vital signs in terms of median and interquartile range in the new balanced training set $\mathcal{D}_{O,1B}$.

Vital Sign, unit	$\mathcal{D}_{O,1B}$
HR, beats/min	86 (73-101)
SBP, mm Hg	121 (105-139)
TEMP, °C	36.4 (36.0-36.9)
RR, breaths/min	18 (16-21)
SPO_2, %	96 (94-98)
AVPU Score	
Alert, n (%)	40,642 (89.6)
Voice, n (%)	2884 (6.4)
Pain, n (%)	981 (2.2)
Unresponsive, n (%)	807 (1.8)
Supplemental Oxygen	18,845 (41.6)

for RR, and $(\mu = 3.5, \sigma = 0.1)$ for TEMP. The lognormal distributions chosen as priors for σ_w were $(\mu = 0.0, \sigma = 4.0)$ for HR, SBP, DBP, and SPO_2, $(\mu = 0.0, \sigma = 0.1)$ for RR, and $(\mu = 1.5, \sigma = 0.1)$ for TEMP. Applying population-based lognormal distributions to the priors of the three hyperparameters enabled us to efficiently fit patient-specific vital-sign GPR models in a large-scale dataset. All GPR models were built using GPy (v 1.9.6) [174]. Sample windows modelled using GPR are shown in Fig. 5.3.

Using the optimised GPR models, we interpolated the posterior mean and variance at every 2 hours across the w hours long window, in keeping with national guidelines of alerting at least at every other hour. We modelled a truncated window of up to w hours to reduce the number of timesteps per input for the recurrent neural network. This reduced the complexity of the architecture, which is essential given our dataset size. Additionally, modelling a large number of timesteps would have required the storage of a subsequently large GPR kernel matrix for each window, which would have imposed computational complexity. The model performed best with windows of length 24 hours i.e. $w = 24$ hours, after evaluating its performance for lengths of 6, 12, and 24 hours.

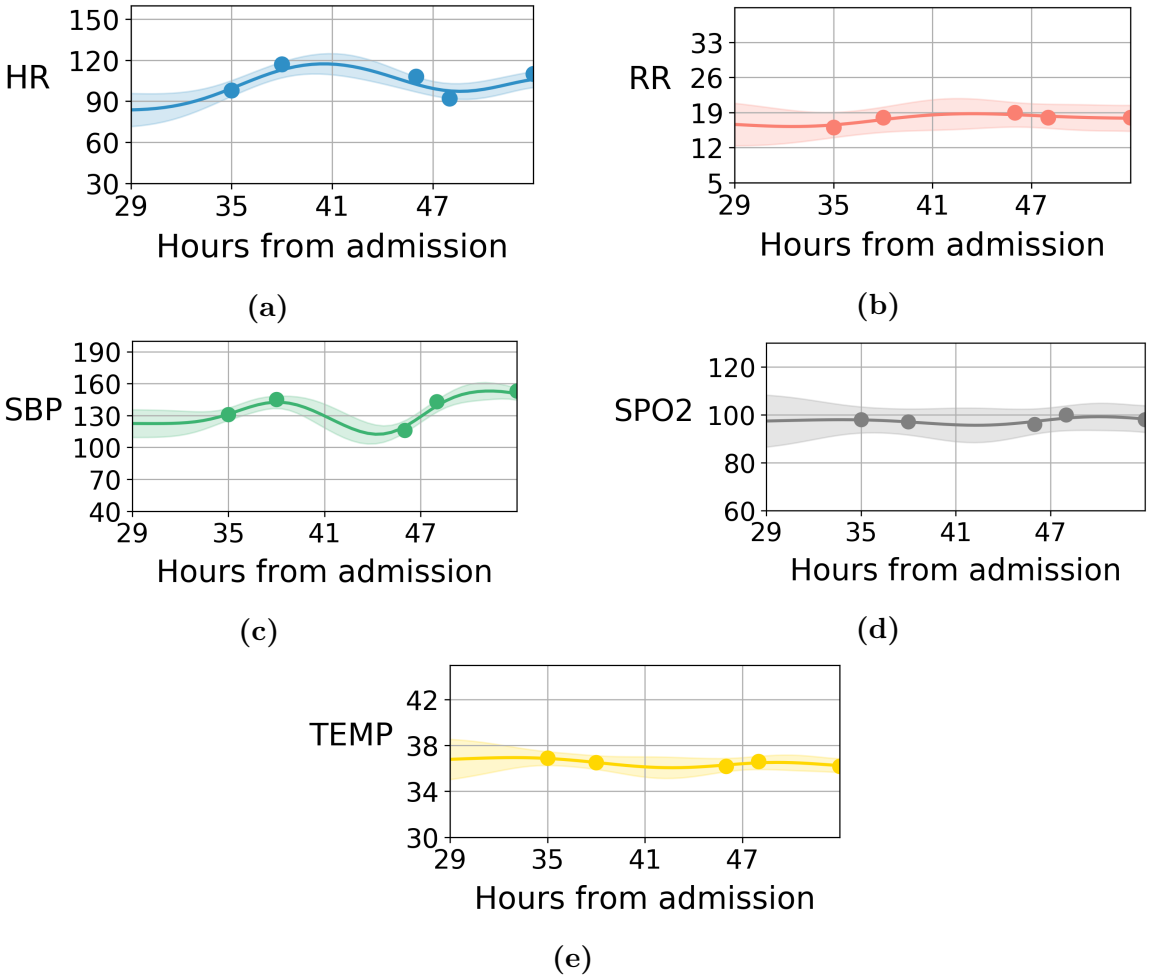


Fig. 5.3: Sample physiological sequences modelled using GPR: (a) HR, (b) RR, (c) SBP, (d) SPO_2, and (e) TEMP.

After sampling equally-spaced measurements, we experimented with standard scaling, min-max scaling and scaling by the maximum absolute value. The best classification performance was achieved through min-max scaling of mean features into the range [-1,1] and maximum absolute scaling of variance features into the range [0,1]. During training, we used 20% of $\mathcal{D}_{O,1B}$ as a validation set, and so the scaling and shifting operations were obtained through the other 80% and then applied to the validation set and the two test set sets $\mathcal{D}_{O,2}$ and $\mathcal{D}_P$.

The mean and standard deviation of the RMSE of the training windows are summarised in Table 5.3 to compare the data interpolation quality of GPR, CF, and LI. In DEWS, AVPU and supplemental oxygen were interpolated using CF as it is less computationally expensive than LI, considering that they both resulted with similar mean RMSE.

Table 5.3: Mean and standard deviation of the Root Mean Squared Error (RMSE) for missing data interpolation using the different modelling techniques for all vital signs: CF, LI, and GPR. This was done by modelling 80% of each training window and holding out 20% of the training window as test points to calculate the RMSE, using $\mathcal{D}_{O,1B}$.

Vital sign	CF	LI	GPR
HR	8.24 ±8.08	7.20 ±7.15	7.49 ±6.96
RR	2.24 ±2.63	2.07 ±2.36	1.97 ±2.15
TEMP	0.41 ±0.37	0.35 ±0.31	0.36 ±0.31
SBP	13.72 ±11.58	12.15 ±10.46	12.79 ±10.21
SPO_2	2.16 ±2.81	2.03 ±2.70	1.97 ±2.46
AVPU	0.10 ±0.33	0.10 ±0.32	-
Supplemental Oxygen	0.12 ±0.29	0.12 ±0.26	-

5.4.3 Model Variants as Baselines

We designed several deep learning architectures to compare to DEWS. The first set of models consisted of simple architectures, namely Logistic Regression (LR), a single-layer LSTM network, and a single-layer BiLSTM network. $\mathbf{Y}_{\mu}$ was the input to the models' first single layer. Despite its simplicity, the BiLSTM architecture lacked interpretability.

Therefore, the second set of models included attention mechanisms applied to $\mathbf{Y}_{\mu}$. BiLSTM-ATT-1 consisted of a simple BiLSTM followed by one attention (ATT) module. This approach was similar to language modelling, since it consisted of a single input feature space, however we were unable to identify the individual contributions of each vital sign. BiLSTM-ATT-2 thus processed each vital sign independently using a

dedicated BiLSTM and attention mechanism. The context vectors of the vital signs were then summed up and decoded.

The third set of models consisted of 'Uncertainty-Aware' (UA) models since attention was not only applied to the mean features, but also to the variance features $\mathbf{Y}_\sigma$. UA-BiLSTM-ATT-1 consisted of two BiLSTM layers, where one BiLSTM processed $\mathbf{Y}_\mu$ and the other processed $\mathbf{Y}_\sigma$. Each BiLSTM was then followed by one ATT module. Finally, UA-BiLSTM-ATT-2, our proposed model DEWS, had one BiLSTM-ATT per mean and variance features of the vital signs as shown in Fig. 5.2. We compared all models to NEWS and a simple logistic regression (LR) which used i.i.d. features as inputs; i.e., the last recorded set of vital-signs measurement.

5.4.4 Deep Learning Experiments

We tried training the model using data from emergency admissions only, as in $\mathcal{D}_{O,2}$ and $\mathcal{D}_P$, yet the model performed best when the exclusion criteria was not applied. Despite differences in characteristics of vital signs in $\mathcal{D}_{O,1B}$ and the testing sets, the model was able to learn complex patterns since we utilised patient-specific modelling.

The hyperparameters of the models, including the number of hidden layers, units per layers, and activation functions were optimised empirically using the training and validation set $\mathcal{D}_{O,1B}$. Within the encoder units, the optimal number of output nodes (e) at each timestep of the BiLSTM that resulted with the best performance was 12. The dense layers that aggregated all context vectors consisted of 5 units, while the final dense layer consisted of 1 unit. For the similarity function of the attention block, we compared the hyperbolic tangent function and ReLU, and the latter performed better for our application. The final complexities of the models, in terms of total number of trainable parameters, are shown in Table C.2.

All deep learning architectures were trained with early stopping by monitoring the loss on the validation set to avoid overfitting, and a batch size of 128, after experimenting with a batch size of 32, 64, and 128. Each batch consisted of sequences of the same length, whereby length of sequence refers to the number of sampled equidistant data points prior to padding. The models were optimised using the Adam optimiser. All deep learning models were implemented using Keras (v 2.2.2) [175] with a TensorFlow backend (v 1.5.0) [176].

5.4.5 Results

Performance Evaluation

Table 5.4 shows the performance results of all classifiers on $\mathcal{D}_{O,2}$. DEWS performed best compared to all other models with an 88.0 AUROC [95% CI 88.0-88.0] and 72.9 sensitivity [95% CI 72.8-0.72.9], exceeding NEWS, LR, and all model variants. The decision threshold of DEWS that achieved a similar specificity to NEWS was 0.66. All LR-based models performed worse than DEWS and NEWS, while all other model variants, with the exception of LI-LSTM, performed similar to NEWS. When only the mean features of the GPR were considered in the BiLSTM-variant models (i.e. BiLSTM, BiLSTM-ATT-1, and BiLSTM-ATT-2), no improvement was observed when compared to other interpolation methods (CF and LI). However, when the variance of the GPR was considered, it further improved the results of GPR-based methods in comparison to other interpolation methods, such as in UA-BiLSTM-ATT-1 and DEWS.

Across different patient sub-populations in $\mathcal{D}_{O,2}$, DEWS performed better than NEWS for both age groups, as shown in Table 5.5, especially for 16-45 years old patients, with 82.0 AUROC [95% CI 81.8-82.2] compared to 76.0 AUROC [95% CI

75.7-76.2], respectively. DEWS also performed better than NEWS across all outcomes, both in terms of AUROC and sensitivity.

In $\mathcal{D}_P$, DEWS had a superior performance over NEWS for the composite outcome, with 88.0 AUROC [95% CI 88.0-88.0] and 86.8 AUROC [95% CI 86.8-86.9] respectively, as shown in Table 5.6. Across all sub-populations, DEWS performed better than NEWS in terms of AUROC. This highlights the generalisability of DEWS on an unseen dataset.

Table 5.4: Performance evaluation of DEWS in comparison to NEWS and all model variants, using i.i.d., CF, LI, or GPR interpolated features in $\mathcal{D}_{O,2}$. The decision threshold of all classifiers was adjusted to achieve a specificity similar to that of NEWS (≈89.0). Mean and confidence intervals were evaluated using a bootstrapping technique ($nb = 1{,}000$).

Model	AUROC	Sensitivity	Specificity	PPV
i.i.d.				
NEWS	86.6 (86.5-86.6)	70.2 (70.2-70.3)	88.8 (88.8-88.8)	5.0 (5.0-5.0)
LR	84.7 (84.6-84.7)	65.1 (65.0-65.2)	88.8 (88.8-88.8)	4.6 (4.6-4.7)
CF				
LR	83.4 (83.4-83.5)	64.1 (64.0-64.2)	88.8 (88.8-88.8)	4.7 (4.6-4.7)
LSTM	86.7 (86.6-86.7)	69.1 (69.0-69.1)	88.7 (88.7-88.7)	4.9 (4.8-4.9)
BiLSTM	86.7 (86.7-86.7)	69.9 (69.9-70.0)	88.8 (88.7-88.8)	5.0 (4.9-5.0)
BiLSTM-ATT-1	86.8 (86.8-86.8)	71.5 (71.4-71.6)	88.7 (88.7-88.7)	5.0 (5.0-5.0)
BiLSTM-ATT-2	86.9 (86.8-86.9)	70.2 (70.1-70.3)	88.8 (88.8-88.8)	5.0 (5.0-5.0)
LI				
LR	84.5 (84.5-84.6)	67.0 (66.9-67.0)	88.7 (88.7-88.7)	4.7 (4.7-4.8)
LSTM	85.7 (85.7-85.8)	66.7 (66.7-66.8)	88.7 (88.7-88.7)	4.7 (4.7-4.7)
BiLSTM	86.0 (86.0-86.0)	69.0 (68.9-69.1)	88.8 (88.8-88.8)	4.9 (4.9-4.9)
BiLSTM-ATT-1	86.8 (86.8-86.8)	71.2 (71.1-71.3)	88.7 (88.7-88.7)	5.0 (5.0-5.0)
BiLSTM-ATT-2	86.6 (86.5-86.6)	70.7 (70.6-70.7)	88.8 (88.7-88.8)	5.0 (5.0-5.0)
GPR				
LR	84.5 (84.4-84.5)	66.7 (66.6-66.8)	88.6 (88.6-88.6)	4.7 (4.7-4.7)
LSTM	86.6 (86.6-86.6)	69.6 (69.6-69.7)	88.7 (88.7-88.7)	4.9 (4.9-4.9)
BiLSTM	87.2 (87.2-87.2)	70.1 (70.0-70.2)	88.7 (88.7-88.7)	5.0 (4.9-5.0)
BiLSTM-ATT-1	86.6 (86.6-86.7)	71.5 (71.5-71.6)	**88.9 (88.9-88.9)**	5.1 (5.1-5.1)
BiLSTM-ATT-2	86.5 (86.5-86.6)	69.3 (69.2-69.4)	88.8 (88.8-88.8)	4.2 (4.2-4.2)
UA-BiLSTM-ATT-1	87.2 (87.2-87.3)	71.7 (71.6-71.8)	88.7 (88.7-88.7)	5.1 (5.1-5.1)
DEWS	**88.0 (88.0-88.0)**	**72.9 (72.8-72.9)**	88.7 (88.7-88.7)	**5.1 (5.1-5.1)**

Table 5.5: Performance evaluation of DEWS in comparison to NEWS across sub-populations of interest; i.e., 16-45 years old, >45 years old, and each of the three events in the composite outcome, in $\mathcal{D}_{O,2}$. The recommended decision threshold for NEWS is 5 and the adjusted decision threshold for DEWS was 0.66, to achieve a similar overall specificity (≈89.0). Mean and confidence intervals were evaluated using a bootstrapping technique ($nb = 1,000$) for the respective sub-population.

Model	AUROC	Sensitivity	Specificity	PPV
16-45 years old				
DEWS	**82.0 (81.8-82.2)**	**53.6 (53.2-53.9)**	92.5 (92.5-92.5)	2.4 (2.4-2.4)
NEWS	76.0 (75.7-76.2)	52.2 (51.9-52.6)	**93.5 (93.4-93.5)**	**2.6 (2.6-2.7)**
> 45 years old				
DEWS	**88.1 (88.1-88.2)**	**74.0 (73.9-74.1)**	**88.1 (88.1-88.1)**	**5.4 (5.4-5.4)**
NEWS	86.9 (86.9-87.0)	71.3 (71.2-71.4)	88.0 (88.0-88.0)	5.2 (5.2-5.2)
Unplanned ICU				
DEWS	**81.1 (81.1-81.2)**	**55.5 (55.4-55.7)**	**90.0 (90.0-90.0)**	**1.4 (1.4-1.4)**
NEWS	77.2 (77.1-77.3)	52.8 (52.7-53.0)	90.0 (90.0-90.0)	1.2 (1.2-1.2)
Cardiac Arrest				
DEWS	**76.7 (76.5-76.9)**	**45.7 (45.4-46.1)**	**90.0 (90.0-90.0)**	**0.3 (0.3-0.3)**
NEWS	75.2 (75.1-75.4)	42.6 (42.3-43.0)	90.0 (90.0-90.0)	0.3 (0.3-0.3)
Mortality				
DEWS	**92.6 (92.6-92.7)**	**83.1 (83.1-83.2)**	**88.8 (88.8-88.8)**	**4.6 (4.6-4.6)**
NEWS	92.2 (92.2-92.2)	80.5 (80.5-80.6)	88.8 (88.8-88.8)	3.8 (3.8-3.8)

Clinical Utility

Fig. 5.4 shows the percentage of triggers, or positive alerts, produced by our best performing model, DEWS, and NEWS at different sensitivity values (horizontal axis) in the two testing sets. In $\mathcal{D}_{O,2}$, across the 16-45 years old patients, Fig. 5.4 (a), NEWS approximately had a 59% trigger rate while DEWS had a 37% trigger rate, at a fixed sensitivity of 80%. This shows that DEWS reduced the trigger rate by approximately 22%, which could directly ease staff burden. Across the >45 years old, DEWS reduced the trigger rate by approximately 3%. In $\mathcal{D}_P$, DEWS had a 15% lower trigger rate than NEWS in the 16-45 years old patients, and an 8% lower trigger rate than NEWS in the elder age group.

Table 5.6: Performance evaluation of DEWS in comparison to NEWS across sub-populations of interest; i.e., 16-45 years old, >45 years old, and each of the three events in the composite outcome, in $\mathcal{D}_P$. The recommended decision threshold for NEWS is 5 and the adjusted decision threshold for DEWS was 0.725, to achieve a similar overall specificity (≈91.5). Mean and confidence intervals were evaluated using a bootstrapping technique ($nb = 1,000$) for the respective sub-population.

EWS	AUROC	Sensitivity	Specificity	PPV
Overall				
DEWS	**88.0 (88.0-88.0)**	**65.4 (65.3-65.4)**	**91.6 (91.6-91.6)**	7.4 (7.4-7.5)
NEWS	86.8 (86.8-86.9)	63.4 (63.3-63.4)	91.5 (91.5-91.5)	**7.7 (7.7-7.7)**
16-45 years old				
DEWS	**90.4 (90.2-90.5)**	**64.8 (64.5-65.1)**	95.4 (95.3-95.4)	**6.4 (6.4-6.5)**
NEWS	87.2 (87.1-87.4)	61.6 (61.3-62.0)	**95.5 (95.5-95.5)**	6.3 (6.2-6.3)
> 45 years old				
DEWS	**87.7 (87.7-87.8)**	**65.4 (65.3-65.5)**	**91.2 (91.2-91.2)**	**7.8 (7.8-7.8)**
NEWS	86.7 (86.7-86.7)	63.4 (63.3-63.5)	91.1 (91.1-91.1)	7.5 (7.5-7.5)
Unplanned ICU				
DEWS	**84.6 (84.6-84.7)**	**52.7 (52.6-52.8)**	**92.9 (92.9-92.9)**	**2.5 (2.5-2.5)**
NEWS	83.5 (83.5-83.6)	53.0 (52.9-53.1)	92.7 (92.7-92.7)	2.6(2.6-2.6)
Cardiac Arrest				
DEWS	**80.6 (80.5-80.7)**	**45.5 (45.4-45.7)**	**92.9 (92.8-92.9)**	**1.2 (1.2-1.2)**
NEWS	79.6 (79.5-79.6)	45.2 (45.1-45.4)	92.7 (92.6-92.7)	1.2 (1.1-1.2)
Mortality				
DEWS	**91.6 (91.6-91.6)**	**73.3 (73.3-73.4)**	**91.7 (91.6-91.7)**	**6.2 (6.2-6.3)**
NEWS	90.2 (90.2-90.3)	70.8 (70.7-70.9)	91.6 (91.6-91.6)	6.0 (6.0-6.0)

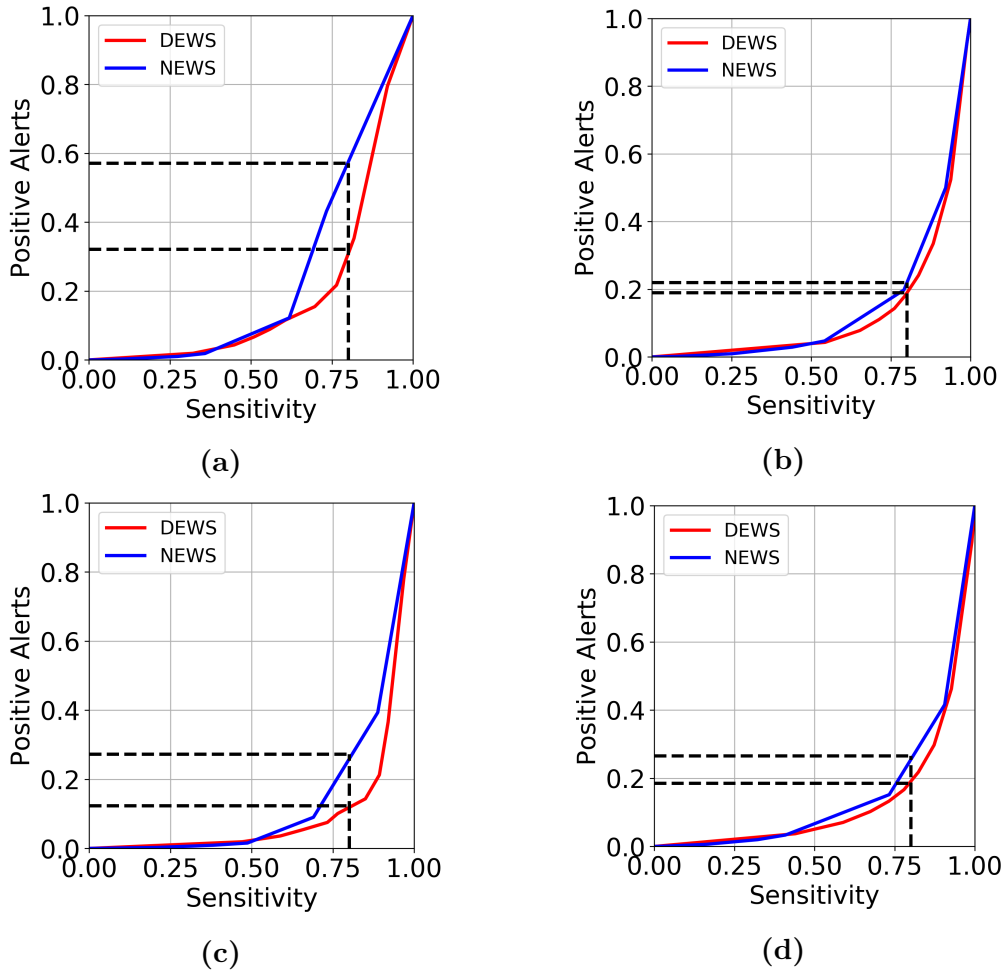


Fig. 5.4: Efficiency curves plotting sensitivity on the horizontal axis against the trigger rate, or the percentage of observations on the vertical axis, with a DEWS probability or normalised NEWS score greater than or equal to a decision threshold, with the decision threshold ranging between 0 and 1, for (a) 16-45 years old patients in $\mathcal{D}_{O,2}$, (b) >45 years old patients in $\mathcal{D}_{O,2}$, (c) 16-45 years old patients in $\mathcal{D}_P$, and (d) >45 years old patients in $\mathcal{D}_P$. The vertical black dashed line represents a fixed sensitivity of 80%.

Case Studies

The attention weights of two windows are visualised in Fig. 5.5, where the box with blue borders shows the vitals signs after feature transformation through GPR modelling and scaling. The two windows belonged to the same patient, where the first row was a non-event window, since the time of prediction was not within 24 hours of an outcome,

while the second row was an event window since the time of prediction was within 24 hours of an outcome. In the attention weights of the non-event window, we observe that time steps 6-10 gained more importance than other time steps for SBP. When compared to the raw data, we notice that this trend corresponds to an increase followed by a decrease in SBP across the respective time steps. All other uniform distributions indicate that the model equally values each time step. The probability of an event produced by DEWS for this window was 28%, compared to a score of 6 by NEWS. This window was thus classified as a true negative by DEWS and a false positive by NEWS.

As for the second row, the event window, RR, SBP, and TEMP varied similarly, with a decreasing attention from left to right. In the original data, RR and TEMP sharply increased in the earlier time steps. SBP, on the other hand, decreased from a high value. In this scenario, DEWS produced a probability of 98%, while NEWS produces a score of 15, and as such both models produced a true positive.

In Fig. 5.6, the mean probability produced by DEWS model and the normalised NEWS score appear to be aligned in terms of overall trends. We notice in (a) and (c) that in the first 24 hours window from admission, where the thick dashed black line represents 24 hours from admission, the probability produced by DEWS decreases, which suggests that DEWS gains confidence as the patient's length of stay increases. In (b), the last 24 hours window prior to an outcome in $\mathcal{D}_{O,2}$, DEWS maintained a mean probability greater than its alerting threshold of 0.66, while NEWS maintained a mean normalised score at around 0.3. In (d), the last 24 hours prior to an outcome in $\mathcal{D}_P$, DEWS maintained a mean probability greater than its alerting threshold of 0.725, while NEWS maintained a mean normalised score at around 0.3.

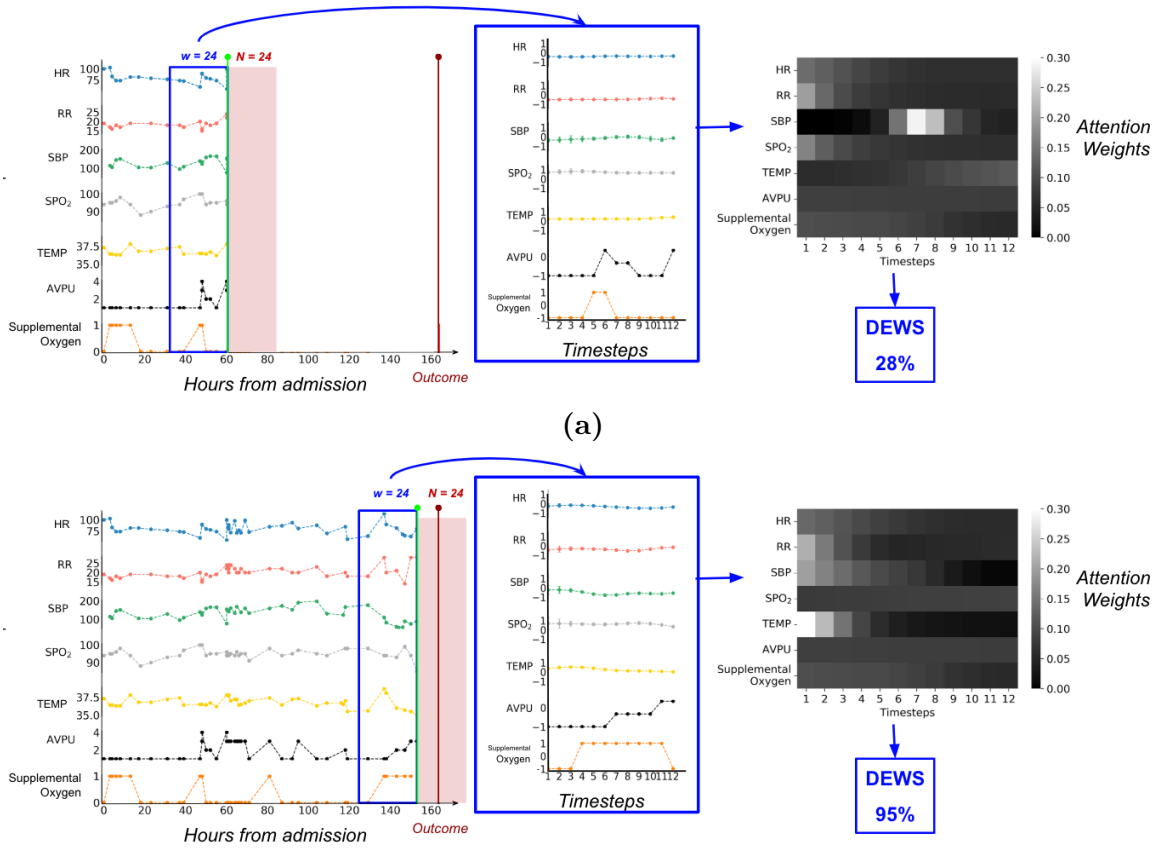


Fig. 5.5: Visualisation of the attention weights learned from the mean and variance of all vital-sign sequences using DEWS, where (a) is a non-event window and (b) is an event window, sampled from the same patient. The leftmost figure shows the raw data, where the green vertical line marks the time at which a prediction is made and the red vertical line marks the time of an outcome. The middle blue bordered figure shows the interpolated features within the window after GPR modelling and scaling.

When comparing the models' AUROC against time to outcome, we observe that DEWS performed better than NEWS across the 24 hours window prior to outcome, as shown in Fig. 5.7.

5.5 Discussion

We propose an attention-based neural network that learns from historical trends of vital signs through interpolated mean and variance features to alert for clinical deterioration.

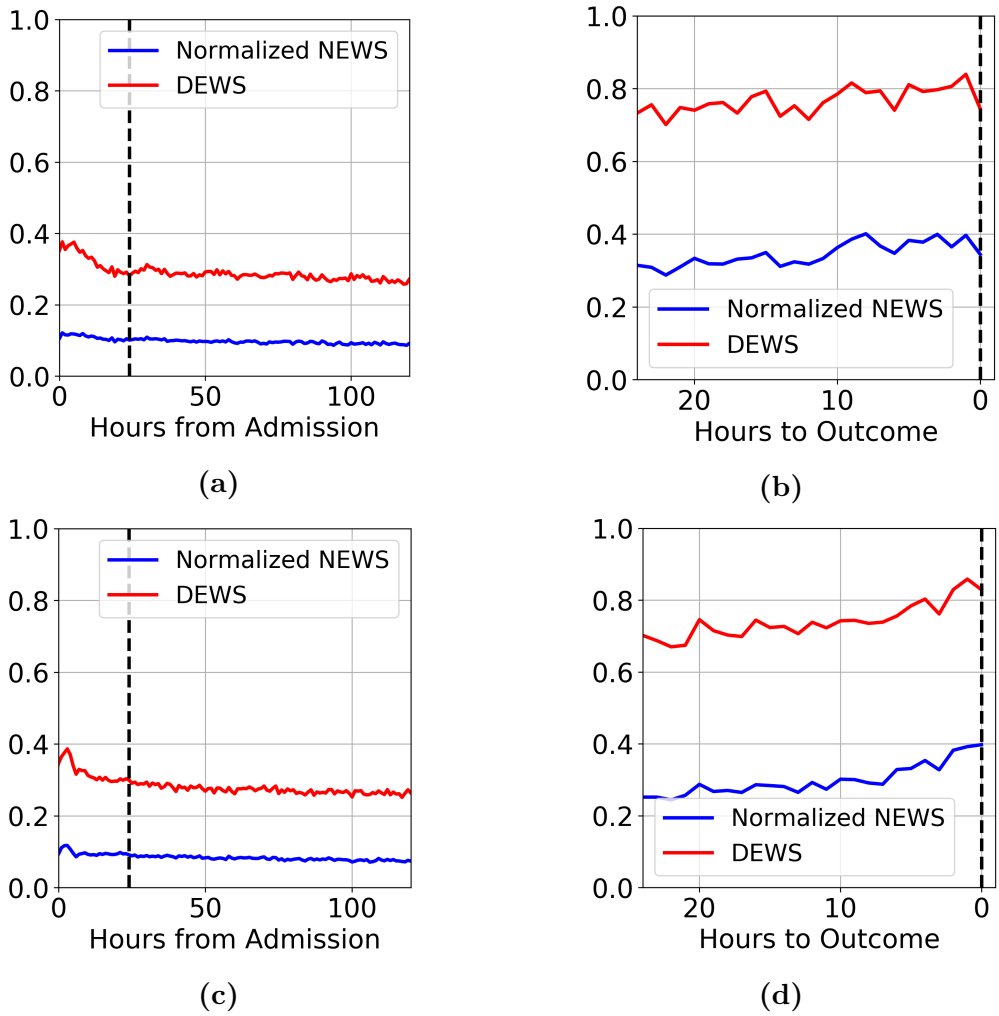


Fig. 5.6: Investigations of the mean probability of DEWS and the mean normalised NEWS score for (a) the first 120 hours from admission in $\mathcal{D}_{O,2}$, (b) the last 24 hours prior to the composite outcome in $\mathcal{D}_{O,2}$, (c) the first 120 hours from admission in $\mathcal{D}_P$, and (d) the last 24 hours prior to the composite outcome in $\mathcal{D}_P$. In (a) and (c), the dashed line indicates 24 hours from admission, and in (b) and (d) the dashed line indicates the time of the composite outcome.

Our proposed architecture DEWS achieved state-of-the-art performance across two independent test sets, even while considering a limited set of features. DEWS decreased the number of triggers in comparison to NEWS, especially amongst younger patients. This would ease the burden on clinical staff in such a demanding environment.

Furthermore, our model performed best across the composite outcome and the three individual outcomes (unplanned ICU admission, cardiac arrest, and mortality).

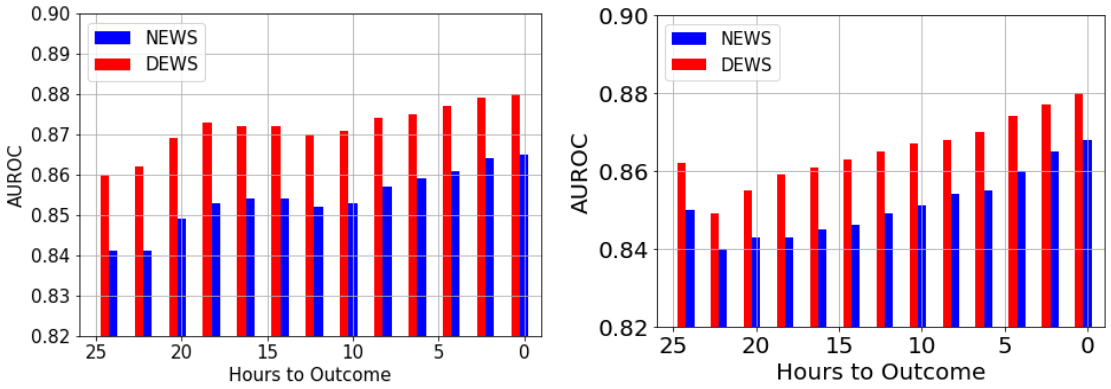


Fig. 5.7: Performance in terms of AUROC (vertical axis) at every two hours to an outcome (horizontal axis) for DEWS and NEWS, in the last 24 hours window prior to the composite outcome, in (a) $\mathcal{D}_{O,2}$ and $\mathcal{D}_P$.

We chose a composite outcome to avoid further class imbalance and as what is being done in the literature [28].

Existing EWS systems only assess the most recently collected vital signs, as i.i.d. data. We demonstrate how historical trends of vital signs can provide potentially beneficial supplementary information. By examining the attention weights assigned for each vital sign, in Fig. 5.5, we were able to demystify the decision-making process of our deep learning model. For example, the clinician could examine why DEWS was alerting by inspecting the time frames where the attention weights were highest in Fig. 5.5, such as increasing RR or decreasing SBP. Such trend analysis can support designing interventions on hospital wards. The alignment of scores between NEWS and DEWS, which is further illustrated in Fig. 5.6, emphasises their supplementary purposes. We envisage the system to provide the NEWS score, DEWS probability, and an overview of the importance of historical trends to the clinicians.

We also accounted for the correlations across vital signs using the trainable weights V_μ and V_σ, which learned the relationships across the aggregated context vectors of the vital signs. This incurred further computational complexity during training of the

deep learning model, but only represented a forward pass during testing. Future work includes experimenting with multi-task GPR (MGP) to account for correlations during feature transformation. However, the computational cost of MGP is $O(m^3n^3)$, which is prohibitive for low-resource settings in comparison to $m \times O(n^3)$ for the univariate GPR [70].

From a deployment perspective, the score can be easily incorporated into existing hospital devices, such as bedside monitors or hand held devices since it uses the same data streams as EWS systems, i.e. NEWS., but it calculates the score differently. Displaying the attention weights on the screen would require ergonomics specific analysis, which is beyond the scope of this thesis.

We focused on utilising physiological time-series data to establish analogous grounds with EWS systems currently deployed in clinical settings. We considered incorporating diagnosis codes as in previous studies [106, 91], however we decided that the inclusion of diagnosis codes may be impractical in real-life clinical settings because they are usually assigned at discharge for billing purposes. One study introduced a relevant attention-based model to predict in-hospital mortality in ICU [53]. However, their proposed attention-based multivariate LSTM model did not achieve a better AUROC than its variant without attention for predicting mortality. Additionally, the paper did not discuss the model's interpretability, despite it being one of their key motives for using attention.

Other scores have incorporated laboratory tests [56, 177], yet one of the main objectives of our work was to inspect trends of vital signs in real-time as in EWS. The reported mean AUROC for mortality in [56] ranged between 0.81 and 0.95, which is comparable to 92.6 AUROC [95% CI 92.6-92.7] in $\mathcal{D}_{O,2}$ and 91.6 AUROC [95% CI 91.6-91.6] in $\mathcal{D}_P$ of DEWS for mortality. While we could not directly compare to [177] since we do not have information available on resuscitation orders, their results for

mortality (AUROC 93.0) and the composite outcome (AUROC 77.0) are comparable to our results of 92.6 AUROC [95% CI 92.6-92.7] and 88.0 AUROC [95% CI 88.0-88.0] in $\mathcal{D}_{O,2}$, respectively. By comparing to previous studies, it can be concluded that vital signs are key predictor variables of clinical deterioration. We hypothesised that incorporating laboratory tests may improve performance, and we explore this in the next chapter.

Chapter 6

Information Fusion in a Multi-modal Early Warning System

6.1 Introduction

Electronic health records contain high-dimensional heterogeneous information that can be used by medical staff to assess patient status. Vital signs are among the most frequently measured variables and are thus widely used in EWS studies. However, there are other collected variables that may be useful predictors of deterioration. While a study published in 1990 showed that routine laboratory tests were not associated with consistent abnormalities prior to cardiac arrest [15], other studies showed that abnormal laboratory tests were associated with a high risk of mortality [99, 29]. The patient's history of illnesses and diseases may also be useful in predicting long term outcomes. This type of information is usually stored in the form of ICD-10 diagnosis codes [178] in the patient's EHR, and has been proven to be a good predictor of future clinical events [77]. We hypothesise that fusing information from multi-modal data present in EHRs may improve the performance of EWS systems in alerting for the composite outcome of unplanned ICU admission, cardiac arrest, and mortality.

In this chapter, we present a framework that maximises information retrieval from a patient record and prioritises interpretability. We developed and validated the *information Fusion in a multi-modal Early Warning System* (iFEWS) using vital signs, laboratory tests, and diagnosis codes. iFEWS combines variable-specific representations in a real-time alerting score using unsupervised representation learning, transfer learning, and supervised learning.

The chapter is organised as follows. Section 6.2 discusses the state-of-the-art for detecting clinical deterioration using multi-modal data. In Section 6.3, we present the algorithmic pipeline and key assumptions of the model. The framework is evaluated for the detection of the composite outcome in Section 6.4. We finally discuss the findings and the strengths and limitations of our model in Section 6.5.

6.2 State-of-the-art

Most of the existing traditional EWS systems primarily use vital-sign variables, such as MCEWS and NEWS [28, 25, 133]. The laboratory decision tree EWS (LDTEWS) was derived using commonly measured laboratory tests for the early discrimination of in-hospital death [29]. The system was later further combined with NEWS and proposed as LDTEWS:NEWS [30], which is a linear time-decay weighted function of the two individual scores. Such systems are generally based on numeric variables and cannot incorporate textual information such as diagnostic codes. One of their main limitations is that they do not learn information regarding the relationships between the different variable types. However, this also illustrates that there is some benefit from independently learning from each variable type, which we incorporate in the training scheme of our proposed model.

Machine learning and deep learning frameworks for clinical tasks also tend to consider sets of homogeneous data types in isolation [38]. Some works have focused on

modelling discrete medical codes only, such as diagnosis codes, medication codes, and procedure codes to predict future clinical events [47, 76, 89], while others used specific numeric features, such as vital signs [179]. While such frameworks perform well, their performance could potentially be improved by considering other variable types since they may contain relevant information.

Some studies have developed and evaluated deep learning frameworks to simultaneously learn from different data types. The most straightforward technique is to concatenate pre-processed features as a single input to a classifier [36, 54]. Although such models are simple and easy to train, it is difficult to interpret the influence of the different data types on the output.

Recently, more complex architectures were proposed to learn from heterogeneous data to predict various clinical end-points [180, 180, 33, 56]. Some of those architectures lack interpretability, since they focus on learning a joint representation to maximise performance [181]. Others investigate feature saliency by masking out individual features [33]. In our work, we sought to develop an inherently interpretable model by incorporating interpretability mechanisms within its architecture. This principled approach provides a better understanding of the relationships between the different variable types than systems that take a more 'black box' approach. Additionally, a real-time multi-modal deep learning-based EWS architecture has not yet been proposed in the literature.

The primary contribution of this chapter is a novel deep learning architecture that fuses information from different variable types through unsupervised and supervised learning techniques. The model learns from a combination of time-series vital-sign variables, laboratory tests, and the diagnosis assigned at admission. We evaluated the model's ability in detecting clinical deterioration prior to the composite outcome of unplanned ICU admission, mortality, and cardiac arrest.

6.3 iFEWS: a multi-modal alerting framework

The problem of detecting clinical deterioration is considered as a binary classification task. For each vital-sign measurement recorded for a patient, our model predicts the probability of a composite outcome within the next N hours. We consider each vital-sign measurement either as an event or non-event window $\mathcal{D}_W = [x_i, \mathbf{y}_i]_{i=1}^n$ and we here choose $N = 24$ hours, as described earlier in chapter 5. We also include a vector of the most recently-measured laboratory tests $\mathcal{D}_L = [x_l, \mathbf{z}]$ in the last k days. We choose $k = 5$ to align with the development of LDTEWS:NEWS [30], since it is considered as the state-of-the-art method a natural comparator for our work. Finally, we include the first ICD-10 diagnosis code d assigned to the patient at admission. In this case, d is a categorical variable. The model then estimates the posterior probability l of being within N hours of an adverse outcome, such that $l \in [0, 1]$.

It is widely acknowledged that the performance of deep learning models relies on the representation of the input data [59]. As summarised in Chapter 2, this involves learning an efficient representation of the explanatory features of the data, which can then be used for subsequent predictive tasks. The data in the EHR is heterogeneous in nature, ranging from both dense and sparse time-series variables, such as vital signs and laboratory tests, respectively, to discrete categorical variables such as diagnosis codes. We therefore treat the different variables based on how and when they were collected relative to the point of prediction as will be described below. We then train our model by learning an efficient representation of each variable type before combining those representations for our classification task. We now describe the data pre-processing and learning techniques for each variable type. The architecture for the vital-signs modality is shown in Fig. 6.1 and the overall architecture of the final model is shown in Fig. 6.2. The final hyperparameters of the models are described in the Deep Learning Experiments section.

6.3.1 Learning from vital signs

Since the vital signs are irregularly sampled, we first apply a patient-specific feature transformation for each window using GPR with a squared-exponential kernel to obtain equally sampled posterior mean and variance estimates. The squared-exponential kernel has been shown to be suitable for modelling physiological data [43]. These posterior mean and variance estimates are concatenated for all the vital signs to obtain: $\mathbf{Y}_\mu = [\mathbf{y}_{\mu,j}]_{j=1}^{m}$ and $\mathbf{Y}_\sigma = [\mathbf{y}_{\sigma,j}]_{j=1}^{m}$, where $\mathbf{Y}_\mu, \mathbf{Y}_\sigma \in \mathbb{R}^{m \times T}$ and $\mathbf{y}_{\mu,j}$ and $\mathbf{y}_{\sigma,j}$ are the GPR mean and variance for the j^{th} vital sign, such that $j = 1, \ldots, m$. Details of this GPR modelling approach were described earlier in Chapter 5.

Multi-channel Autoencoder

An autoencoder learns an efficient lower-dimensional representation of the (higher-dimensional) data through unsupervised learning. The basic architecture consists of an *encoder* that learns the compact latent representation from the input data, and a *decoder* that reconstructs the input using the latent representation.

To learn a joint latent representation of the vital signs, we jointly reconstruct the m vital signs by proposing a multi-channel attention-based autoencoder (MC-AE-ATT) that consists of m attention-based encoders and a single decoder as shown in Fig. 6.1. A single-channel encoder $E_{\mu,j}$ first processes a vital-sign sequence j independently using a BiLSTM network in order to maximise information retrieval in the forward and backward directions. The average of the forward and backward hidden-state outputs for vital sign j, $\overline{h_{t,j}}$, is then processed using an attention-based block $A_{\mu,j}$ to encode interpretability and compute the context vector:

$$\mathbf{c}_j = \sum_{t=1}^{T} \alpha_{t,j} \overline{h_{t,j}} \tag{6.1}$$

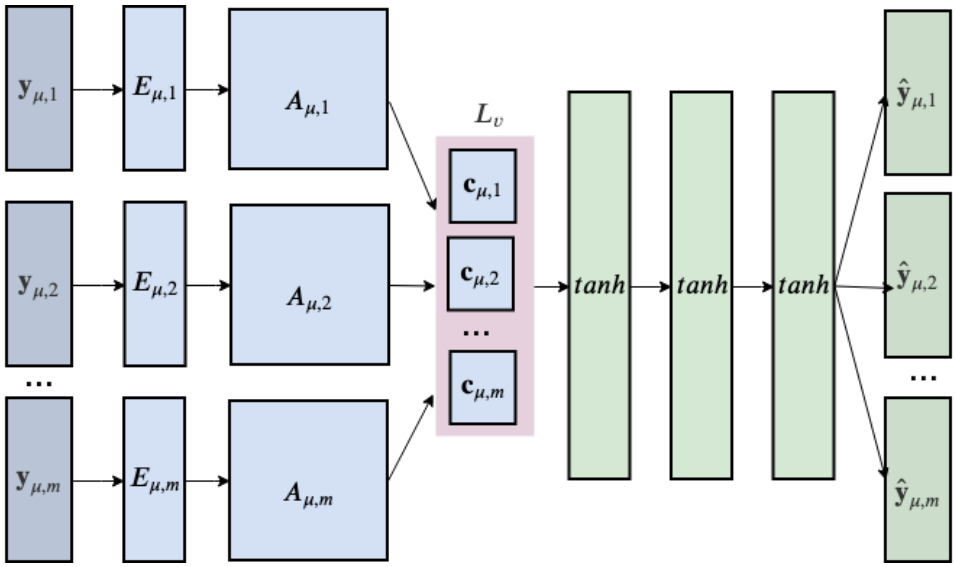


Fig. 6.1: Architecture of MC-AE-ATT for unsupervised feature learning from the vital signs. The multichannel encoder E_μ (*blue*) first processes the regularly sampled inputs $\mathbf{y}$ $(T \times 1)$ using a BiLSTM $(T \times e)$; and an attention unit (A_μ) that computes a context vector $(e \times 1)$ for each input. The context vectors are then concatenated to obtain the latent space representation L_v. L_v is then processed by a subsequent set of dense layers to obtain the reconstructions of the original inputs, denoted as $\hat{\mathbf{y}}_\mu$.

The details of the attention computations were previously described in more detail in Chapter 5. The context vectors of the m vital signs are concatenated to obtain the latent representation L_v:

$$L_v = [\mathbf{c}_1^T, ..., \mathbf{c}_m^T] \tag{6.2}$$

The latent representation is then mapped using three dense layers, as shown in Fig. 6.1, and a sigmoid function is applied to obtain the reconstructed input of all vital signs $\hat{\mathbf{y}}$:

$$\hat{\mathbf{y}} = \sigma(W_4\ g_3(W_3\ g_2(W_2\ g_1(W_1\ L_v + \mathbf{b_1}) + \mathbf{b_2}) + \mathbf{b_3}) + \mathbf{b_4}), \tag{6.3}$$

where W_1, W_2, and W_3 are the weight matrices and $\mathbf{b_1}$, $\mathbf{b_2}$, and $\mathbf{b_3}$ are the bias vectors of the dense layers. W_4 is the weight matrix and $\mathbf{b_4}$ is the bias vector of the final sigmoid layer. The activation functions of the dense layers are g_1, g_2, and g_3.

The parameters of the multi-channel autoencoder are optimised by minimising the binary cross-entropy loss for all the vital-sign sequences:

$$\mathcal{L}_{RL}(y,\hat{y}) = -\frac{1}{m \times T} \sum_{i=1}^{m \times T} y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i), \tag{6.4}$$

where $m \times T$ is the total number of input features from all of the vital-sign sequences.

Classification of the Latent Space

To predict the probability of an outcome within the next N hours using the vital-sign data l_v, the latent representation L_v is further processed using a multi-layer perceptron with a final sigmoid layer:

$$l_v = \sigma(W_v L_v + \mathbf{b}_v), \tag{6.5}$$

where W_v is the weights matrix and $\mathbf{b}_v$ is the bias vector. We denote this component of iFEWS as MC-AE-ATT-CL$_v$, corresponding to our multichannel autoencoder with attention (MC-AE-ATT) with subsequent (-CL$_v$) classification of the latent representation.

6.3.2 Learning from Laboratory Tests

In comparison to vital signs, laboratory tests are less frequently measured. We pre-process the laboratory tests in order to yield a real-time alerting score equivalent to the vital-sign measurements. First, we associated each vital-sign measurement with the most recently-collected set of laboratory tests $\mathcal{D}_L = [x_l, \mathbf{z}]$ during the previous $N \times k$ hours, where k is the number of days, x_1 is the time the laboratory tests were measured, and $\mathbf{z} \in \mathbb{R}^q$ is a vector of q (scalar-valued) laboratory-test measurements. The time between a vital-sign measurement and the laboratory test measurements is denoted as $t_{v-l} = x_n - x_l$, where x_n is the time of prediction based on the vital-sign

measurements. Physiologically implausible and missing values were replaced by the mean of the respective variable in the training set and the features were then scaled to obtain the final feature set $\hat{\mathbf{z}}$.

As a motive for allowing interpretability, we modelled the laboratory tests using a logistic regression model where we could simply assess the learned coefficients assigned to each laboratory-test variable:

$$l_l = \sigma(W_l \hat{\mathbf{z}} + \mathbf{b_l}), \tag{6.6}$$

where W_l is the weights matrix, $\hat{\mathbf{z}}$ is the vector of processed laboratory tests, and $\mathbf{b}_l$ is the vector of biases. We denote this module with the suffice $-CL_l$.

To account for the time difference between the vital-sign measurements and the laboratory-test measurements t_{v-1}, we further process l_l using an exponential decay model, such that the updated label becomes:

$$\hat{l_l} = l_l \exp^{-\lambda \frac{t_{v-1}}{Nk}}, \tag{6.7}$$

where λ is learned during training of the model. This equation adjusts the posterior probability of an outcome computed using the laboratory tests using the exponential decay model.

We considered two sets of laboratory tests as input variables: (1) set S consisted of 8 laboratory tests considered in our clinical baselines LDTEWS and LDTEWS:NEWS, and (2) set U consisted of 4 additional laboratory-test variables. (Set $S \cup U$ therefore contains 11 variables in total).

6.3.3 Learning from the first diagnosis

To ensure that our model is feasible, we only considered the diagnosis code (i.e. ICD-10 format [178]) assigned at admission as the patient's primary diagnosis. Each diagnosis code consists of several characters that represent a particular disease or illness. We first grouped the diagnosis codes into 21 groups based on the high-level grouping of the ICD-10 codes [178]. An additional group was created to represent missing or incorrect diagnosis codes that do not map to the ICD-10 dictionary [178]. Thus, in total there were 22 possible diagnosis categories. To learn a representation of the discrete diagnosis codes, we incorporated an embedding module with a non-negativity constraint. The embedding module thus maps each discrete code d into a latent vector of positive real numbers $\hat{\mathbf{d}}$. The latent vector $\hat{\mathbf{d}}$ is then used to estimate the probability of deterioration l_d:

$$l_d = \sigma(W_d l_d + \mathbf{b}_d), \tag{6.8}$$

where W_d is the weights matrix and $\mathbf{b}_d$ is the bias vector.

6.3.4 Label Fusion

The final output l was then computed using the three auxiliary outputs: the vital-signs label l_v, the time-adjusted laboratory-tests label $\hat{l}_l$, and the diagnosis label l_d, such that,

$$l = \sigma(W_o[l_v; \hat{l}_l; l_d] + \mathbf{b}_o). \tag{6.9}$$

Fig. 6.2 shows a schematic of the proposed model iFEWS. The inputs are first processed with different feature learning techniques to compute the auxiliary labels shown in the white squares (l_d, $\hat{l}_l$, and l_v). The final output l is then computed to indicate the occurrence of a composite outcome within the next N hours of a vital-sign measurement.

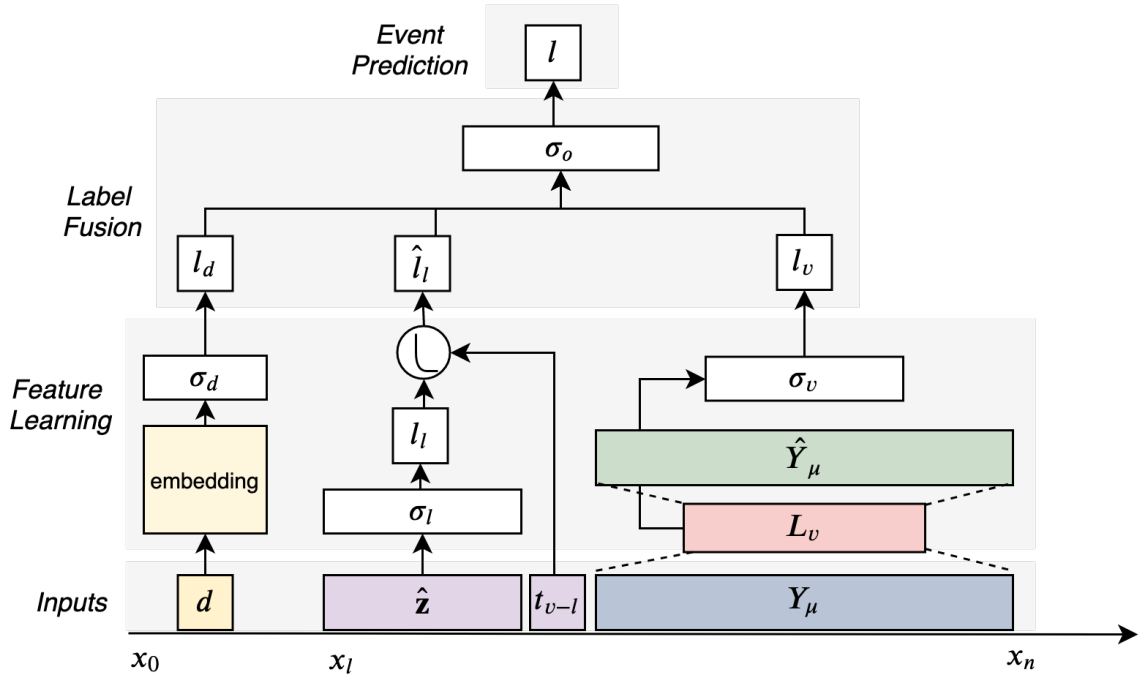


Fig. 6.2: Schematic of the iFEWS model that learns from the processed vital signs Y_μ, laboratory tests $\hat{\mathbf{z}}$, and the first diagnosis assigned at admission d, and fuses the auxiliary labels of each variable type to compute the posterior probabilities of an outcome within the next N hours (l_v, l_l, and l_d, respectively). t_{v-l} is used in the exponential decay function of the laboratory-tests label.

6.3.5 Continued Training

Unlike data encountered in computer vision and natural language processing, clinical datasets tend to be smaller in magnitude. Given the size of our proposed model, we optimise its performance by first pre-training its components independently and then fine-tuning their parameters as part of the larger model. Thus, in order to improve the efficiency of training our large model with the small training dataset, we train iFEWS in a two-fold process.

First, we pre-train MC-AE-ATT independently with the loss function in Eq. 6.4. We also pre-train CL_l with the binary cross-entropy loss but with a newly defined output $\tilde{l}_l \in (0,1)$, which indicates the occurrence of an event *at any time in the future during the current admission.*

Algorithm 2 Proposed framework of iFEWS that predicts whether a window $\mathcal{D}_W$ of vital-sign measurements is within N hours of an adverse outcome given a vector of the most recent laboratory-test measurements $\mathcal{D}_L$ in the last k days and the diagnosis code d that was first assigned at admission.

Input: Unlabelled window $\mathcal{D}_W = [x_i, \mathbf{y}_i]_{i=1}^n$, laboratory-test measurements $\mathcal{D}_L = [x_l, \mathbf{z}]$, and diagnosis code d.
Outputs: Label $l \in [0, 1]$.

Learning from Vital Signs

For $j = 1$ to m: ▷ *j is the index of the vital-sign variables.*
    $GPR_j \leftarrow$ Fit a GPR using $[x_i, y_i]_{i=1}^n$
    $\mathbf{y}_{\mu,j}$, $\mathbf{y}_{\sigma,j} \leftarrow GPR_j([x_t^*]_{t=1}^T)$
    $\mathbf{c}_\mu \leftarrow Encoder(\mathbf{y}_{\mu,j})$ ▷ *Encoder is the attention-based BiLSTM module.*
$L_v \leftarrow [\mathbf{c}_1, ..., \mathbf{c_m}]$ ▷ *L_v is the latent space of vital signs.*
$l_v \leftarrow MLP(L_v)$ ▷ *l_v is the posterior probability of an outcome computed using vital signs.*

Learning from Laboratory Tests

$\hat{\mathbf{z}} \leftarrow Pre-processing(\mathbf{z})$
$l_l \leftarrow LR(\hat{\mathbf{z}})$
$t_{v-1} \leftarrow x_n - x_l$
$\hat{l}_l \leftarrow l_l \exp^{-\{\lambda \frac{t_{v-1}}{Nk}\}}$ ▷ *l_l is the probability of an outcome computed using laboratory tests.*

Learning from the Diagnosis

$\hat{\mathbf{d}} \leftarrow Embedding(d)$
$l_d \leftarrow LR(\hat{\mathbf{d}})$ ▷ *l_d is the posterior probability of an outcome computed the diagnosis code assigned at admission.*

Label Fusion

$l \leftarrow LR(l_v, l_l, l_d)$ ▷ *l is the overall posterior probability of being within N hours of an adverse outcome.*

The pre-trained weights of MC-AE-ATT and CL_l were then used to initialise their corresponding weights in iFEWS. The classification objective of iFEWS is the binary cross-entropy loss of the true labels $\tilde{l}$ and the predicted labels:

$$\mathcal{L}_{CL}(\tilde{l}, l) = -\frac{1}{N}\sum_{i=1}^{N} \tilde{l}_i \cdot \log(l_i) + (1 - \tilde{l}_i) \cdot \log(1 - l_i), \tag{6.10}$$

where N is the number of training samples.

The final objective function of iFEWS consisted of the joint loss function:

$$\mathcal{L}_{JT} = \mathcal{L}_{RL}(y, \hat{y}) + \mathcal{L}_{CL}(\tilde{l}, l). \tag{6.11}$$

We included the reconstruction loss function of the MC-AE-ATT component, since it contains the majority of parameters that compute the latent representation of the vital-sign measurements. (We note that losses $\mathcal{L}_{RL}$ and $\mathcal{L}_{CL}$ could be combined in the affine $\beta\mathcal{L}_a + (1-\beta)\mathcal{L}_b$, and $\beta = 0.5$ performed best empirically for our task.)

6.4 Empirical Study

In this section, we summarise the key results of our experimentation and design choices. As in Chapter 5, due to class imbalance, we used the balanced dataset $\mathcal{D}_{O,1B}$ obtained from OUH for training and validation. The validation set hereafter is denoted as $\mathcal{D}_{O,1V}$. We used the imbalanced OUH dataset $\mathcal{D}_{O,2}$ and the imbalanced PH dataset $\mathcal{D}_P$ for testing the performance of the models.

6.4.1 Model Variants as Baselines

To evaluate the effect of the design choices on the overall performance of the model, and to justify model complexity, we assess several simpler variants of iFEWS. For learning the representation of the vital signs, we first developed and evaluated a single-channel autoencoder (SC-AE) that simply concatenated all the vital-sign sequences as one input. The inputs were then processed by three dense layers. In order to encode temporal information, we then designed the multichannel autoencoder (MC-AE) that processed each vital-sign sequence independently using an BiLSTM network. Since the BiLSTM network lacks interpretability, we finally incorporated the attention mechanism in each channel (MC-AE-ATT).

We also compared our model to LDTEWS and LDTEWS:NEWS as standard clinical benchmarks. Both LDTEWS and LDTEWS:NEWS only included 8 routinely collected laboratory tests (i.e. Hb, WCC, U, ALB, CR, NA, and K) as included in set S. We further included TROP, HCT, TBIL, and CRP in set U and evaluated our deep learning models using both sets.

6.4.2 Evaluation Metrics

We evaluated the performance of our models using several metrics based on the respective task. For the autoencoders, we measured the mean squared error (MSE) on the both testing sets to assess the reconstruction quality.

During model development and validation, we assessed the model variants and components using AUROC and AUPRC. For our proposed iFEWS model and other classifiers, we used the AUROC, sensitivity, specificity, and PPV evaluated on the testing sets. All metrics were performed using a bootstrapping technique with replacement with a fixed number of bootstraps (nb). We compared the performance of the models across the patients aged 16-45 years and $>$ 45 years, and across the three outcomes (unplanned ICU admission, cardiac arrest, and mortality) independently.

6.4.3 Deep Learning Experiments

All hyperparameters of the model were optimised empirically using the balanced training and validation set $\mathcal{D}_{O,1B}$. The regularly-spaced mean vital-sign measurements ($\mathbf{y}_\mu$) were transformed with min-max scaling of [0,1]. All of the vital-sign autoencoder models were trained with 20 epochs, with early stopping by monitoring the loss on the validation set. The encoder module of the SC-AE consisted of four dense layers with 64 nodes followed by a latent-space dense layer consisting of 12 nodes. The decoder module of the SC-AE consisted of four dense layers with 64 nodes and a final sigmoid

layer with 84 output nodes (corresponding to the 12 equidistant timesteps of the 7 vital signs). The encoder of the MC-AE model consisted of a BiLSTM with 5 output nodes at each timestep and the decoder consisted of four dense layers with 64 nodes each. The classifier consisted of five dense layers and a final sigmoid layer.

To assess the predictive power of vital signs and the continued learning scheme, we trained MC-AE-ATT-CL$_v$ independently using three different training schemes. The first training scheme involved pre-training MC-AE-ATT independently, and then fixing its weights during the training of the latent space classifier $-CL_v$. The second scheme involved joint training of MC-AE-ATT and the latent space classifier $-CL_v$ with random initialisation of weights. The third scheme, continued learning, involved pre-training the MC-AE-ATT independently followed by joint learning with the latent space classifier $-CL_v$.

The laboratory-test measurements were transformed using standardisation with a zero mean and unit variance. For the models using laboratory tests, we trained and evaluated our models for the original label $\tilde{l}$ (i.e. the vital-sign measurements are within N hours of an outcome). The models were trained with 100 epochs with early stopping by monitoring the classification loss on the validation set in order to avoid overfitting.

The diagnosis codes embedding module performed best when it computed 3-dimensional vector representations. We also compared embeddings to one-hot encoding, and we found that (in experiments not shown here for brevity) that the model using embeddings performed better. We also did not pre-train the embedding in the continued learning training scheme because it did not show any predictive power when learning in isolation of components of the larger models.

Table 6.1: Mean and standard deviation of the mean squared error on the training set $\mathcal{D}_{O,1B}$ and the testing sets $\mathcal{D}_{O,2}$ and $\mathcal{D}_P$ using the different autoencoder architectures for reconstructing all vital signs. All values are on a scale of 10^{-3}.

Model	$\mathcal{D}_{O,1B}$	$\mathcal{D}_{O,2}$	$\mathcal{D}_P$
SC-AE	2.00 ± 3.60	1.08 ± 2.06	2.24 ± 15.66
MC-AE	2.60 ± 4.69	1.39 ± 2.79	2.49 ± 15.27
MC-AE-ATT	4.14 ± 6.32	2.72 ± 4.45	3.70 ± 16.17

Weights that were not pre-initialised with the continued learning scheme were randomly initialised. All the models were optimised using the Adam optimiser and implemented using Keras (v 2.2.2) [175] with a TensorFlow backend (v 1.5.0) [176].

6.4.4 Results

Feature Learning of Vital Signs

The reconstruction errors in terms of the MSE of the vital-sign sequences in $\mathcal{D}_{O,1B}$, $\mathcal{D}_{O,2}$ and $\mathcal{D}_P$ are shown in Table 6.1. The MSE increases as the model complexity increases across all datasets. While MC-AE-ATT is the most interpretable since it incorporates an attention mechanism, it yields the highest reconstruction error in all datasets. When comparing the independent testing sets, we note that $\mathcal{D}_P$ has the highest standard deviation of errors. This may be because the autoencoder module is overfitting to the data in $\mathcal{D}_{O,1B}$. Additionally, the vital-sign sequences in $\mathcal{D}_P$ were scaled using transformations learned from the training dataset $\mathcal{D}_{O,1B}$. On the other hand, the errors are lower in the $\mathcal{D}_{O,2}$ testing set because $\mathcal{D}_{O,1B}$ and $\mathcal{D}_{O,2}$ belong to the same distributions as they were both obtained from the OUH.

Fig. 6.3 shows reconstructed respiratory rate windows from a sample window in (a) $\mathcal{D}_{O,1B}$ and (b) $\mathcal{D}_{O,2}$. We notice that the three autoencoder models learn general trends of the input data, with a better reconstruction quality for the former (training set) than the latter (testing set).

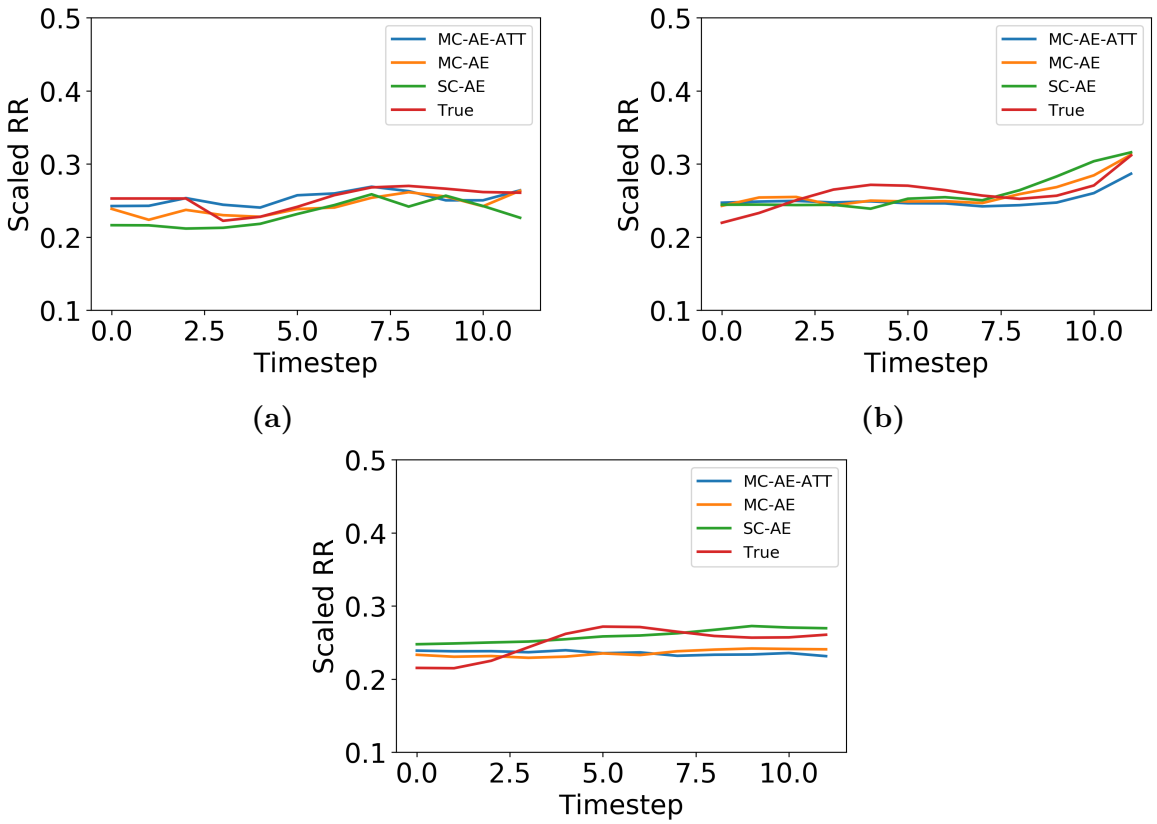


Fig. 6.3: Reconstructed examples of a respiratory rate sequence using the different autoencoder architectures in comparison to the ground-truth (in red) from (a) $\mathcal{D}_{O,1B}$, (b) $\mathcal{D}_{O,2}$, and (c)$\mathcal{D}_P$.

Table 6.2: Performance on the validation set $\mathcal{D}_{O,V}$ using the MC-AE-ATT with classification of the latent space ($-CL_v$) and the respective numbers of trainable and non-trainable parameters. Mean and confidence intervals were evaluated using a bootstrapping technique ($nb = 1,000$).

Training Scheme	Trainable parameters	Non-trainable parameters	AUROC	AUPRC
Pre-initialisation	12,147	33,775	85.8 (85.7-85.8)	86.4 (86.3-86.4)
Joint Learning	45,992	0	88.0 (87.9-88.0)	**89.2 (89.2-89.2)**
Continued Learning	45,992	0	**89.4 (89.3-89.4)**	87.9 (87.8-87.9)

Table 6.2 presents the performance of the different training schemes on the validation set $\mathcal{D}_{O,V}$. Pre-initialisation has the lowest number of trainable parameters, since it only involves training of the latent space classifier. It also achieves the lowest AUROC [95%

CI 85.7-85.8] and AUPRC [95% 86.3-86.4] values across all schemes. Continued learning achieves the highest AUROC [95% 89.3-89.4] across all schemes; we choose to adopt it for training our overall model. We note that the AUPRC values are considerably high since the validation set $\mathcal{D}_{O,V}$ is balanced as is the training set from which it was derived.

Predictive Power of Laboratory Tests

Table 6.3 summarises the performance of LDTEWS and the LR models on the validation set $\mathcal{D}_{O,1V}$ using the two sets of laboratory-test variables, S and U. LDTEWS achieves the lowest performance for both labels in terms of AUROC [95% CI 67.9-68.0] and AUPRC [95% CI 67.1-67.2]. We also observe that LR achieves the highest AUROC [95% 72.3-72.4] and AUPRC [95% CI 73.1-73.2] when using the laboratory-tests dataset U. This suggests that incorporating the additional variables in set U over set S improves the predictive performance of a laboratory-tests based classifier.

Table 6.3: Performance evaluation of simple logistic regression using laboratory tests in comparison to the clinical baseline (LDTEWS) on the validation set $\mathcal{D}_{O,1V}$. Note that S denotes the set of variables considered in LDTEWS and U denotes includes four additional laboratory tests. Mean and confidence intervals were evaluated using a bootstrapping technique ($nb = 1,000$).

Model	AUROC	AUPRC
LDTEWS_S	67.9 (67.9-68.0)	67.1 (67.1-67.2)
LR_S	70.6 (70.5-70.7)	71.4 (71.3-71.4)
LR_U	**72.3 (72.3-72.4)**	**73.1 (73.1-73.2)**

Performance Evaluation of iFEWS

Table 6.4 summarises the performance results of the final models on $\mathcal{D}_{O,2}$. The iFEWS variant without attention (iFEWS_{MC-AE}) achieved the highest AUROC [95% 90.2-90.3] and sensitivity [95% 77.7-77.8]. Compared to the clinical baseline

that is currently adopted in practice, NEWS, iFEWS$_{MC-AE}$ achieves increases the AUROC by approximately 4%. iFEWS$_{SC-AE}$ achieved the lowest AUROC [95% CI 89.6-89.7] across the three autoencoder-based variants of iFEWS. We note that MC-AE-ATT has the highest reconstruction error (as shown in Table 6.1), which implies that incorporating attention may lead to information loss during reconstruction and explains the lower performance of iFEWS$_{MC-AE-ATT}$ than iFEWS$_{MC-AE}$. All models achieved a comparable PPV since the specificity was set to 89.0.

Table 6.4: Performance evaluation of the different classifiers on $\mathcal{D}_{O,2}$. The decision threshold of all classifiers was adjusted to achieve a specificity similar to that of NEWS ($\approx$ 89.0). The subscripts indicate (i) what types of features were used in the LR model and (ii) the type of autoencoder in iFEWS. Mean and confidence intervals were evaluated using a bootstrapping technique ($nb = 1,000$).

Model	AUROC	Sensitivity	Specificity	PPV
i.i.d.				
NEWS	86.6 (86.5-86.6)	70.2 (70.2-70.3)	88.8 (88.8-88.8)	5.0 (5.0-5.0)
LDTEWS:NEWS	88.5 (88.5-88.6)	72.8 (72.7-72.8)	88.8 (88.8-88.8)	5.1 (5.1-5.1)
LR$_{L,V,D}$	88.6 (88.6-88.7)	74.3 (74.2-74.3)	88.8 (88.8-88.8)	5.3 (5.2-5.3)
Model Variants				
iFEWS$_{SC-AE}$	89.6 (89.6-89.6)	75.1 (75.1-75.2)	88.8 (88.8-88.8)	5.3 (5.3-5.3)
iFEWS$_{MC-AE}$	**90.3 (90.2-90.3)**	**77.7 (77.7-77.8)**	**88.9 (88.9-88.9)**	**5.5 (5.5-5.5)**
iFEWS$_{MC-ATT-AE}$	90.0 (89.9-90.0)	76.5 (76.4-76.6)	88.8 (88.7-88.8))	5.4 (5.4-5.4)

Table 6.5 compares the performance of iFEWS$_{MC-AE-ATT}$ on sub-populations in $\mathcal{D}_{O,2}$ with LDTEWS:NEWS. We refer to iFEWS$_{MC-AE-ATT}$ as iFEWS hereafter for brevity. Across the younger patients, iFEWS achieved a higher AUROC than LDTEWS:NEWS, [95% CI 87.6-87.9] and [95% CI 82.8-83.2] respectively. The performance of iFEWS for 16-45 years old patients is also superior to that of the supervised learning model DEWS presented in Chapter 5 (AUROC [95% CI 81.8-82.2]) and NEWS (AUROC [95% CI 75.7-76.2]). This represents more than 10% increase in the performance of the current state-of-the-art (i.e. NEWS) for the younger patient group. For the group of elder patients, unplanned ICU admission, and cardiac arrest, iFEWS consistently performed better than LDTEWS:NEWS in terms of the AUROC. For

mortality, iFEWS achieved a similar AUROC to LDTEWS:NEWS, [95% CI 93.7-93.8] and [95% CI 93.6-93.7] respectively. However, iFEWS had a higher sensitivity, [95% CI 84.5-84.7] compared to [95% CI 83.0-83.2].

Table 6.5: Performance evaluation of iFEWS in comparison to LDTEWS:NEWS across sub-populations of interest i.e. 16-45 years old, > 45 years old, and each of the three events in the composite outcome, in $\mathcal{D}_{O,2}$. The adjusted decision threshold for iFEWS was 0.63, to achieve a similar overall specificity of the clinical benchmark NEWS (≈ 89.0). Mean and confidence intervals were evaluated using a bootstrapping technique ($nb = 1,000$) for the respective sub-population.

Model	AUROC	Sensitivity	Specificity	PPV
16-45 years old				
iFEWS	**87.8 (87.6-87.9)**	**60.0 (59.7-60.4)**	93.8 (93.8-93.9)	**3.2 (3.2-3.2)**
LDTEWS:NEWS	83.0 (82.8-83.2)	50.0 (49.6-50.3)	**94.3 (94.2-94.3)**	2.9 (2.9-3.0)
> 45 years old				
iFEWS	**89.9 (89.9-89.9)**	**77.5 (77.4-77.6)**	**88.0 (87.9-88.0)**	**5.6 (5.5-5.6)**
LDTEWS:NEWS	88.7 (88.6-88.7)	74.2 (74.1-74.2)	87.9 (87.9-87.9)	5.3 (5.3-5.3)
Unplanned ICU				
iFEWS	**84.8 (84.7-84.8)**	**63.9 (63.7-64.1)**	90.1 (90.1-90.1)	**1.5 (1.5-1.5)**
LDTEWS:NEWS	80.7 (80.6-80.8)	54.7 (54.5-54.8)	**90.1 (90.1-90.1)**	1.5 (1.5-1.5)
Cardiac Arrest				
iFEWS	**81.1 (81.1-81.3)**	**53.1 (52.8-53.4)**	**90.2 (90.2-90.2)**	**0.4 (0.3-0.4)**
LDTEWS:NEWS	78.1 (77.9-78.2)	49.9 (49.5-50.2)	90.1 (90.1-90.1)	0.4 (0.3-0.4)
Mortality				
iFEWS	**93.7 (93.7-93.8)**	**84.6 (84.5-84.7)**	**88.9 (88.9-88.9)**	**3.9 (3.9-4.0)**
LDTEWS:NEWS	93.6 (93.6-93.7)	83.1 (83.0-83.2)	88.8 (88.8-88.8)	4.0 (4.0-4.0)

Table 6.6 presents the performance of iFEWS across the different patient sub-populations in $\mathcal{D}_P$. For the overall dataset, iFEWS achieved a higher AUROC than LDTEWS:NEWS, [95% CI 89.5-89.5] and [95% CI 88.5-88.6] respectively. As for the 16-45 years old, iFEWS achieved a higher AUROC [95% CI 93.9-94.0] than LDTEWS:NEWS [95% 88.5-88.8]. For the elder patient group and across all outcomes, iFEWS had a higher AUROC than LDTEWS:NEWS. Thus, even on a completely independent testing set, we conclude that iFEWS had superior discriminatory performance than the baseline EWS.

Table 6.6: Performance evaluation of iFEWS in comparison to LDTEWS:NEWS across sub-populations of interest i.e. 16-45 years old, > 45 years old, and each of the three events in the composite outcome, in $\mathcal{D}_P$. The adjusted decision threshold for iFEWS is 0.63 to achieve a similar overall specificity of the clinical benchmark NEWS ($\approx$ 89.0). Mean and confidence intervals were evaluated using a bootstrapping technique ($nb = 1,000$) for the respective sub-population.

Model	AUROC	Sensitivity	Specificity	PPV
Overall				
iFEWS	**89.5 (89.5-89.5)**	**72.6 (72.5-72.6)**	89.3 (89.3-89.4)	6.9 (6.9-6.9)
LDTEWS:NEWS	88.6 (88.5-88.6)	68.6 (68.5-68.6)	**90.6 (90.6-90.6)**	**7.3 (7.3-7.3)**
16-45 years old				
iFEWS	**94.0 (93.9-94.0)**	**75.5 (75.2-75.8)**	93.5 (93.5-93.5)	6.2 (6.2-6.3)
LDTEWS:NEWS	88.7 (88.5-88.8)	60.5 (60.2-60.9)	**95.5 (95.5-95.5)**	**6.2 (6.2-6.3)**
> 45 years old				
iFEWS	**89.2 (89.2-89.2)**	**72.9 (72.9-73.0)**	89.0 (89.0-89.0)	**7.3 (7.3-7.3)**
LDTEWS:NEWS	88.4 (88.4-88.5)	68.9 (68.8-68.9)	**90.2 (90.1-90.2)**	7.0 (6.9-7.0)
Unplanned ICU				
iFEWS	**88.5 (88.5-88.6)**	**65.4 (65.3-65.5)**	91.0 (91.0-91.0)	2.5 (2.5-2.5)
LDTEWS:NEWS	85.0 (85.0-85.1)	57.4 (57.3-57.5)	**92.1 (92.1-92.1)**	2.5 (2.5-2.5)
Cardiac Arrest				
iFEWS	**82.9 (82.8-83.0)**	**56.6 (56.4-56.7)**	91.0 (91.0-91.1)	1.2 (1.2-1.2)
LDTEWS:NEWS	82.4 (82.3-82.4)	52.5 (52.3-52.6)	**92.1 (92.1-92.1)**	1.2 (1.2-1.3)
Mortality				
iFEWS	**92.1 (92.1-92.1)**	**79.0 (78.9-79.0)**	89.6 (89.6-89.6)	5.4 (5.4-5.4)
LDTEWS:NEWS	91.7 (91.6-91.7)	75.1 (75.1-75.1)	**90.8 (90.8-90.8)**	**5.8 (5.8-5.8)**

Features Saliency

To get a better understanding of the decision-making process of our model, we examined feature saliency of the LR components. This involved investigating the weights assigned to the features after model training in the sigmoid-based layers. Fig. 6.4 visualises the magnitude of the weights in W_l of the LR of the laboratory tests. We notice that three of the four additional variables considered in U are ranked within the top six weights. Additionally, UR and CR are assigned the highest absolute weights in comparison to the other variables. This is aligned with the clinical literature where abnormal UR levels are associated with heart failure, whereas small increases in CR levels have been shown to be significantly associated with mortality [146]. On the other hand, HCT, K, and Na are associated with the smallest coefficients.

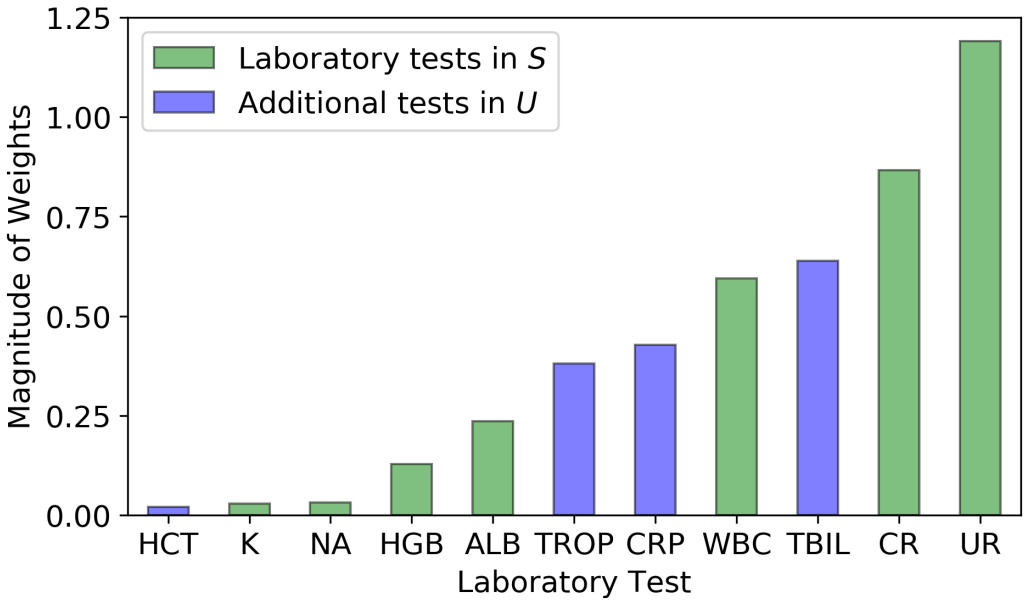


Fig. 6.4: The absolute value of the weights assigned to the laboratory-test variables with the LR model trained with set U in iFEWS. The variables originally included in LDTEWS and LDTEWS:NEWS are shown in green. The four additional variables included in U are shown in blue.

One interesting result is that TROP was assigned a higher coefficient than ALB, HGB, NA, K, and HCT, despite it being the least collected lab-test variable ($< 1\%$)

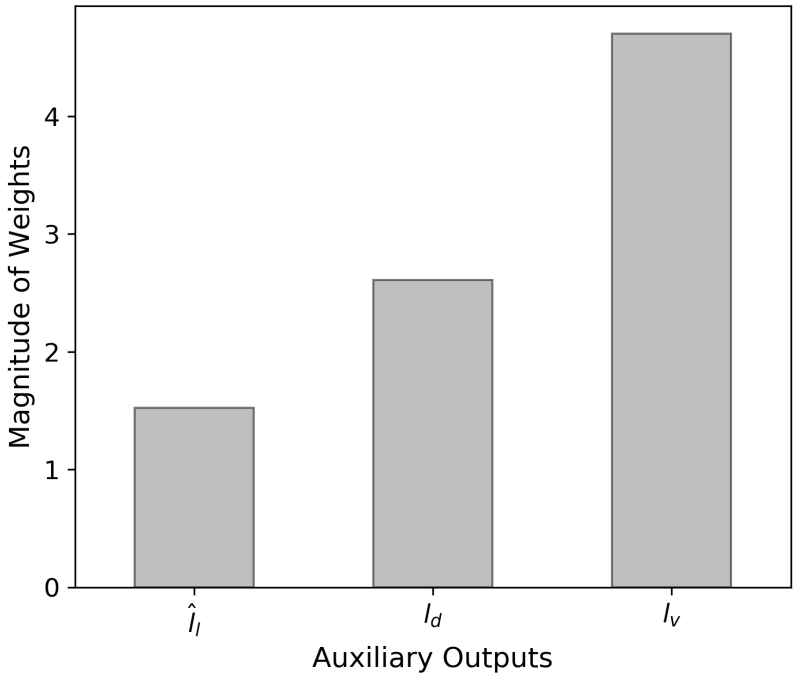


Fig. 6.5: Visualisation of the magnitude of coefficients assigned to the auxiliary outputs to compute the overall escalation label l, where l_v was computed using vital-sign variables, l_d was computed using the diagnosis code assigned at admission, and $\hat{l}_l$ was computed using the laboratory-test variables.

across all datasets as shown in Table 3.4. TROP is a protein in the blood that measures contractions in the heart muscle and is an indicator of heart muscle damage [148]. It is usually ordered by the clinician when the patient has chest pain or other symptoms of a heart attack. Despite its low frequency, the model was able to learn a strong association between TROP and the risk of the composite outcome since TROP is a high indicator of heart illness.

We also examined the weights assigned to the auxiliary outputs using the different variable types. Fig. 6.5 visualises the magnitude of the weights in the form of a bar chart. We observe that the highest absolute weight is assigned to the label computed using the vital-sign variables (l_v), which is approximately double the absolute weights assigned to the other variable types.

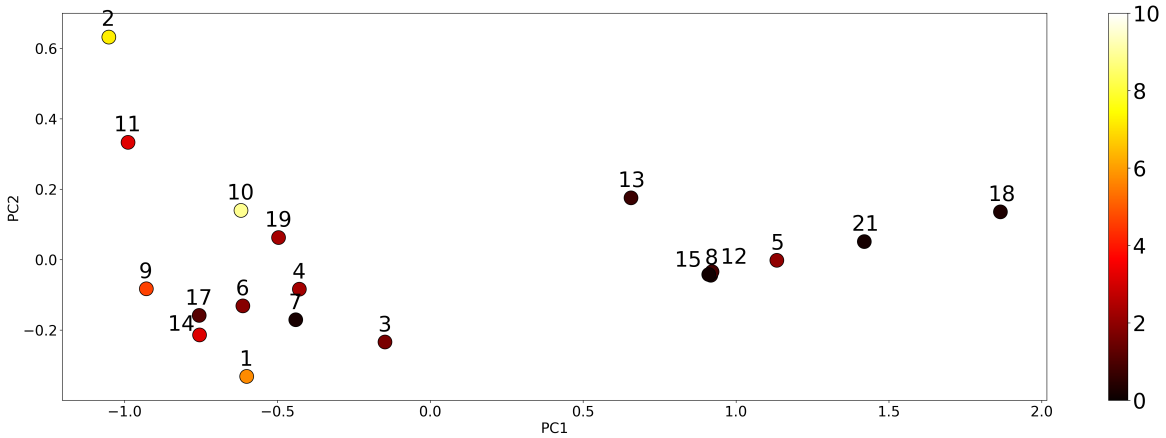


Fig. 6.6: Visualisation of the embeddings learned from the diagnosis groups in iFEWS. The horizontal and vertical axis represent the two principal components of the originally 3-dimensional embedding space, obtained using PCA. The annotations represent the diagnosis groups and the colour bar represents the proportion of patients who experienced an outcome within the patients who were admitted due to the respective diagnosis group.

We also investigated what the model learned through its embedding module, which converted grouped diagnosis codes into 3-dimensional vector. To do so, we first used PCA, a standard statistical procedure that are used to project the 3-dimensional vectors into a 2-dimensional space. The results are shown in Fig. 6.6. The annotations on the scatter-plot represent the diagnosis groups. The colour bar represents the proportion of patients who experienced a composite outcome and were assigned with the respective diagnosis group, across all patients who were admitted with the respective diagnosis group. We observe that the diagnosis groups that have a higher proportion of patients experiencing the composite outcome are clustered closer to each other on the left side of the plot. The diagnosis groups with the highest proportion of the outcome are 2 (diagnosed with cancer tumours) and 10 (diagnosed with diseases of the respiratory system).

Clinical Utility

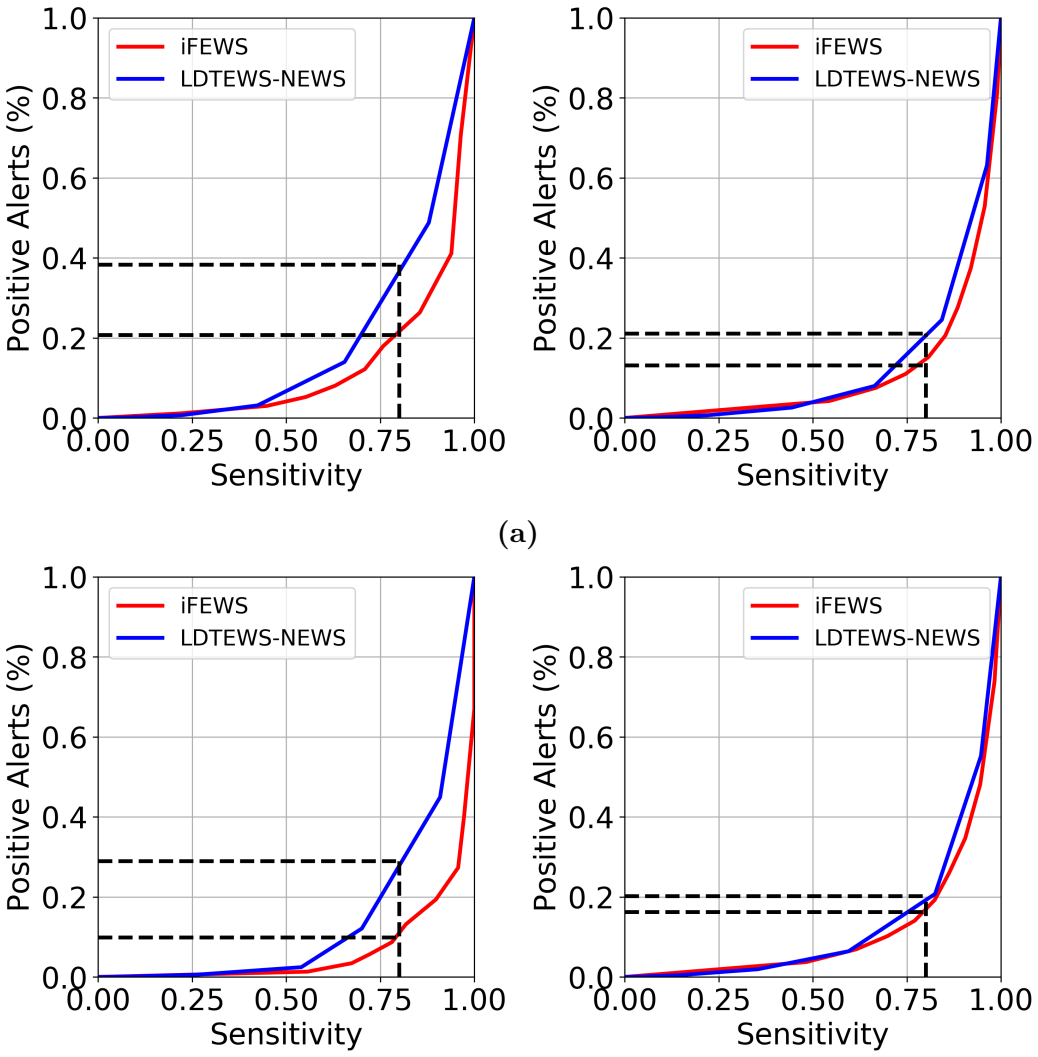
Fig. 6.7 shows the percentage of triggers, or positive alerts, produced by iFEWS in comparison to LDTEWS-NEWS at different sensitivity values (horizontal axis) in the two testing sets. For the 16-45 years old patients (Fig. 5.4 (a) left column), iFEWS produces approximately 14.5% fewer positive alerts than LDTEWS:NEWS to achieve the same level of sensitivity. Across the $>$ 45 years old patients (right column), iFEWS has approximately a 6% lower trigger rate than LDTEWS:NEWS.

In $\mathcal{D}_P$, the trigger rate of iFEWS is approximately 17% lower than that of LDTEWS:NEWS across the 16-45 years old patients, and approximately 2% lower across the >45 years old patients. The performance of iFEWS in comparison to LDTEWS:NEWS in terms of the trigger rate and the AUROC presented earlier highlights our model's ability in easing staff burden of positive alerts and superior discrimination ability.

6.5 Discussion

In this chapter, we propose iFEWS, a deep early warning system that learns from multi-modal data in EHRs. The proposed model achieved state-of-the-art performance across two unseen testing sets, and a superior performance over the clinical state-of-the-art NEWS and a recently proposed clinical score LDTEWS:NEWS.

iFEWS is also inherently interpretable, which is an essential characteristic of a model that could be potentially deployed in clinical practice. For example, the coefficients of the laboratory-tests classifier and the label fusion module were helpful in understanding the decision-making process of the overall model. However, although the incorporation of attention on the vital-sign variables may lead to better interpretability, as shown in



(a)

(b)

Fig. 6.7: Efficiency curves plotting sensitivity (horizontal axis) against the trigger rate, or the percentage of observations (vertical axis) with an iFEWS probability or normalised LDTEWS:NEWS score greater than or equal to a decision threshold, with the decision threshold ranging between 0 and 1, for 16-45 years old patients (left column) and $>$ 45 years old patients (right column) in $\mathcal{D}_{O,2}$ (top row) and $\mathcal{D}_P$ (bottom row)

. The vertical black dashed line represents a fixed sensitivity of 80%.

the previous chapter, it may not necessarily yield a better predictive performance as shown in Table 6.4. This needs to be further investigated in future studies.

There are several areas of improvement for our current approach. First, our framework, as in NEWS, LDTEWS, and LDTEWS:NEWS, assumes that patient

admissions are independent. This approach may be neglecting relevant information from previous admissions, such as comorbidities. Processing diagnosis codes from a previous admission, in particular, could be improved using natural language processing techniques. Secondly, our model does not learn any temporal information from the laboratory tests. However, this was a challenging task due to the low sampling frequency and will be further investigated in our future work.

Another area of future work relates to improving the performance of the model by exploring other design choices. Such design choices include incorporating the variance of the interpolated vital-sign sequences and presence vectors for the other features. We hypothesise that this may improve the performance of the model, as it improved the performance of DEWS in the previous chapter.

We would also like to use a variational autoencoder (VAE), which is a generative model that learns a probabilistic latent space. Such a model could potentially be extended to other relevant clinical applications, such as for generating synthetic vital-sign sequences to increase the dataset size.

Chapter 7

Conclusion

7.1 Thesis Summary

The detection of clinical deterioration amongst emergency in-hospital admissions is a key priority for developed healthcare systems [182]. Existing methods, derived either heuristically [23] or by univariate methods [27], do not account for individual variations and trends in the data. Increased access to medical data enables the development of relevant machine learning models that can assist in the clinical decision-making process. In this thesis, we focused on developing improved early warning models that continuously assess the patient's status for signs of deterioration. We have designed several novel end-to-end architectures and assessed their performance for heterogeneous patient populations. The contributions are summarised in the following paragraphs.

In Chapter 4, we developed a statistical model that considers patient-specific attributes, mainly demographics. Our results suggest that incorporating age in the design of standard rule-based systems can improve the performance for young patients, while remaining at par with NEWS for elderly patients. While the guidelines of the Royal College of Physicians for using NEWS claim that the system would become complex and cumbersome to use if additional demographic-related parameters are

included in the model [20], we argue that the benefit of incorporating age for younger patients outweighs the need for a simple system. Additionally, our proposed system (AEWS) can be easily remodelled using other external datasets, unlike heuristically-developed systems, such as NEWS, that rely on clinical judgement [25].

In Chapter 5, we proposed an end-to-end learning framework that uses Bayesian non-parametrics and deep learning. The framework first pre-processes sparse time-series vital-sign data using patient-specific Gaussian Process Regression. We have also shown that adopting prior distributions for the hyperparameters enables efficient modelling of a large-scale dataset. The recurrent network classifier then predicts whether the patient is deteriorating or not. While the model is considered to be less interpretable than rule-based systems, we offer an alternative analysis of interpretability by incorporating attention mechanisms in the classifier. The attention weights indicate which patterns are most contributing to the decision of the classifier, which we consider to be an essential component of the clinical utility of our model.

While the first two approaches focused on the use of routinely collected vital-sign data, in Chapter 6 we described how we modelled other heterogeneous information, such as results of laboratory tests and diagnosis code assigned at admission. The main theme of the chapter is information fusion using representation learning and transfer learning, since we pre-trained the sub-components of the model and used an auto-encoder to learn a general representation of the patient's vital-sign data. We then jointly optimised the learning of the representation and the classification of deterioration.

7.1.1 Limitations

This thesis has several limitations. The first limitation is that all of the proposed systems were developed using the outcome prediction framework, where the outcome

was defined as the composite outcome of unplanned ICU admission, cardiac arrest or mortality. Using EWS systems in the context of outcome prediction models can blur the original discourse that introduced EWS systems with the priority of securing timely medical response [183], rather than predicting a particular outcome. While securing timely medical response remained as the overarching priority, the predictive framework allowed us to classify the data samples to either the deteriorating class or non-deteriorating class in order to develop models using supervised learning. While we chose those outcomes because they were available in our datasets and are commonly used in related studies, they are not the only outcomes that define a 'deteriorating patient'. This becomes especially important when the same EWS systems are also used to alert for other life-threatening conditions, such as sepsis. We also did not perform multi-class classification and simply aggregated the three outcomes due to class imbalance.

Another limitation of the current framework is the strict timeframe cut-off; i.e., the model estimates whether or not an event occurs within the next N hours. Thus, if an observation is collected at a time $>N$ hours to an event, then the patient is assumed to be normal. This loses a lot of information and is an unrealistic assumption since patients may show signs of deterioration more than N hours in advance of an adverse event. An alternative framework would be the time-to-event (survival) analysis that instead measures the difference between the time of the observations and the time of the outcome [184].

The third limitation is overfitting, which is a common problem in machine learning research, including machine learning in healthcare [37]. Overfitting occurs when the model fits the training data too well and is unable to generalise to unseen data during testing. This is especially highlighted in Chapter 4, where AEWS only produced marginal improvement over NEWS (0.1% increase) on the external testing set ($\mathcal{D}_P$).

The deep learning models and the training strategies alleviated the problem better than the simple statistical model in Chapters 5 and 6. However, the performance gains for $\mathcal{D}_{O,2}$ were consistently higher than the gains for $\mathcal{D}_P$. Since the training dataset $\mathcal{D}_{O,1}$ and the testing set $\mathcal{D}_{O,2}$ originally belong to the same dataset, this discrepancy in performance gains across the testing sets implies that the model has overfit to training data from OUH. Overfitting in this thesis mostly occurred due to the limited size of the data, especially the minority class.

7.2 Recommendations for Early Warning Systems

To maximise the potential of EWS systems, several considerations should be accounted for. First, the performance of EWS systems heavily relies on data quality, as discussed in Chapter 3. Data quality is the core of predictive modelling in healthcare, as the performance of risk prediction systems heavily relies on the quality of the data items used as input variables. This in turn depends on staffing, nursing protocols, instruments used for data collection, or other possibly confounding factors. For example, cognition-related assessment is often subjectively performed by the attending medical staff, which may incur bias in the data as it may not reflect the true state of the patient. Other text-based entries may contain typographical errors and would thus require careful data cleansing. The availability of high quality data decreases the overhead of data pre-processing within AI systems. The data must thus be collected carefully to support the development of more accurate data-dependent predictive systems [129].

While the majority of current systems only assess the patient's vital signs, other variables may be better indicators of severity of illness. As summarised in our literature review (Chapter 2) and based on the results presented in Chapter 6, there are obvious benefits of incorporating other variables, such as laboratory tests, demographics,

or diagnoses history. This should be easy to implement since EWS systems are implemented digitally and can easily retrieve information from a patient's EHR.

Diagnostic codes are usually assigned for billing purposes when a patient is discharged, unlike physiological variables that are routinely collected during the hospital stay. Thus, the incorporation of diagnostic codes in a real-time alerting model must be carefully performed. We therefore encourage related studies to avoid the misuse of diagnosis codes to ensure the approach is feasible in real-time clinical practice.

Current existing systems tend to assign a binary label to each patient based on an optimised alerting threshold. Due to the high class imbalance across the positive and negative classes, this may lead to high false alarm rates (i.e., a low PPV). False alarm rates could then negatively affect staff compliance, disturb patients, or lead to unnecessary interventions. One solution that would allow more flexibility is to optimise a different alerting threshold for different patient sub-groups. Additionally, the hospital-wide alerting system should consider a patient ranking system based on the assigned probability score or normalised EWS instead of a binary label. This could help in prioritising patients that require the most urgent help.

Most of the related literature performs retrospective evaluation of recently proposed EWS systems. Only a few studies investigated the impact of EWS system implementation on health outcomes and resource utilisation. Such studies of the latter variety have had mixed results on mortality, admission to ICU, cardiac arrest, and length of hospital stay [185]. To better evaluate EWS systems, high-quality multi-center clinical trials must be designed in the near future.

Today, deep learning presents the opportunity to learn complex spatiotemporal relationships across a variety of data variables, as shown in this thesis. Modern risk prediction systems should take advantage of deep neural networks to maximise information retrieval from the different data modalities present within an EHR.

The performance of risk predictions systems is also generally assessed across the overall patient cohort, yet their performance may not be consistent across different patient sub-groups as shown across our results. The distributions of the characteristics of an in-hospital patient cohort, such as age or sex, may vary from one hospital to another. This in turn affects the transferability of risk prediction systems, as they may not generalise well across unseen patient populations. We therefore highly encourage data sharing and collaborative modelling efforts within the research community to allow for robust multicentre validation of risk prediction scores across heterogeneous patient cohorts.

A key concern with current EWS studies is the variable quality of derivation, validation and reporting [186]. Researchers should be encouraged to follow Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis TRIPOD guidance [130] and the future TRIPOD ML when it becomes available [187].

In summary, the success of a risk prediction system is highly dependent on the quality of its data streams, its ability to learn from multi-modal data, its generalisability across heterogeneous patient cohorts, and its ease of deployment. EWS systems currently used in practice, such as the NEWS, do not reflect the recent developments of related AI technologies. With rigorous multicentre validation, the rise in AI presents unprecedented opportunities to secure timely response for deteriorating patients and improve patient outcomes.

7.3 Pathway to Translation

The AI risk prediction models presented in this thesis present an opportunity to bring hospital alerting systems to the next level, since they outperform the EWS systems that are currently used in practice in terms AUROC and sensitivity. However, the performance of the models was only assessed through retrospective analysis and the

success of the models depends on several factors. The first factor is the performance of the models in real clinical settings. This entails evaluating the models in a long-term prospective study, which would require funding support. Key performance indicators will include false positive rate, accuracy, and time to compute a prediction, compared to the performance of the state-of-the-art. Other clinical metrics should also be assessed to highlight public value, such as patient outcomes.

Another factor is the ease and cost of deploying the models in practice. The proposed models can be easily incorporated in existing in-hospital devices, such as bedside monitors or hand-held devices, since they use the same data streams as existing scores. They also do not have large storage requirements, since the most complex deep learning architecture (i.e., iFEWS) only has 45k parameters, compared to millions of parameters in other domains such as computer vision and natural language processing. The scores can also be deployed as 'scoring engines' in back-end software modules to process the data. However, the engine would require maintenance, since the score may need to be validated from time to time and fine-tuned using recently collected data to account for dataset shift.

The translation of our models into clinical practice also relies on how the results of the engine are presented to the end-user (i.e., clinicians and nurses). The user-interface must be user-friendly and interpretable and must not cause disruption to existing clinical workflows. This requires extensive interface development with a focus on human factors and a controlled trial to evaluate the system's usability. Successful prospective studies will support the translation of clinically validated AI into practice, including our presented work which presents high potential clinical impact.

One challenge that may hinder translation into practice is the threat of liability, which may arise when patient injury occurs due to an incorrect prediction of the system. Within AI in healthcare, liability is complex since multiple parties may be

held liable, including the hospital, clinician, and the provider. However, our models aim to be decision support tools that clinicians can refer to rather than strictly follow as a standard of care.

7.4 Future Work

The work in this thesis presents many ideas for future research. We focus on two main themes that revolve around the use of machine learning to solve important clinical problems based on the available datasets.

7.4.1 Learning Optimal Treatment Strategies

Reinforcement Learning (RL) is a machine learning tool that learns an optimal set of rules to maximise an expected return. In comparison to supervised learning models that output a prediction given an input, RL makes a set of sequential decisions to achieve a particular task [188]. It has been very successful in gaming, such as for learning how to play chess, Atari video games, or Go.

RL has been recently used within different clinical contexts mainly to optimise treatment strategies. This includes learning the optimal dose of intravenous fluids and vasopressors for septic ICU patients [189], recommending a personalised regime of ventilator support and sedation dosage [190], or identifying the best drug regimens for epilepsy patients [191]. Each decision within the learned strategies aims to achieve the best patient outcomes.

In the context of our work, we might first aim to model the effect of medication on the vital-sign trajectory, since medication data is available in our datasets. In particular, as an example, we could learn effective therapy using vasopressors. This will involve the definition of a proper reward function and a careful design of a feasible

RL framework in collaboration with our clinical partners. While doing so, it will also be important to consider confounding factors and the effective sample size [192]. While RL is expected to improve existing clinical practice rather than suggest new treatments, it is an exciting field that could largely impact clinical decision-making.

7.4.2 Interpretability, Fairness, and Privacy

Another future research direction could consider ethics as a central concern in the development of our models. In this thesis, we considered two frameworks of interpretability: rule-based systems and attention-based deep learning models. While there are conflicting schools of thoughts concerning which framework is more interpretable, this could only be validated by testing both systems in a real clinical setting. Another area of future research is to consider both inherently interpretable models and post-hoc interpretations, since post-hoc interpretations on their own can potentially be misleading [116].

One important ethical dimension is fairness of machine learning models. This concerns developing models that promote health equity and address existing biases, from model design and data collection to model evaluation [193]. Within the context of patient monitoring, we could first evaluate the performance of our models for minority groups based on age or ethnicity. We might aim to improve our existing models and design new models that perform equally for all patients.

Finally, developing machine learning models relies on handling large-scale clinical data that is sensitive in nature. The current practice is to organise the data, such as EHR, in a centralised server for further research activity. To ensure data privacy, several machine learning approaches have been suggested in recent years. Federated learning offers a decentralised approach where worker nodes contribute to the training of a global model using their own local datasets [194]. Along with Secure Multiparty

Computation and Differential Privacy [195], our future work will investigate such frameworks to train deep learning models while maintaining data privacy.

In summary, we have contributed to applications of machine learning in the medical domain to improve the detection of clinical deterioration. Our work highlights the potential of machine learning in learning complex patterns from large-scale medical data and lays the foundations for future work.

References

[1] Timothy J Hodgetts, Gary Kenward, Ioannis Vlackonikolis, Susan Payne, Nicolas Castle, Robert Crouch, Neil Ineson, and Loua Shaikj. Incidence, location and reasons for avoidable in-hospital cardiac arrest in a district general hospital. *Resucitation*, 54:115–123, 2002.

[2] Ken M. Hillman, P. J. Bristow, T. Chey, K. Daffurn, T. Jacques, S. L. Norman, G. F. Bishop, and G. Simmons. Antecedents to hospital deaths. *Internal Medicine Journal*, 31(6):343–348, 2001.

[3] Office for National Statistics. Avoidable mortality in the UK: 2017. Technical report, Office for National Statistics, 2019.

[4] Louise S. Van Galen, Patricia W. Struik, Babiche E.J.M. Driesen, Hanneke Merten, Jeroen Ludikhuize, Johannes I. Van Der Spoel, Mark H.H. Kramer, and Prabath W.B. Nanayakkara. Delayed recognition of deterioration of patients in general wards is mostly caused by human related monitoring failures: A root cause analysis of unplanned ICU admissions. *PLoS ONE*, 2016.

[5] Christopher M Bishop. *Pattern Recognition and Machine Learning*, volume 4. Springer, 2006.

[6] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future, 2017.

[7] Paul Sajda. Machine Learning for Detection and Diagnosis of Disease. *Annual Review of Biomedical Engineering*, 8:537–65, 2006.

[8] Bryce Goodman and Seth Flaxman. EU regulations on algorithmic decision-making and a "right to explanation". In *ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, 2016.

[9] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019.

[10] Lucian L Leape, Troyen A Brennan, Nan Laird, Ann G Lawthers, Russel Localio, Benjamin A Barnes, Leisi Herbert, Joseph P Newhouse, Paul C Weiler, and Howard Hiatt. The Nature of Adverse Events in Hospitalized Patients. *The New England Journal of Medicine*, 324(6):377–384, 1991.

[11] Daryl Jones, Imogen Mitchell, Ken Hillman, and David Story. Defining clinical deterioration. *Resuscitation*, 84(8):1029–1034, 2013.

[12] Matthew M. Churpek, Trevor C. Yuen, and Dana P. Edelson. Predicting clinical deterioration in the hospital: The impact of outcome selection. *Resuscitation*, 84(5):564–568, 2013.

[13] Ross McL Wilson, William B Runciman, Robert W Gibberd, Bernadette T Harrison, Liza Newby, and John D Hamilton. The Quality in Australian Health Care Study. *The Medical Journal of Australia*, 163:458–471, 1995.

[14] Graham Neale, Maria Woloshynowych, and Charles Vincent. Exploring the causes of adverse events in NHS hospital practice. *Journal of the Royal Society of Medicine*, 94(7):322–30, 2001.

[15] Roland M H Schein, Nelson Hazday, Maria Pena, Bradley H. Ruben, and Charles L. Sprung. Clinical antecedents to in-hospital cardiopulmonary arrest. *Chest*, 98(6):1388–1392, 1990.

[16] Ken M. Hillman, Peter J. Bristow, Tien Chey, Kathy Daffurn, Theresa Jacques, Sandra L. Norman, Gillian F. Bishop, and Grant Simmons. Duration of life-threatening antecedents prior to intensive care admission. *Intensive Care Medicine*, 28(11):1629–1634, 2002.

[17] Juliane Kause, Gary Smith, David Prytherch, Michael Parr, Arthas Flabouris, and Ken Hillman. A comparison of Antecedents to Cardiac Arrests, Deaths and Emergency Intensive care Admissions in Australia and New Zealand, and the United Kingdom - The ACADEMIA study. *Resuscitation*, 62(3):275–282, 2004.

[18] Susanna E Bedell, David C Deitz, David Leeman, and Thomas L Delbanco. Incidence and characteristics of preventable iatrogenic cardiac arrests. *Journal of the American Medical Association*, 265:2815–2820, 1991.

[19] RJM Morgan, F Williams, and MM Wright. An early warning scoring system for detecting developing critical illness. *Clinical Intensive Care*, 8:100, 1997.

[20] Royal College of Physicians. National Early Warning Score (NEWS) - Standardising the assessment of acute-illness severity in the NHS. Technical report, Royal College of Physicians, London, 2012.

[21] Gary B. Smith, David R. Prytherch, Paul E. Schmidt, and Peter I. Featherstone. Review and performance evaluation of aggregate weighted 'track and trigger' systems, 2008.

[22] Gary B. Smith, David R. Prytherch, Paul Schmidt, Peter I. Featherstone, Debbie Knight, Gill Clements, and Mohammed A. Mohammed. Hospital-wide physiological surveillance - A new approach to the early identification and management of the sick patient. *Resuscitation*, 2006.

[23] Royal College of Physicians. National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS. Technical report, 2017.

[24] India Wheeler, Charlotte Price, Alice Sitch, Peter Banda, John Kellett, Mulinda Nyirenda, and Jamie Rylance. Early Warning Scores Generated in Developed Healthcare Settings Are Not Sufficient at Predicting Early Mortality in Blantyre, Malawi: A Prospective Cohort Study. *PLoS ONE*, 2013.

[25] David R. Prytherch, Gary B. Smith, Paul E. Schmidt, and Peter I. Featherstone. ViEWS-Towards a national early warning score for detecting adult inpatient deterioration. *Resuscitation*, 81(8):932–937, 2010.

[26] Gary B. Smith, David R. Prytherch, Paul Meredith, Paul E. Schmidt, and Peter I. Featherstone. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation*, 84(4):465–470, 2013.

[27] Lionel Tarassenko, David A. Clifton, Michael R. Pinsky, Marilyn T. Hravnak, John R. Woods, and Peter J. Watkinson. Centile-based early warning scores derived from statistical distributions of vital signs. *Resuscitation*, 82(8):1013–1018, 2011.

[28] Peter J. Watkinson, Marco A.F. Pimentel, David A. Clifton, and Lionel Tarassenko. Manual centile-based early warning scores derived from statistical distributions of observational vital-sign data. *Resuscitation*, 129:55–60, 2018.

[29] Stuart W. Jarvis, Caroline Kovacs, Tessy Badriyah, Jim Briggs, Mohammed A. Mohammed, Paul Meredith, Paul E. Schmidt, Peter I. Featherstone, David R. Prytherch, and Gary B. Smith. Development and validation of a decision tree early warning score based on routine laboratory test results for the discrimination of hospital mortality in emergency medical admissions. *Resuscitation*, 84(11):1494–1499, 2013.

[30] Oliver C. Redfern, Marco A.F. Pimentel, David Prytherch, Paul Meredith, David A. Clifton, Lionel Tarassenko, Gary B. Smith, and Peter J. Watkinson. Predicting in-hospital mortality and unanticipated admissions to the intensive care unit using routinely collected blood tests and vital signs: Development and validation of a multivariable model. *Resuscitation*, 2018.

[31] Naveed Afzal, Vishnu Priya, Sunghwan Sohn, Hongfang Liu, Rajeev Chaudhry, Christopher G Scott, Iftikhar J Kullo, and Adelaide M Arruda-olson. Natural language processing of clinical notes for identification of critical limb ischemia. *International Journal of Medical Informatics*, 111:83–89, 2018.

[32] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C Nelson, Jessica L Mega, and Dale R Webster. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *Journal of the American Medical Association*, 316(22):2402–2410, 2016.

[33] Nenad Tomašev, Xavier Glorot, Jack W. Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk,

Alistair Connell, Cían O. Hughes, Alan Karthikesalingam, Julien Cornebise, Hugh Montgomery, Geraint Rees, Chris Laing, Clifton R. Baker, Kelly Peterson, Ruth Reeves, Demis Hassabis, Dominic King, Mustafa Suleyman, Trevor Back, Christopher Nielson, Joseph R. Ledsam, and Shakir Mohamed. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, 2019.

[34] Hyojeong Lee, Soo-Yong Shin, Myeongsook Seo, Gi-Byoung Nam, and Segyeong Joo. Prediction of Ventricular Tachycardia One Hour before Occurrence Using Artificial Neural Networks. *Scientific Reports*, 6, 2016.

[35] M Aczon, D Ledbetter, L Ho, A Gunny, A Flynn, J Williams, and R Wetzel. Dynamic Mortality Risk Predictions in Pediatric Critical Care Using Recurrent Neural Networks. *arXiv*, pages 1–18, 2017.

[36] Scott B. Hu, Deborah J L Wong, Aditi Correa, Ning Li, and Jane C. Deng. Prediction of clinical deterioration in hospitalized adult patients with hematologic malignancies using a neural network model. *PLoS ONE*, 11(8):1–12, 2016.

[37] Daniele Ravi, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-perez, Benny Lo, and Guang-zhong Yang. Deep Learning for Health Informatics. *Ieee Journal of Biomedical and Health Informatics*, 21(1):4–21, 2017.

[38] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics*, pages 1–16, 2017.

[39] Hanqing Cao, Larry Eshelman, Nicolas Chbat, Larry Nielsen, Brian Gross, and Mohammed Saeed. Predicting ICU hemodynamic instability using continuous multiparameter trends. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society.*, pages 3803–3806, 2008.

[40] Joon Lee and Roger G Mark. An investigation of patterns in hemodynamic data indicative of impending hypotension in intensive care. *BioMedical Engineering OnLine*, 9(1):62, 2010.

[41] Rob Donald, Tim Howells, Ian Piper, I. Chambers, G. Citerio, P. Enblad, B. Gregson, K. Kiening, J. Mattern, P. Nilsson, A. Ragauskas, Juan Sahuquillo, R. Sinnot, and A. Stell. Early Warning of EUSIG-Defined Hypotensive Events Using a Bayesian Artificial Neural Network Article. *Acta Neurochirurgica Supplementum*, 114:87–91, 2012.

[42] Marcus Eng Hock Ong, Christina Hui Lee Ng, Ken Goh, Nan Liu, Zhi Xiong Koh, Nur Shahidah, Tong Tong Zhang, Stephanie Fook-Chong, and Zhiping Lin. Prediction of cardiac arrest in critically ill patients presenting to the emergency department using a machine learning score incorporating heart rate variability compared with the modified early warning score. *Critical Care*, 16(3):R108, 2012.

[43] Lei Clifton, David A. Clifton, Marco A.F. Pimentel, Peter J. Watkinson, and Lionel Tarassenko. Gaussian Processes for Personalized e-Health Monitoring With Wearable Sensors. *IEEE Transactions on Biomedical Engineering*, 60(1):193–197, 2013.

[44] Thomas A. Lasko, Joshua C. Denny, and Mia A. Levy. Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data. *PLoS ONE*, 8(6), 2013.

[45] Shima Ghassempour, Federico Girosi, and Anthony Maeder. Clustering multivariate time series using Hidden Markov Models. *International Journal of Environmental Research and Public Health*, 11(3):2741–2763, 2014.

[46] Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. Unfolding Physiological State: Mortality Modelling in Intensive Care Units. *Bone*, 23(1):1–7, 2014.

[47] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In *Machine Learning for Healthcare Conference*, pages 301–318, 2015.

[48] Curtis E Kennedy, Noriaki Aoki, Michele Mariscalco, and James P Turley. Using Time Series Analysis to Predict Cardiac Arrest in a PICU. *Pediatric Critical Care Medicine*, 16(9):332–9, 2015.

[49] Marzyeh Ghassemi, Tristan Naumann, Thomas Brennan, David a Clifton, and Peter Szolovits. A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 446–453, 2015.

[50] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. Risk Prediction with Electronic Health Records: A Deep Learning Approach. In *Proceedings of the 2016 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2016.*, pages 432–440, 2016.

[51] Riccardo Miotto, Li Li, Brian A. Kidd, and Joel T. Dudley. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific reports*, 6:26094, 2016.

[52] Jonathan Rubin, Cristhian Potes, Minnan Xu-Wilson, Junzi Dong, Asif Rahman, Hiep Nguyen, and David Moromisato. An ensemble boosting model for predicting transfer to the pediatric intensive care unit. *International Journal of Medical Informatics*, 112:15–20, 2018.

[53] Huan Song, Deepta Rajan, Jayaraman J. Thiagarajan, and Andreas Spanias. Attend and Diagnose: Clinical Time Series Analysis using Attention Models. *arXiv*, 2017.

[54] Joseph Futoma, Sanjay Hariharan, and Katherine Heller. Learning to Detect Sepsis with a Multitask Gaussian Process RNN Classifier. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[55] Edward Choi, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2):361–370, 2017.

[56] Alvin Rajkomar and Others. Scalable and accurate deep learning for electronic health records. *Nature Digital Medicine*, 1(1):1–10, 2018.

[57] Joon myoung Kwon, Youngnam Lee, Yeha Lee, Seungwoo Lee, Hyunho Park, and Jinsik Park. Validation of deep-learning-based triage and acuity score using a large national dataset. *PLoS ONE*, 2018.

[58] Sebastian Ruder. An overview of gradient descent optimization algorithms. 2016.

[59] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[60] Jimeng Sun, Fei Wang, Jianying Hu, and Shahram Edabollahi. Clustering Overly-Specific Features in Electronic Medical Records. *ACM SIGKDD Explorations Newsletter*, 14(1):16, 2012.

[61] G. Skolidis, R. H. Clayton, and G. Sanguinetti. Automatic Classification of Arrhythmic Beats Using Gaussian Processes. *Computers in Cardiology*, 35:921–924, 2008.

[62] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering - A decade review. *Information Systems*, 53(October 2016):16–38, 2015.

[63] John A. Quinn, Christopher K.I. Williams, and Neil McIntosh. Factorial switching linear dynamical systems applied to physiological condition monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1537–1551, 2009.

[64] Ioan Stanculescu, Christopher K I Williams, and Yvonne Freer. A Hierarchical Switching Linear Dynamical System Applied to the Detection of Sepsis in Neonatal Condition Monitoring. *UAI'14 Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 752–761, 2014.

[65] Li Wei H. Lehman, Shamim Nemati, Ryan P. Adams, and Roger G. Mark. Discovering shared dynamics in physiological signals: Application to patient monitoring in ICU. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 5939–5942, 2012.

[66] Zachary C. Lipton, John Berkowitz, and Charles Elkan. A Critical Review of Recurrent Neural Networks for Sequence Learning. 2015.

[67] Rasmussen and Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[68] Marco A.F. Pimentel, David A. Clifton, and Lionel Tarassenko. Gaussian process clustering for the functional characterisation of vital-sign trajectories. In *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, 2013.

[69] Glen Wright Colopy, Stephen J. Roberts, and David A. Clifton. Gaussian Processes for Personalized Interpretable Volatility Metrics in the Step-Down Ward. *IEEE Journal of Biomedical and Health Informatics*, 2019.

[70] Robert Dürichen, Marco A F Pimentel, Lei Clifton, Achim Schweikard, and David A. Clifton. Multitask Gaussian processes for multivariate physiological time-series analysis. *IEEE Transactions on Biomedical Engineering*, 62(1):314–322, 2015.

[71] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Ali Eslami, and Yee Whye Teh. Neural Processes. In *ICML Workshop on Theoretical Foundations and Applications of Deep Generative Models*, 2018.

[72] T Jayalakshmi and A. Santhakumaran. Statistical Normalization and Back Propagationfor Classification. *International Journal of Computer Theory and Engineering*, 3(1):89–93, 2011.

[73] Patrick Schwab, Gaetano Scebba, Jia Zhang, Marco Delai, and Walter Karlen. Beat by Beat: Classifying Cardiac Arrhythmias with Recurrent Neural Networks. *arXiv*, 2017.

[74] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical Intervention Prediction and Understanding using Deep Networks. *arXiv*, pages 1–16, 2017.

[75] Narges Razavian, Jake Marcus, and David Sontag. Multi-task Prediction of Disease Onsets from Longitudinal Lab Tests. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, pages 1–27, 2016.

[76] Youngduck Choi, Chill Yi-i Chiu Ms, and David Sontag. Learning Low-Dimensional Representations of Medical Concepts. *AMIA Joint Summits on Translational Science proceedings*, pages 41–50, 2016.

[77] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer Representation Learning for Medical Concepts. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pages 1495–1504, 2016.

[78] Hugo Larochelle, Yoshua Bengio, Jérôme Louradour, and Pascal Lamblin. Exploring strategies for training deep neural networks. *Journal of Machine Learning Research*, 2009.

[79] George E. Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 2012.

[80] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Computing Surveys*, 2019.

[81] Lindsay I Smith. A tutorial on Principal Components Analysis. Technical report, 2002.

[82] Hayden Wimmer and Loreen Powell. Principle Component Analysis for Feature Reduction and Data Preprocessing in Data Science. In *Proceedings of the Conference on Information Systems Applied Research*, pages 1–6, 2016.

[83] Songhee Cheon, Jungyoon Kim, and Jihye Lim. The Use of Deep Learning to Predict Stroke Patient Mortality. *International journal of environmental research and public health*, 16(11), 2019.

[84] Denis Krompaß, Cristóbal Esteban, Volker Tresp, Martin Sedlmayr, and Thomas Ganslandt. Exploiting Latent Embeddings of Nominal Clinical Data for Predicting Hospital Readmission. *KI - Künstliche Intelligenz*, 29(2):153–159, 2015.

[85] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 2000.

[86] Cristóbal Esteban, Oliver Staeck, Yinchong Yang, and Volker Tresp. Predicting Clinical Events by Combining Static and Dynamic Information Using Recurrent Neural Networks. In *IEEE International Conference on Healthcare Informatics (ICHI)*, pages 93–101, 2016.

[87] Magnus Sahlgren. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–53, 2008.

[88] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. *arXiv*, 2013.

[89] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. GRAM: Graph-based Attention Model for Healthcare Representation Learning. *arXiv*, pages 1–15, 2016.

[90] Cristóbal Esteban, Danilo Schmidt, Denis Krompaß, and Volker Tresp. Predicting sequences of clinical events by using a personalized temporal latent embedding model. *Proceedings - 2015 IEEE International Conference on Healthcare Informatics, ICHI 2015*, pages 130–139, 2015.

[91] Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In *NIPS Proceedings*, 2016.

[92] Aaron Courville Ian Goodfellow, Yoshua Bengio. Deep Learning Book. *Deep Learning*, 2015.

[93] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre Antoine Manzagol. Stacked denoising autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*, 2010.

[94] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Distilling Knowledge from Deep Networks with Applications to Healthcare Domain. 2015.

[95] Carl Doersch. Tutorial on Variational Autoencoders. 2016.

[96] Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. Deepr: A Convolutional Net for Medical Records. *IEEE Journal of Biomedical and Health Informatics*, 21(1):22–30, 2017.

[97] A. J. Scott, D. W. Hosmer, and S. Lemeshow. Applied Logistic Regression. *Biometrics*, 2006.

[98] Evangelia Christodoulou, Jie Ma, Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, and Ben Van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110:12–22, 2019.

[99] Elsa Loekito, James Bailey, Rinaldo Bellomo, Graeme K. Hart, Colin Hegarty, Peter Davey, Christopher Bain, David Pilcher, and Hans Schneider. Common laboratory tests predict imminent death in ward patients. *Resuscitation*, 84(3):280–285, 2013.

[100] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.

[101] Christopher J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 1998.

[102] Anneleen Daemen, Dirk Timmerman, Thierry Van den Bosch, Cecilia Bottomley, Emma Kirk, Caroline Van Holsbeke, Lil Valentin, Tom Bourne, and Bart De Moor. Improved modeling of clinical data with kernel methods. *Artificial Intelligence in Medicine*, 54(2):103–114, 2012.

[103] Yukun Chen, Robert J Carroll, Eugenia R McPeek Hinz, Anushi Shah, Anne E Eyler, Joshua C Denny, and Hua Xu. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *Journal of the American Medical Informatics Association*, 20:e253–e259, 2013.

[104] Sepp Hochreiter and J Urgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[105] Zachary C. Lipton, David C. Kale, Charles Elkan, and Randall Wetzel. Learning to Diagnose with LSTM Recurrent Neural Networks. In *Proceedings of ICLR 2016*, pages 1–18, 2015.

[106] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

[107] Tingting Zhu, Glen Wright Colopy, Clare MacEwen, Katherine Niehaus, Yang Yang, Chris W. Pugh, and David A. Clifton. Patient-Specific Physiological Monitoring and Prediction Using Structured Gaussian Processes. *IEEE Access*, 7:58094–58103, 2019.

[108] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 2007.

[109] Marco A.F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

[110] Muhammad Aurangzeb Ahmad, Ankur Teredesai, and Carly Eckert. Interpretable machine learning in healthcare. In *Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018*, 2018.

[111] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[112] Gary B. Smith, David R. Prytherch, Paul E. Schmidt, and Peter I. Featherstone. Review and performance evaluation of aggregate weighted 'track and trigger' systems. *Resuscitation*, 77(2):170–179, 2008.

[113] You Jin Kim, Yun-Geun Lee, Jeong Whun Kim, Jin Joo Park, Borim Ryu, and Jung-Woo Ha. High Risk Prediction from Electronic Medical Records via Deep Attention Networks. In *NIPS Proceedings*, 2017.

[114] Marko Hoikka, Tom Silfvast, and Tero I. Ala-Kokko. Does the prehospital National Early Warning Score predict the short-term mortality of unselected emergency patients? *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 2018.

[115] Santiago Romero-Brufau, Jeanne M. Huddleston, Gabriel J. Escobar, and Mark Liebow. Why the C-statistic is not informative to evaluate early warning scores and what metrics to use. *Critical Care*, 19(1):19–24, 2015.

[116] Zachary C. Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10):35–43, 2018.

[117] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. 2017.

[118] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Interpretable Deep Models for ICU Outcome Prediction. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2016:371–380, 2016.

[119] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[120] Shane Nanayakkara, Sam Fogarty, Michael Tremeer, Kelvin Ross, Brent Richards, Christoph Bergmeir, Sheng Xu, Dion Stub, Karen Smith, Mark Tacey, Danny Liew, David Pilcher, and David M. Kaye. Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study. *PLoS Medicine*, 15(11):1–16, 2018.

[121] Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–158, 2012.

[122] George Hripcsak and David J Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–21, 2013.

[123] Jon D Patrick, Dung H M Nguyen, Yefeng Wang, and Min Li. A knowledge discovery and reuse pipeline for information extraction in clinical notes. *Journal of the American Medical Informatics Association*, 18(5):574–579, 2011.

[124] William R Hogan and Michael M Wagner. Accuracy of data in computer-based patient records. *Journal of the American Medical Informatics Association*, 4(5):342–355, 1997.

[125] E. M. Mirkes, T. J. Coats, J. Levesley, and A. N. Gorban. Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes. *Computers in Biology and Medicine*, 75:203–216, 2016.

[126] Nicole Gray Weiskopf and Chunhua Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20:144–151, 2012.

[127] NHS. NHS Data Model and Dictionary, 2019.

[128] NHS Trust Oxford University Hospitals. Biochemistry Reference Ranges Document. Technical report, 2019.

[129] David Wong, Timothy Bonnici, Julia Knight, Lauren Morgan, Paul Coombes, and Peter Watkinson. SEND: a system for electronic notification and documentation of vital sign observations. *BMC medical informatics and decision making*, 15:68, 2015.

[130] Karel G.M. Moons, Douglas G. Altman, Johannes B. Reitsma, John P.A. Ioannidis, Petra Macaskill, Ewout W. Steyenberg, Andrew J. Vickers, David Ransohoff, and Gary S. Collins. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Disagnosis (TRIPOD): Explanantion and Elaboration. *Annals of Internal Medicine*, 162(1):W1–W74, 2015.

[131] Charlson Index Online Calculator.

[132] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

[133] Matthew M Churpek, Trevor C Yuen, Christopher Winslow, Jesse Hall, and Dana P Edelson. Differences in vital signs between elderly and nonelderly patients prior to ward cardiac arrest. *Critical care medicine*, 43(4):816–22, 2015.

[134] Malcolm Elliott and Alysia Coventry. Critical care: the eight vital signs of patient monitoring. *British Journal of Nursing*, 2012.

[135] John F. Fieselmann, Michael S. Hendryx, Charles M. Helms, and Douglas S. Wakefield. Respiratory rate predicts cardiopulmonary arrest for internal medicine inpatients. *Journal of General Internal Medicine*, 1993.

[136] David R. Goldhill and A. F. McNarry. Physiological abnormalities in early warning scores are related to mortality in adult inpatients. *British Journal of Anaesthesia*, 2004.

[137] Daniel T. Lackland and Michael A. Weber. Global Burden of Cardiovascular Disease and Stroke: Hypertension at the Core. *Canadian Journal of Cardiology*, 2015.

[138] Frank H. Wians. Clinical Laboratory Tests: Which, Why, and What Do The Results Mean? *Laboratory Medicine*, 2009.

[139] David Tanne, Noa Molshatzki, Oleg Merzeliak, Rakefet Tsabari, Maya Toashi, and Yvonne Schwammenthal. Anemia status, hemoglobin concentration and outcome after acute stroke: A cohort study. *BMC Neurology*, 2010.

[140] J Z Ma, J Ebben, H Xia, and A J Collins. Hematocrit level and associated mortality in hemodialysis patients. *Journal of the American Society of Nephrology*, 1999.

[141] Göran Nilsson, Pär Hedberg, and John Öhrvik. White Blood Cell Count in Elderly Is Clinically Useful in Predicting Long-Term Survival. *Journal of Aging Research*, 2014.

[142] Arun Chawla, Richard H. Sterns, Sagar U. Nigwekar, and Joseph D. Cappuccio. Mortality and serum sodium: Do patients die from or with hyponatremia? *Clinical Journal of the American Society of Nephrology*, 2011.

[143] Yan Chen, Yingying Sang, Shoshana H. Ballew, Adrienne Tin, Alex R. Chang, Kunihiro Matsushita, Josef Coresh, Kamyar Kalantar-Zadeh, Miklos Z. Molnar, and Morgan E. Grams. Race, Serum Potassium, and Associations With ESRD and Mortality. *American Journal of Kidney Diseases*, 2017.

[144] Jesse Adams and Fred Apple. New Blood Tests for Detecting Heart Disease. *Circulation*, 2004.

[145] Chris Higgins. Urea and the clinical value of measuring blood urea concentration. *Acutecaretesting.Org*, 2016.

[146] Glenn M. Chertow, Elisabeth Burdick, Melissa Honour, Joseph V. Bonventre, and David W. Bates. Acute Kidney Injury, Mortality, Length of Stay, and Costs in Hospitalized Patients. *Journal of the American Society of Nephrology*, 2005.

[147] Alicia M. Alvarez and Debabrata Mukherjee. Liver abnormalities in cardiac diseases and heart failure. *International Journal of Angiology*, 2011.

[148] Scott M. Wells and Meg Sleeper. Cardiac troponins. *Journal of Veterinary Emergency and Critical Care*, 18(3):235–245, 2008.

[149] Noël J.M. Cano, Claude Pichard, Hubert Roth, Isabelle Court-Fortuné, Luc Cynober, Michèle Gérard-Boncompain, Antoine Cuvelier, Jean Pierre Laaban, Jean Claude Melchior, Jean Claude Raphaël, and Christophe M. Pison. C-reactive protein and body mass index predict outcome in end-stage respiratory failure. *Chest*, 2004.

[150] Don D. Sin and S. F. Paul Man. Why are patients with chronic obstructive pulmonary disease at increased risk of cardiovascular diseases? The potential role of systemic inflammation in chronic obstructive pulmonary disease. *Circulation*, 2003.

[151] E G Lakatta. Age-associated cardiovascular changes in health: impact on cardiovascular disease in older persons. *Heart Failure Reviews*, 7(1):29–49, 2002.

[152] Jennifer Gonik Chester and James L. Rudolph. Vital Signs in Older Patients: Age-Related Changes, 2011.

[153] S S Franklin, W Gustin, N D Wong, M G Larson, M A Weber, W B Kannel, and D Levy. Hemodynamic patterns of age-related changes in blood pressure. The Framingham Heart Study. *Circulation*, 96(1):308–15, 1997.

[154] Maria Grazia Modena. Hypertension in postmenopausal women: How to approach hypertension in menopause. *High Blood Pressure and Cardiovascular Prevention*, 21(3):201–204, 2014.

[155] Helga Gudmundsdottir, Aud Høieggen, Aud Stenehjem, Bård Waldum, and Ingrid Os. Hypertension in women: latest findings and clinical implications. *Therapeutic advances in chronic disease*, 3(3):137–46, 2012.

[156] Elisabete Pinto. Blood pressure and ageing. *Postgraduate medical journal*, 83(976):109–114, 2007.

[157] J.P. Janssens, J.C. Pache, and L.P. Nicod. Physiological changes in respiratory function associated with ageing. *European Respiratory Journal*, 13(1):197–205, 1999.

[158] Gulshan Sharma and James Goodwin. Effect of aging on respiratory system physiology and immunology. *Clinical interventions in aging*, 1(3):253–260, 2006.

[159] Satoshi Yamamoto, Shumpei Watanabe, Izumi Fujiwara, Norimichi Tsumura, Tetsuo Akiba, and Keiko Ogawa-Ochiai. Measuring oxygen saturation of skin with advancing age using iterative optical path-length matrix method. *Artificial Life and Robotics*, 18(3-4):165–171, 2013.

[160] C. P J Quarterman, Antony N. Thomas, M. McKenna, and R. McNamee. Use of a patient information system to audit the introduction of modified early warning scoring. *Journal of Evaluation in Clinical Practice*, 11(2):133–138, 2005.

[161] J. Kellett and B. Deane. The Simple Clinical Score predicts mortality for 30 days after admission to an acute medical unit. *Qjm*, 99(11):771–781, 2006.

[162] Thomas Olsson, Andreas Terent, and Lars Lind. Charlson Comorbidity Index can add prognostic information to Rapid Emergency Medicine Score as a predictor of long-term mortality. *European Journal of Emergency Medicine*, 12:220–224, 2005.

[163] Alan Monaghan. Detecting and managing deterioration in children. *Paediatric nursing*, 17(1):32–35, 2005.

[164] Douglas C. Montgomery and Runger C. George. *Applied Statistics and Probability for Engineers*. John Wiley & Sons, 6 edition, 2014.

[165] Matthew M. Churpek, Trevor C. Yuen, Christopher Winslow, David O. Meltzer, Michael W. Kattan, and Dana P. Edelson. Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. *Crit*, 44(2):368–374, 2016.

[166] Jos van der Westhuizen and Joan Lasenby. Bayesian LSTMs in medicine. In *NeurIPS*, 2018.

[167] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2007.

[168] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *NIPS Proceedings*, 2017.

[169] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the International Conference on Learning Representations*, 2015.

[170] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

[171] Supreeth P. Shashikumar, Amit J. Shah, Gari D. Clifford, and Shamim Nemati. Detection of Paroxysmal Atrial Fibrillation using Attention-based Bidirectional Recurrent Neural Networks. *arXiv*, 2018.

[172] M. Schuster and K. K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[173] Qiong Wei and Roland L. Dunbrack. The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. *PLoS ONE*, 2013.

[174] GPyOpt. A Bayesian Optimization framework in python, 2016.

[175] François Chollet and others. Keras, 2015.

[176] Abadi Martin and Others. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015.

[177] Matthew M. Churpek, Trevor C. Yuen, Christopher Winslow, Ari A. Robicsek, David O. Meltzer, Robert D. Gibbons, and Dana P. Edelson. Multicenter development and validation of a risk stratification tool for ward patients. *American Journal of Respiratory and Critical Care Medicine*, 2014.

[178] ICD-10-CM/PCS Medical Coding Reference.

[179] Sara Khalid, David A. Clifton, Lei Clifton, and Lionel Tarassenko. A two-class approach to the detection of physiological deterioration in patient vital signs, with clinical label refinement. *IEEE Transactions on Information Technology in Biomedicine*, 16(6):1231–1238, 2012.

[180] Luchen Liu, Haoran Li, Zhiting Hu, Haoran Shi, Zichang Wang, Jian Tang, and Ming Zhang. Learning Hierarchical Representations of Electronic Health Records for Clinical Outcome Prediction. 2019.

[181] Luchen Liu, Jianhao Shen, Ming Zhang, Zichang Wang, and Jian Tang. Learning the Joint Representation of Heterogeneous Temporal Events for Clinical Endpoint Prediction. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[182] Gary Smith. Improving the detection and response to patient deterioration, 2011.

[183] RJM Morgan and MM Wright. In defence of early warning scores. *British Journal of Anaesthesia*, 99(5):747–748, 2007.

[184] Douglas G. Altman and JM Bland. Time to event (survival) data. *British Medical Journal*, 317:468–469, 1998.

[185] M. E.Beth Smith, Joseph C. Chiovaro, Maya O'Neil, Devan Kansagara, Ana R. Quiñones, Michele Freeman, Makalapua L. Motu'apuaka, and Christopher G. Slatore. Early warning system scores for clinical deterioration in hospitalized patients: A systematic review. *Annals of the American Thoracic Society*, 11(9):1454–1465, 2014.

[186] Stephen Gerry, Jacqueline Birks, Timothy Bonnici, Peter J. Watkinson, Shona Kirtley, and Gary S. Collins. Early warning scores for detecting deterioration in adult hospital patients: A systematic review protocol, 2017.

[187] Gary S. Collins and Karel G.M. Moons. Reporting of artificial intelligence prediction models, 2019.

[188] Ann Nowè and Tim Brys. A gentle introduction to reinforcement learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016.

[189] Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018.

[190] Niranjani Prasad, Corey Chivers, Michael Draugelis, and Barbara E Engelhardt. A Reinforcement Learning Approach to Weaning of Mechanical Ventilation in Intensive Care Units. 2013.

[191] Orrin Devinsky, Cynthia Dilley, Michal Ozery-Flato, Ranit Aharonov, Ya'ara Goldschmidt, Michal Rosen-Zvi, Chris Clark, and Patty Fritz. Changing the approach to treatment choice in epilepsy using big data. *Epilepsy and Behavior*, 56:32–37, 2016.

[192] Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, and David Sontag. The promises and pitfalls of reinforcement learning in healthcare. 2019.

[193] Alvin Rajkomar, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12):866–872, 2018.

[194] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, 54, 2017.

[195] Theo Ryffel, Andrew Trask, Morten Dahl, Bobby Wagner, Jason Mancuso, Daniel Rueckert, and Jonathan Passerat-Palmbach. A generic framework for privacy preserving deep learning. 2018.

Appendix A

We include here supplementary information regarding the analysis of the datasets ($\mathcal{D}_{O,1}$, $\mathcal{D}_{O,2}$, and $\mathcal{D}_P$) retrieved from the HAVEN database. Table A.1 summarises the plausible ranges of the vital signs and laboratory tests considered during data pre-processing. Fig. A.1 shows the probability density function of all vital-sign variables for the three datasets. Fig. A.2 shows the probability density function of all the laboratory-test measurements for the three datasets. Table A.2 shows how the ICD-10 codes were grouped based on the ICD-10 dictionary [178].

Table A.1: Summary of vital signs abbreviations, units and plausible ranges used for inclusion criteria.

Variable	Abbreviation	Units	Plausible Ranges
Vital Signs			
Heart Rate	HR	beats/min	15 - 250
Respiratory Rate	RR	breaths/min	3 - 50
Systolic Blood Pressure	SBP	mm Hg	30 - 300
Temperature	TEMP	°C	28 - 45
Oxygen Saturation	SPO_2	%	40 - 100
AVPU score	AVPU	NA	1 - 4
Oxygen Support	O_2S	NA	0/1
Laboratory Tests			
Hemoglobin	HGB	g/L	NA
Sodium	Na	mmol/L	100 - 200
White Blood Count	WBC	10^9/L	NA
Potassium	K	mmol/L	1 - 15
Albumin	ALB	g/L	10-70
Creatinine	CR	μmol/L	8.8 - 2210
Urea	UR	mmol/L	0.4-107.1
Total Bilirubin	TBIL	μmol/L	NA
Hematocrit	HCT	L/L (%)	3-76
Troponin	TROP	μg/L	> 0
C-reactive Protein	CRP	mg/L	> 0

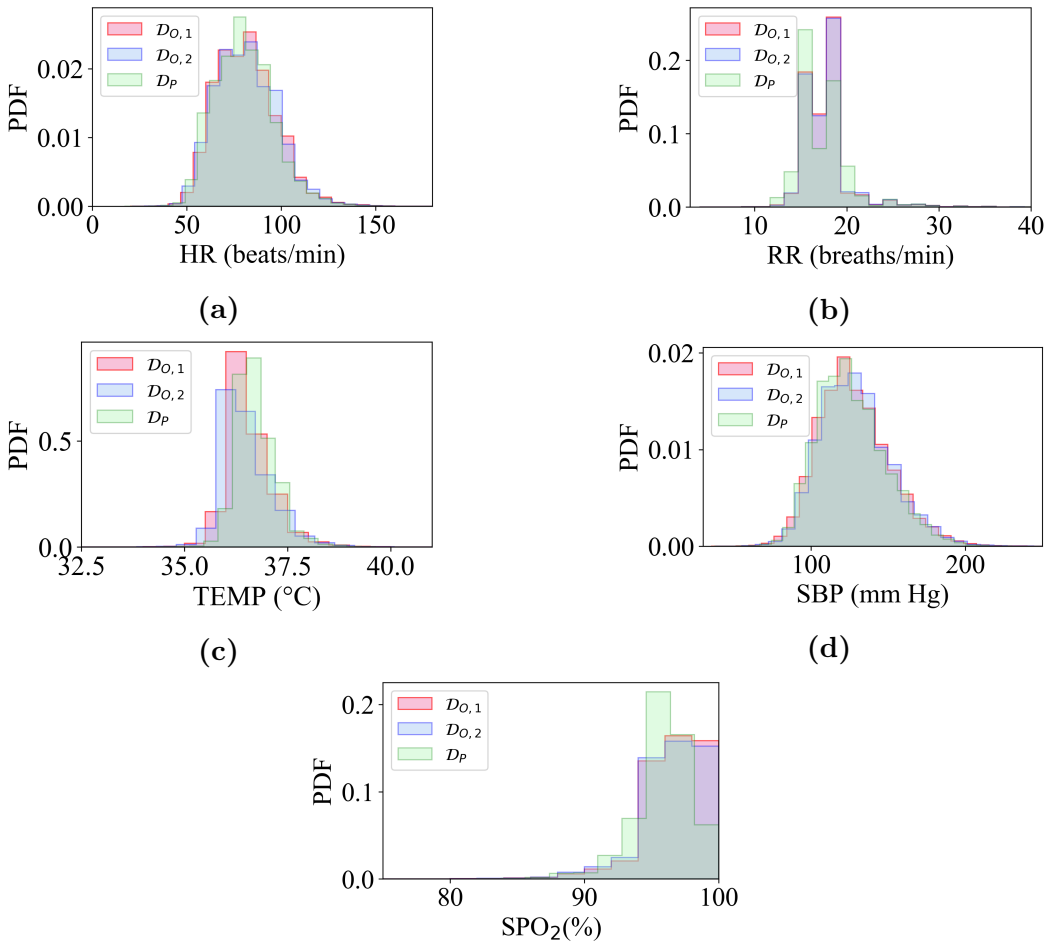


Fig. A.1: Probability Density Function (PDF) to visualise the distributions of the continuous vital signs for the $\mathcal{D}_{O,1}$, $\mathcal{D}_{O,2}$, and $\mathcal{D}_P$ datasets, indicated by the red, blue, and green colours respectively. The two-sided vital signs are (a) Heart rate (HR), (b) Respiratory Rate (RR), (c) Temperature (TEMP), and (d) Systolic Blood Pressure (SBP). The single-sided distribution is (e) Oxygen Saturation (SPO_2).

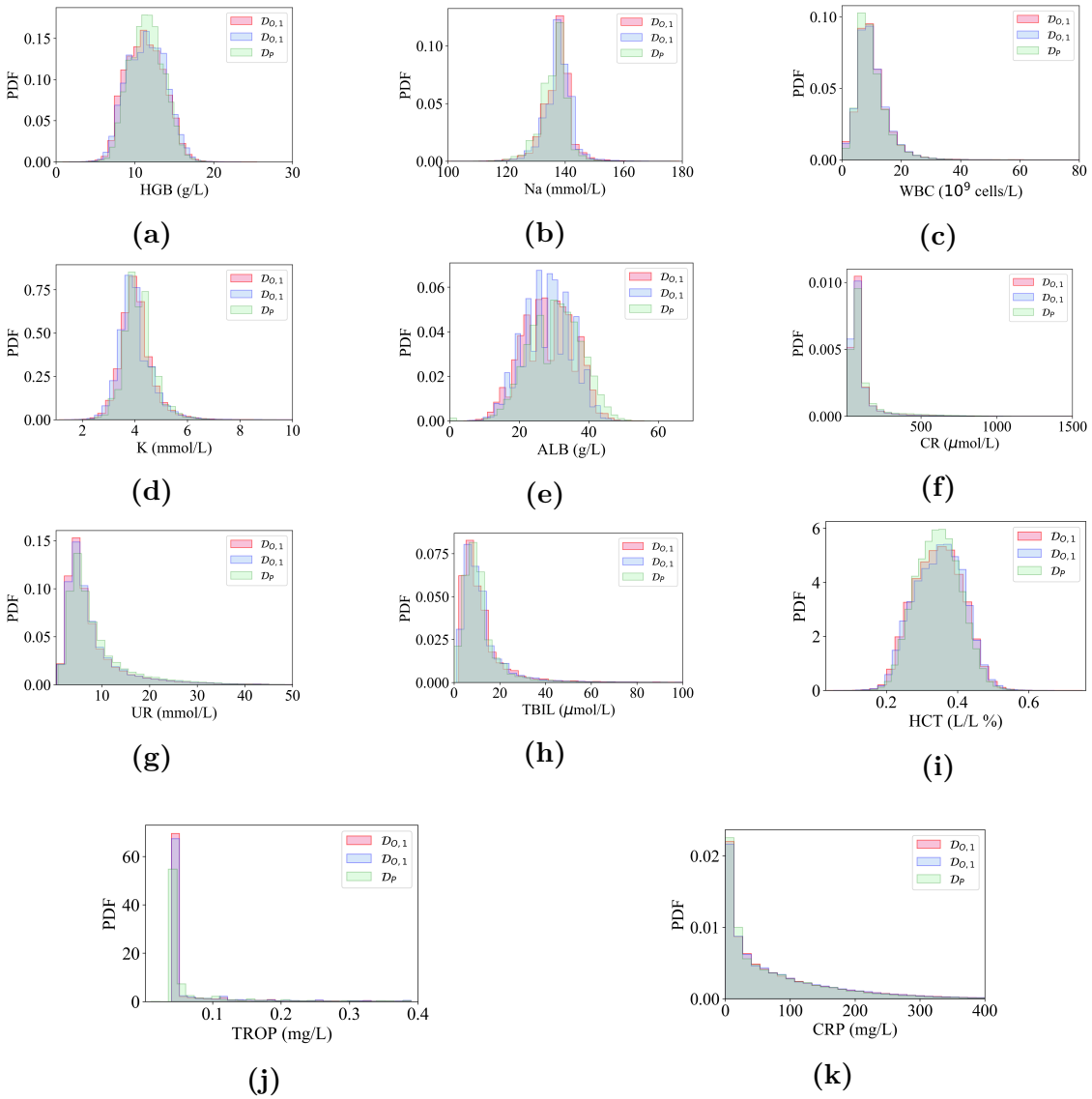


Fig. A.2: Probability Density Function (PDF) to visualise the distributions of the laboratory tests for the $\mathcal{D}_{O,1}$, $\mathcal{D}_{O,2}$, and $\mathcal{D}_P$ datasets indicated by the red, blue, and green colours respectively.

Table A.2: Reference of the ICD-10 group numbers and their description.

ICD-10 Group	Description
1	Certain infectious and parasitic diseases
2	Neoplasms
3	Diseases of the blood and the immune system
4	Endocrine, nutritional and metabolic diseases
5	Mental, behavioural, and neurodevelopmental disorders
6	Diseases of the nervous system
7	Diseases of the eye
8	Diseases of the ear
9	Diseases of the circulatory system
10	Diseases of the respiratory system
11	Diseases of the digestive system
12	Diseases of the skin and subcutaneous tissue
13	Diseases of the musculoskeletal system and connective tissue
14	Diseases of the genitourinary system
15	Pregnancy and childbirth
16	Conditions in the perinatal period
17	Congenital malformations, deformations and chromosomal abnormalities
18	Symptoms, signs and abnormal clinical and laboratory findings
19	Injury, poisoning and other external causes
20	External causes of morbidity
21	Factors influencing health status and contact with health services

Appendix B

In this appendix, we include the alerting thresholds of existing EWS systems. Table B.1 shows the alerting thresholds of NEWS. Table B.2 shows the alerting thresholds of MCEWS, and Table B.3 shows the scores assigned to thresholds based on centile cut-offs of the vital-sign distributions.

Table B.1: Alerting thresholds of NEWS.

Variable	3	2	1	0	1	2	3
HR	≤40		41-50	51-90	91-110	111-130	≥131
RR	≤8		9-11	12-20		21-24	≥25
TEMP	≤35.0		35.1-36.0	36.1-38.0	38.1-39.0	≥39.1	
SBP	≤90	91-100	101-110	111-219			≥220
SPO_2	≤91	92-93	94-95	≥96			
AVPU				A			V,P,U
Supplemental Oxygen				Not provided		Provided	

Table B.2: Alerting thresholds of MCEWS.

Variable	3	2	1	0	1	2	3
HR	≤42	43-49	50-53	54-104	105-112	113-127	≥128
RR	≤7	8-11	11-12	13-21	22-23	24-28	≥ 29
TEMP	≤35.4		35.5-35.9	36.0-37.3	37.4-38.3		≥38.4
SBP	≤83	84-90	91-100	101-157	158-167	168-184	≥185
SPO_2	≤84	85-90	91-93	≥94			
AVPU				A	V		P,U
Supplemental Oxygen				Not provided		Provided	

Table B.3: Scores assigned to measured vital sign values in MCEWS based on alerting thresholds obtained from centiles of cdf, computed from either a symmetrical or skewed distribution.

Centile Range (%)	Score
Symmetrical Distribution	
<1	3
1-5	2
5-10	1
10-90	0
90-95	1
95-99	2
>99	3
Skewed Distribution	
<80	0
80-90	1
90-98	2
>98	3

Appendix C

In this appendix, we include any additional results for the technical chapters.

Additional Results for Chapter 5

Table C.1: Summary of the complexity of the models used as baselines in Chapter 5 in comparison to the proposed model DEWS. The complexity is represented by the total number of trainable parameters.

Model	No. of trainable parameters
LR	85
LSTM	973
BiLSTM	1,933
BiLSTM-ATT-1	627
BiLSTM-ATT-2	10,466
UA-BiLSTM-ATT-1	4,001
DEWS	17,986

Additional Results for Chapter 6

Table C.2: Summary of the complexity of the models in Chapter 6. The complexity is represented by the total number of trainable parameters.

Model	No. of trainable parameters
$\text{LR}_{L,V,D}$	
iFEWS_{SC-AE}	50,805
iFEWS_{MC-AE}	86,289
$\text{iFEWS}_{MC-AE-ATT}$	51,779