DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE

ICT International Doctoral School

Machine learning applications in Intensive Care Unit

Seyedmostafa Sheikhalishahi

Advisor

Dr. Venet Osmani

Fondazione Bruno Kessler

January 2022

Abstract

The rapid digitalization of the healthcare domain in recent years highlighted the need for advanced predictive methods particularly based upon deep learning methods. Deep learning methods which are capable of dealing with time-series data have recently emerged in various fields such as natural language processing, machine translation, and the Intensive Care Unit (ICU). The recent applications of deep learning in ICU have increasingly received attention, and it has shown promising results for different clinical tasks; however, there is still a need for the benchmark models as far as a handful of public datasets are available in ICU. In this thesis, a novel benchmark model of four clinical tasks on a multi-center publicly available dataset is presented; we employed deep learning models to predict clinical studies. We believe this benchmark model can facilitate and accelerate the research in ICU by allowing other researchers to build on top of it. Moreover, we investigated the effectiveness of the proposed method to predict the risk of delirium in the varying observation and prediction windows, the variable ranking is provided to ease the implementation of a screening tool for helping caregivers at the bedside. Ultimately, an attention-based interpretable neural network is proposed to predict the outcome and rank the most influential variables in the model predictions' outcome. Our experimental

findings show the effectiveness of the proposed approaches in improving the application of deep learning models in daily ICU practice.

Keywords: [Intensive care unit, Machine learning, Interpretable deep learning, Mortality prediction, Delirium prediction, Phenotyping, Decompensation]

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The recent advances in informatics systems have enabled engineers to record clinical data digitally; thus, this digitalization led to the rapid growth of clinical data in the last decade. The vast amount of digitally recorded data allowed researchers to access clinical data such as electronic health records (EHRs), genomics data, and medical images. Specifically, critical care units are designed for patients with a life-threatening illness, those who have undergone a major surgical procedure, or the ones with trauma. Therefore requiring 24-hour monitoring for all the patients and, in some cases, intervention [1]. The critical care unit is equipped with many devices responsible for various tasks such as patient monitoring, respiratory and cardiac support, pain management, and other life support equipment for people with trauma. In this context, the continuous monitoring and care of ICU patients have resulted in enormous data, which leads to many opportunities and challenges [2].

The vast amount of unexplored information in EHRs and the existence of conventional machine learning methods has enabled opportunities for re-

searchers and clinicians to extract precious knowledge from EHRs, which is particularly relevant in the critical care setting [3] as decision-making needs to be precise and quick as the decision is associated with patient's life. In this regard, there are a handful number of publicly available critical care datasets such as Medical Information Mart for Intensive Care (MIMIC-III) [4], eICU Collaborative Research Database (eICU-CRD) [5], and The Amsterdam University Medical Centers Database (AmsterdamUMCdb) [6].

In this regard, varying machine learning techniques such as Logistic Regression (LR), Support Vector Machines (SVM), and Random Forest (RF) techniques have been employed to potentially identify patients at risk of acute or chronic diseases in critical care [7, 8]. Conventional machine learning (ML) techniques such as LR and RF are interpretable due to the linear nature of the models. However, the complexity of critical illness makes the conventional ML approaches to medical research insufficient [3]. The methods mentioned above poorly perform on high-dimensional data and time-series data; thus, they cannot provide accurate predictive performance [9, 10]. The massive amount of high-dimensional data and the time-series type of data require more advanced techniques (deep learning techniques) such as recurrent neural networks (RNNs) or Transformers, which can handle high dimensional time-series data. These deep learning techniques have outperformed conventional machine learning techniques in many domains which deal with time-series data, such as natural language processing, machine translation, and the clinical field. Therefore, although deep learning techniques outperform traditional machine learning

techniques, the lack of interpretability of deep learning techniques remains a considerable challenge, specifically in the healthcare domain which the decision-making by the model should be transparent to the clinicians.

1.1 Problem Statement

EHRs include structured and unstructured data; there are challenges and shortcomings in predictive models using structured or unstructured data. This thesis targets structured data to extract the information incorporated in EHRs and build predictive models.

There are some well-known and well-defined clinical tasks in ICU, such as mortality prediction, Length of Stay (LoS) prediction, Phenotyping, physiologic decompensation, and delirium prediction. However, the lack of publicly available benchmarks regarding the application of machine learning techniques on these tasks hinders this area's progression. In this regard, we provide a benchmark on applying machine learning techniques on varying clinical tasks to help researchers accelerating progress in this field by building on top of these studies.

As mentioned earlier, the interpretability of deep learning methods has received attention in the machine learning domain and specifically in the application of machine learning in the clinical domain due to the importance of the model transparency for clinicians and researchers [11]. Considering this, we provide an interpretable deep learning model in the ICU domain, which provides some degree of interpretability related to clinical variable ranking.

1.2 Motivations

As mentioned earlier, the critical care unit deals with human life, and therefore the decision needs to be taken quickly and accurately. Moreover, the vast amount of available data in the domain makes the decision-making procedure challenging for clinical staff. Therefore, machine learning algorithms have a great potential to assist clinicians and clinical staff in making the decision and intervention procedure.

The main objective of this thesis is to employ and advance machine learning research in the clinical domain, specifically in the critical care domain. Extensively, we exploit machine learning techniques to improve the quality of patient care in the critical care domain and optimize healthcare costs and resources, in other words, to shift from reactive to proactive care. This thesis addresses the prediction of five clinical tasks in ICU using real-world clinical data. In this regard, advanced machine learning models are adopted for all tasks. Additionally, an interpretable model is proposed to provide a degree of interpretability to assist intensivists.

1.3 Contributions

This thesis's main contributions are divided into two categories: i) clinical contributions, and ii) technical contributions. We present several examples of creating actionable clinical insights from machine learning.

- **Clinical contribution:**

  ICUs cost $81.7 billion in the US, accounting for 13.4% of hospital

costs and 4.1% of national health expenditures [12]. Between 2000 and 2005, the number of hospital beds in the United States shrank by 4.2%, but the number of critical care beds increased by 6.5%, with occupancy increasing by 4.5%. Resource allocation and identifying patients with unexpected extended ICU stays and high mortality risk would help decision-making systems to improve the quality of care and ICU resource allocation. Therefore forecasting the mortality and LoS in ICU would be significantly important to provide high-quality care to a patient. It would avoid extra costs for care providers. Similarly, early detection of acute or chronic diseases in the ICU would significantly impact the shift from reactive care to proactive care, which is desirable for patients, clinicians, and healthcare providers. With that in mind, we aim to facilitate the research to ease shifting from reactive care to proactive care.

- **Technical contribution:** This thesis adopted deep learning methods for several clinical tasks such as in-hospital mortality prediction, forecasting LoS in the ICU, patient phenotyping, physiologic decompensation, and delirium prediction. The specific technical contributions in this thesis with referenced publications are:

  - We have provided clinical models and conventional machine learning models as baselines, and the state-of-the-art deep learning models such as long-short term memory (LSTM) RNNs; to capture better the evolution of time and build a more robust data representation. At the time of writing this thesis, there are a

handful of publicly available datasets in ICU, such as MIMIC-III [4], eICU-CRD [5], and Amsterdam UMCdb [6] but there are only benchmarks on the MIMIC dataset. By providing the Benchmark on the eICU-CRD, we aim to improve the research in the ICU field. We aim to help researchers explore this field by employing more advanced methods, using more datasets for training and validation, and achieving the utmost advantages of the available datasets. This work has lead to the following publication:

- **Seyedmostafa Sheikhalishahi**, Vevake Balaraman, and Venet Osmani. Benchmarking machine learning models on multicentre eicu critical care dataset. *PloS one*, 15(7):e0235424, 2020

– Delirium occurrence is common and preventive strategies are resource intensive. Screening tools can prioritize patients at risk. Using machine learning we can capture time and treatment effects that pose a challenge to delirium prediction. We aim to develop a delirium prediction model that can be used as a screening tool, we have developed such a tool with varying observation and prediction windows on the eICU-CRD and MIMIC-III datasets. Moreover, the variable ranking is provided to fit the clinical needs as an effective screening tool. The delirium prediction task has resulted in the following publications:

- Anirban Bhattacharyya, **Seyedmostafa Sheikhalishahi**, Siddharth Dugar, Sudhir Krishnan, Abhijit Duggal, and Venet Osmani. 400: Predicting delirium risk for the following 24

hours in critically ill patients using deep learning. *Critical Care Medicine*, 48(1):182, 2020

- Anirban Bhattacharyya, **Seyedmostafa Sheikhalishahi** , Heather Torbic, Wesley Yeung, Tiffany Wang, Jennifer Birst, Abhijit Duggal, Leo Anthony Celi, and Venet Osmani. Delirium prediction in the icu - designing a screening tool for preventive interventions (under review). *Journal of the American Medical Informatics Association Open*

– The use of deep learning models has been questionable in the clinical domain regarding the transparency of these models. In that context, a new interpretable model is proposed for the clinical time-series data, which considers the time-dependencies and variable-dependencies to achieve a rich variable ranking. The proposed model aims to provide a certain degree of interpretability to clinicians. The proposed interpretable model is applied to delirium prediction, which provides the most influential variables for delirium prediction by considering both time and variable level importance.

- **Seyedmostafa Sheikhalishahi**, Anirban Bhattacharyya, Leo Anthony Celi, and Venet Osmani. An interpretable deep learning model for time-series electronic health records: Case study of delirium prediction in critical care (under review). *Artificial Intelligence In Medicine*

1.4 Structure of the Thesis

This thesis captures the evolution in machine learning applications in the critical care unit, starting with clinical methods and gradually shifting toward machine learning, deep learning methods, and interpretable deep learning methods. This evolution is organized into six chapters as follows:

- **Chapter 1** provides an introduction to the problem statement, motivations and lists the specific contributions discussed in this thesis.
- **Chapter 2** discusses the general machine learning applications in the clinical domain and specifically in the ICU domain. This Chapter introduces conventional machine learning and deep learning algorithms for ICU, such as Logistic Regression, Random Forest, and LSTM-RNN. The evaluation metrics which have been used in this study are provided as well. Eventually, an introduction to interpretability methods and different types of interpretable techniques is provided.
- **Chapter 3** presents recently developed benchmark model on the eICU-CRD. In this Chapter, we provide extracted dataset for four different clinical tasks. A comparison between clinical models, conventional machine learning models, and deep learning models is also provided.
- **Chapter 4** discusses delirium and its importance in the clinical domain, two datasets are being employed, namely eICU-CRD and MIMIC-III. Delirium prediction is performed using varying observation and prediction windows; this chapter presents the results com-

paring conventional machine learning techniques and deep learning techniques. Moreover, variable ranking is provided based on three post-hoc interpretable methods to show the most and least influential variables in predicting delirium.

- **Chapter 5** presents an interpretable approach that has been adapted for the clinical time-series data based on the self-attention mechanism. Moreover, the proposed approach considers the dependencies among varying time-step, varying variables, and between variables and time steps. Detailed comparisons and analyses are provided for the delirium prediction task.

- **Chapter 6** concludes by summarizing the contributions of the thesis, outlining the potential direction for future work.

Chapter 2

Background

In this chapter, we begin with introducing conventional machine learning algorithms, followed by a discussion of deep learning algorithms and the interpretability of deep learning algorithms. Besides, we explore the evaluation metrics to evaluate model performance. Moreover, machine learning applications in the clinical domain are discussed.

First, we introduce conventional machine learning models, namely, linear regression, logistic regression, and random forest. Then, we explore the simplest form of artificial neural network called feed-forward neural network. Afterward, we introduce more robust deep learning models for time-series modeling called *recurrent neural networks* (RNNs) and, accordingly, one of RNNs' variants, which is called *long-short term memory* (LSTM). Additionally, an introduction to the varying interpretable techniques is provided, in which we review two types of interpretable algorithms, namely, ad-hoc and post-hoc interpretable algorithms.

In the rest of this chapter, we focus on the application of machine learning algorithms in the clinical domain by describing the required procedures

to build and evaluate predictive models in the clinical field. Moreover, an introduction to evaluation metrics methods for clinically relevant tasks is provided at the end of this chapter.

2.1 Machine learning

Machine learning is a discipline of computer science in which a model learns from examples rather than rules. More formally, a computer program is supposed to learn from experience E with respect to task T and performance measure P when its performance measure P at task T improves with experience E [17].

In this context, machine learning algorithms based on the availability of data and training strategies are divided into three categories: *supervised learning*, *unsupervised learning*, and *semi-supervised learning*. *Supervised learning* is defined as given a set of input associated with a set of outcomes. Given the set of input data, the model learns to predict the set of outcomes. *Unsupervised learning* is a technique given a set of input data, and the algorithm learns the hidden patterns in the input data. *Semi-supervised learning* is considered a stage between supervised learning and unsupervised learning, in which there exists a small amount of input data associated with a set of outcomes and the remaining set of input data has no association with a known set of outcomes.

In addition to that, there have been rapid developments in machine learning methodology. The machine learning methods are categorized into *conventional machine learning* algorithms and *deep learning* algorithms.

Deep learning can model complex and non-linear effects and improve predictive performance. However, due to non-linearity, deep learning methods are considered as "black-boxes".

2.1.1 Linear Regression

Linear regression is an approach to model a linear relationship between a dependent variable (target) and one or more independent variables; in case of more than one independent variable, the method is called multiple linear regression, formally:

$$y = \sum_{i=0}^{n} \beta_i * x_i + \epsilon \tag{2.1}$$

x_0 is equal to 1, β_i are regression coefficients, and ϵ provides for random variation in the dependent variable (y) not explained by the independent variables (x). The main assumptions made on the training are that the outcome can be *discrete* or *continuous*. Linear regression is applied and ranked as one of the most applied models in diverse areas such as biology and sociology to describe potential relationships between variables.

2.1.2 Logistic Regression

Logistic regression [18] is a predictive analysis model that uses the logistic function to fit the models to data. Unlike linear regression, the outcome in logistic regression must be discrete. Logistic regression is used to explain the relationship between the dependent variable and the independent variables. It models the probability of an event occurring, depending on the values of independent variables, which can be numerical or categorical.

Logistic regression estimates the probability of an event occurring versus the probability of the event not occurring and predicts the effect of a series of variables. The logistic regression model formally is defined as,

$$\log \frac{p(x)}{1-p(x)} = \beta_0 + x \cdot \beta \tag{2.2}$$

Solving,

$$p(x) = \frac{e^{(\beta_0 + x\cdot\beta)}}{1 + e^{(\beta_0 + x\cdot\beta)}} = \frac{1}{1 + e^{-(\beta_0 + x\cdot\beta)}} \tag{2.3}$$

The function $\sigma(z) = \frac{1}{1+e^{-z}}$ is often referred to as sigmoid or logistic function which restricts the value of $(-\beta x)$ in the range [0,1]. The logistic regression gives a linear classifier decision boundary that separates the two classes. The class probabilities also depend on the distance from the boundary.

In the next section, we explore a more advanced method, namely random forest, which could deal with classification and regression tasks.

2.1.3 Random Forest

A decision tree uses a set of rules to make a specific decision. Decision trees extract knowledge from a large amount of data, while it has a simple and compact form that can efficiently classify new data. Each node in the decision tree is represented as one of the following three nodes: leaf nodes, decision nodes, and root nodes. Over-fitting is a significant practical difficulty for decision tree models and many other predictive models [19]. Random forest techniques are developed by Breiman [20] to address the over-fitting of the decision tree. The random forests model is built by an

ensemble of DTs grown from a randomized variant of the tree induction algorithm. Moreover, as it is outlined in Figure 2.1 the data is divided into subsets, and for each subset of data, a DT is applied, and each tree provides a classification then the random forest selects the classification having the most votes, and in case of regression, it takes the average of the outputs provided by different trees.

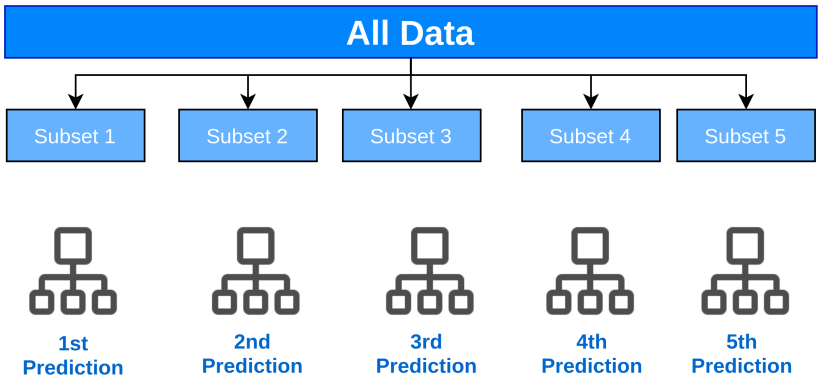


Figure 2.1: *Schematic of a random forest model*

Varying artificial neural networks, namely feed-forward neural networks and recurrent neural networks, are explained in the following sections.

2.2 Artificial Neural Networks

Artificial neural networks (ANNs) are inspired by the computational model of biological neural networks. ANNs are composed of perceptrons (neurons) and bias. Each neuron partitions the space into parts using linear function features. The perceptron receives multiple inputs depending on the weighted sum of inputs and the threshold; the output is 0 or 1. Bias is required to move threshold as needed by the loss function and helps in

improving the accuracy of the model [21].

In the past decade, there has been quick progress in ANNs, dealing with different data types such as image and text, aiming to capture the ideal representation from data to improve the predictive performance. Generally, ANNs are composed of many layers of neurons, in which each layer is connected to the next layer. In ANNs, The first layer is called the input layer, and the last layer is considered the output layer, and all the layers between input and output layers are considered hidden layers. Feed-forward neural networks (FNNs) are composed of multiple layers of neurons that are stacked together [22].

As mentioned in Chapter 1, patient vital signs are monitored and measured continuously in the critical care unit, shaping sequences of data. Thus, EHRs data can be seen as time-series sequences. FNNs can approximate the available non-linear relationships in the dataset, but FNNs cannot capture the temporal dependencies in time series data. As a result, Elman et al. [23] proposed recurrent neural networks (RNNs) to handle more effectively time-series data.

This section provides a more detailed introduction based on FNNs, RNNs, and self-attention mechanisms.

2.2.1 Feed-Forward Neural Networks

The feed-forward neural networks (FNNs) are composed of multiple layers, including multiple artificial neurons. It consists of an input layer and an output layer with one or more hidden layers connecting the input and output layers. Each artificial neuron is connected to the next layer of

artificial neurons, making the network fully connected. What makes ANNs family different from conventional machine learning models is a non-linear activation function applied to neurons. This activation function helps to create more complex decision boundaries where different data points are not linearly separable. As outlined in Figure 2.2, a multi-layer perceptron with a single hidden layer is discussed.

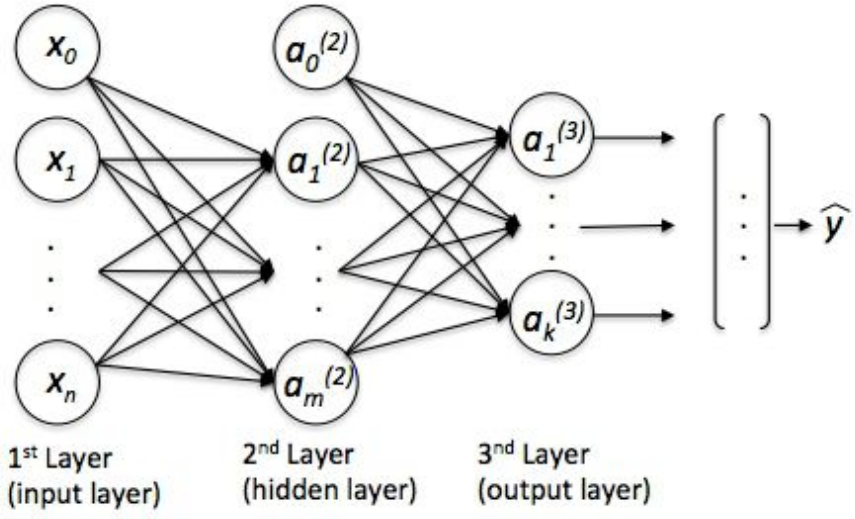


Figure 2.2: *Schematic of a simple feed-forward neural network*

All input of a neuron is multiplied by associated weights and summed up to a bias term. Then, a non-linear activation function is applied to the result before passing it to the next layer. An affine transformation followed by a non-linear function σ is applied on the input $\mathbf{x}$ which produce a hidden representation $\mathbf{h}$ in the hidden layer, formally:

$$\mathbf{h} = \sigma(\mathbf{x}\mathbf{W} + \mathbf{b}) \tag{2.4}$$

where $\mathbf{x} \in \mathbb{R}^{d_i}$, $\mathbf{W} \in \mathbb{R}^{d_i \times d_h}$, $\mathbf{b} \in \mathbb{R}^{d_h}$, $\mathbf{h} \in \mathbb{R}^{d_h}$, with d_i and d_h being

the dimension of the input and the hidden layers. The output layer then applies another affine transformation on the hidden representation followed by an activation function g:

$$\mathbf{o} = g(\mathbf{h}\mathbf{U} + \mathbf{c}) \tag{2.5}$$

where $\mathbf{U} \in \mathbb{R}^{d_h \times d_o}$, $\mathbf{c} \in \mathbb{R}^{d_h}$, $\mathbf{o} \in \mathbb{R}^{d_o}$, with d_o as the dimension of the output layer. The prediction $\hat{y}$ is compared with the original label, and the loss function is then applied to calculate the error, which helps in learning the network parameters such as $\mathbf{W}, \mathbf{U}, \mathbf{b}$, and $\mathbf{c}$. These parameters need to be learned and updated during the training phase until the best predictive performance is achieved.

2.2.2 Recurrent Neural Networks

The main drawback of FNNs for time-series data is the assumption of time independence between time steps. In other words, FNNs do not take into account the dependency between varying time steps that is available in data. Recurrent neural networks (RNNs) were proposed to deal with the temporal dependencies in the time-series data such as text, speech, and EHRs data.

Figure 2.3 outlines a simplified version of RNN followed by an unrolled schematic of a RNN. As it is shown in Figure 2.3 a simplified RNN conform the following stage to create the current hidden state representation $\mathbf{h_t}$:

1. Takes as input $\mathbf{x_t}$ at time t

2. Weights the input with $\mathbf{W}$, results in $\mathbf{W}\mathbf{x_t}$

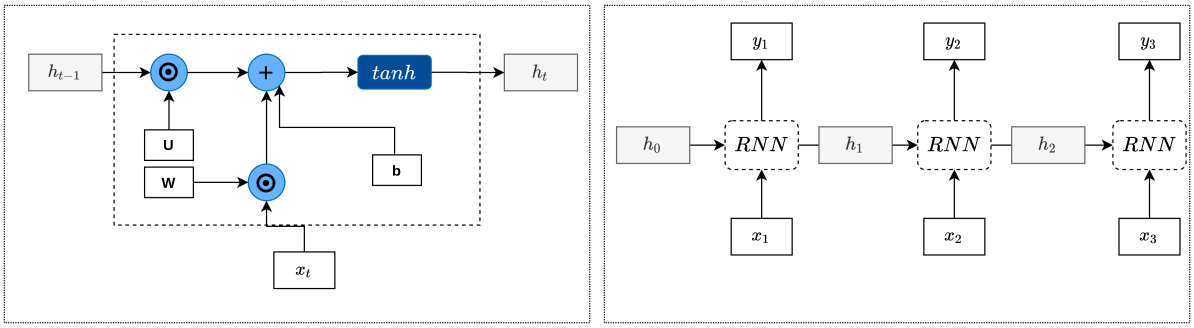


Figure 2.3: *A simplified version of RNN (left), an unrolled version of RNN (right)*

3. Weights the previous hidden state $\mathbf{h_{t-1}}$ with $\mathbf{U}$, formally $\mathbf{Uh_{t-1}}$

4. Sum up the results which are produced by (2) and (3), $\mathbf{Wx_t} + \mathbf{Uh_{t-1}}$

5. Apply non-linear activation function *tanh* on (4) to compute $\mathbf{h_t}$, formally:

$$\mathbf{h_t} = tanh(\mathbf{Wx_t} + \mathbf{Uh_{t-1}}) \tag{2.6}$$

The recurrent connection makes RNNs use information from the past and use it to predict the current time-step. However, RNNs are prone to *exploding gradient* and *vanishing gradient* problem. *Vanishing gradient* occurs in the long sequences tasks, in which RNNs are unable to relate the correlation between temporally distant events [24], *vanishing gradient* leads to long training times and poor predictive performance. In contrary to *vanishing gradient*, *exploding gradient* occurs when the computed gradients are exponentially growing, which results in large updates to model wights, and consequently, the weights can become very large, resulting in Not-a-Number value [25]. In this regard, in the next section, we discuss Long Short-Term Memory RNNs, which can handle *exploding gradient* and *vanishing gradient* resulting in higher predictive performance.

Long Short-Term Memory

Long Short Term Memory networks (LSTMs) are a particular type of RNNs which are designed to address *exploding gradient* and *vanishing gradient* problems in RNNs. Integrating the memory cell [26] in LSTM makes it a better candidate to deal with long-distance temporal dependencies, as the memory cell decides which information to forget [27] and which information to keep and write in the current memory cell.

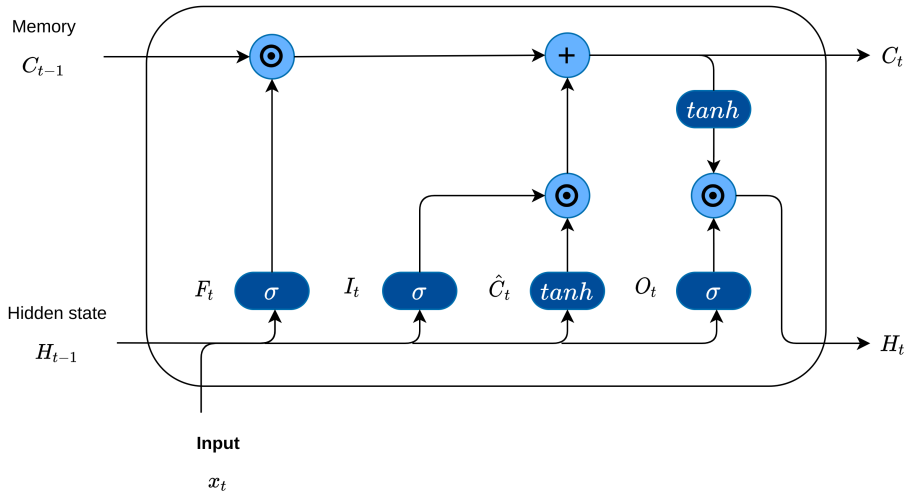


Figure 2.4: *LSTM cell*

As it is outlined in 2.4, an LSTM cell is composed by a forget gate $\mathbf{F_t}$, an input gate $\mathbf{I_t}$, a candidate memory $\hat{\mathbf{C}}_\mathbf{t}$, and an output gate $\mathbf{O_t}$. Forget gate decides which information to discard, input gate decides which values from the current input x_t. will be stored in the cell state, candidate memory decides which new values could be added to the state.

The input of the current time step and hidden state of the previous time step is fed into these gates, each gate composed by a fully connected layer

network and applies a sigmoid function to compute the values of forget gate $\mathbf{F_t}$, input gate $\mathbf{I_t}$ and output gate $\mathbf{O_t}$. Different from the other three gates $\hat{\mathbf{C}}_\mathbf{t}$ applies **tanh** as activation function, formally LSTM network computes the current cell content (C_t) and the current hidden state (H_t) as follows:

$$\begin{aligned}
\mathbf{F_t} &= \sigma(\mathbf{W_{xf}} \cdot \mathbf{x_t} + \mathbf{W_{hf}} \cdot \mathbf{H_{t-1}}) \\
\mathbf{I_t} &= \sigma(\mathbf{W_{xi}} \cdot \mathbf{x_t} + \mathbf{W_{hi}} \cdot \mathbf{H_{t-1}}) \\
\hat{\mathbf{C}}_\mathbf{t} &= \tanh(\mathbf{W_{xc}} \cdot \mathbf{x_t} + \mathbf{W_{hc}} \cdot \mathbf{H_{t-1}}) \\
\mathbf{O_t} &= \sigma(\mathbf{W_{xo}} \cdot \mathbf{x_t} + \mathbf{W_{ho}} \cdot \mathbf{H_{t-1}}) \\
\mathbf{C_t} &= \mathbf{F_t} \odot \mathbf{C_{t-1}} + \mathbf{I_t} \odot \hat{\mathbf{C}}_\mathbf{t} \\
\mathbf{H_t} &= \mathbf{O_t} \odot \tanh(\mathbf{C_t})
\end{aligned} \tag{2.7}$$

where $\mathbf{W_{xf}}, \mathbf{W_{xi}}, \mathbf{W_{xc}}$, and $\mathbf{W_{xo}}$ are the network weight parameters respectively related to forget gate, input gate, candidate memory, and output gate. $\mathbf{H_{t-1}}$ represents the previous hidden state, similarly $\mathbf{C_{t-1}}$ denotes previous cell content, and $\mathbf{x_t}$ is the input at current time step.

2.2.3 Self-attention Mechanisms

As it is demonstrated in Figure 5.4, self-attention is an attention mechanism that relates different positions of a single sequence in order to compute a representation of the sequence [28]. The self-attention mechanism has been employed successfully in a variety of tasks, including machine translation [29], abstractive summarization [30], and textual entailment [31]. Formally,

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.8}$$

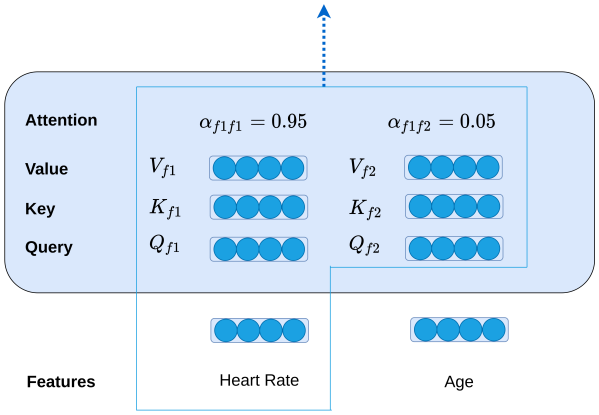


Figure 2.5: *The architecture of self-attention*

Where Q, K, V are computed by multiplying input with the learned matrices W_Q, W_k, W_V during training.
This thesis employs multi-head self-attention, which projects queries, keys, and values h times with different, learned linear projections. The scores are computed in parallel and are concatenated to get one matrix score, Formally:

$$Multihead(Q, K, V) = Concat(head_1, head_2)W^O \tag{2.9}$$

Where

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

and matrices such as W_i^Q, W_i^K, W_i^V, and W^O are the projections [32].

In the next section, we explore varying interpretable techniques used in the deep learning field, especially those employed in this thesis.

2.3 Interpretablity

As shown in Figure 2.6, even though deep learning techniques outperform the conventional machine learning techniques, interpretability of deep learning techniques remains one of the biggest challenges in the machine learning domain. In this context, interpretability and explainability concepts are often used interchangeably within the general Artificial Intelligence (AI) community.

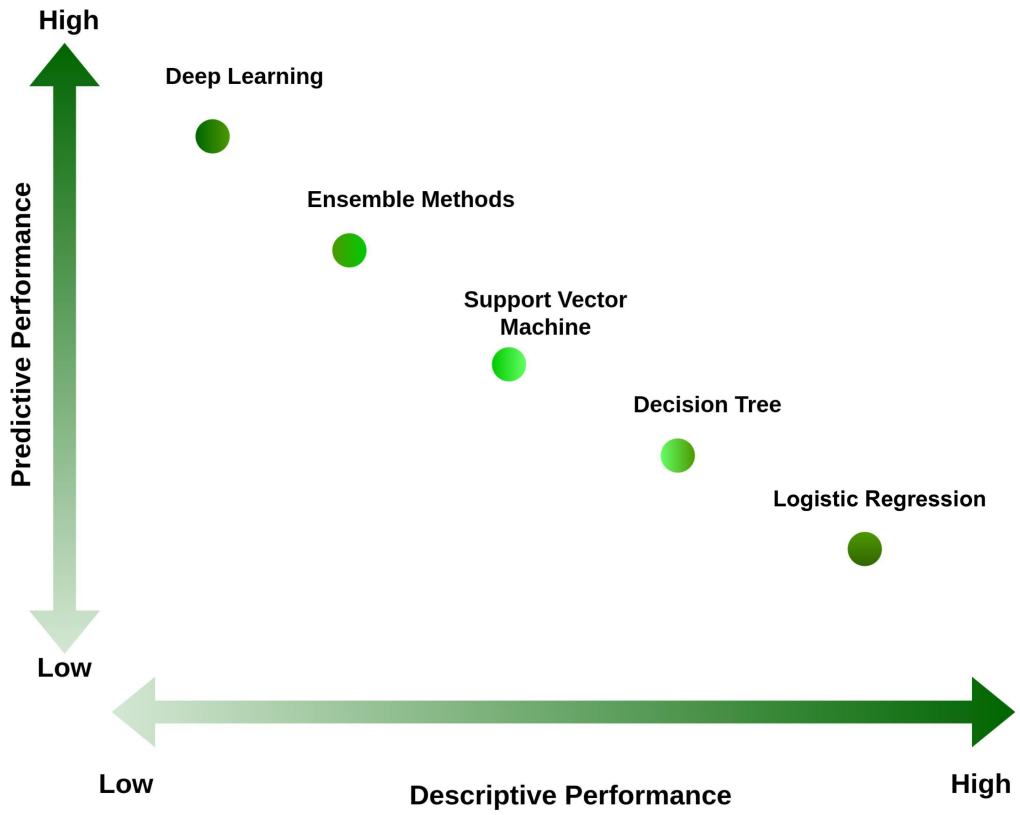


Figure 2.6: *Trade-off between Descriptive and Predictive Performance*

As outlined in Figure 2.7, interpretable methods can be categorized depending on the various criteria. Interpretable models fall in the two

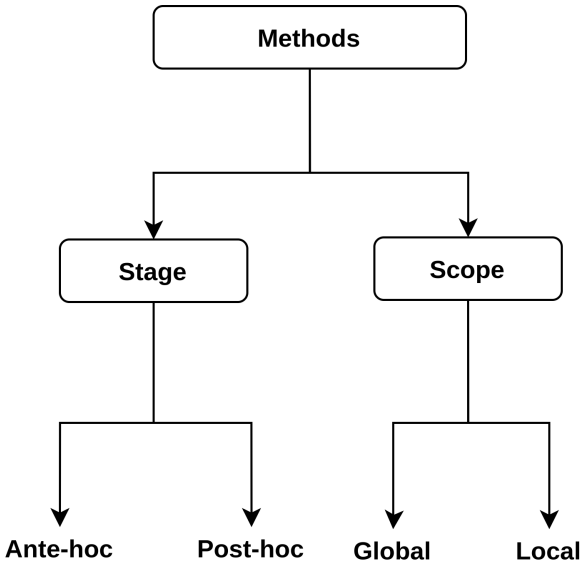


Figure 2.7: *Methods hierarchy for interpretability*

main categories in which each method can be described by using its *stage*, and *scope*. In the following sections, we explore the categories mentioned earlier while the main focus is on the *stage* categories, namely, Ante-hoc and Post-hoc stages.

2.3.1 Stage

The association of interpretable techniques concerning the predictive model is defined as the stage in which we characterize two main relations between a predictive model and an interpretable technique, namely ante-hoc and post-hoc techniques. In the following sections, the concepts and differences between ante-hoc and post-hoc approaches are provided.

Ante-hoc interpretable approach

Ante-hoc approaches incorporate interpretable modules into the predictive model from the beginning. As shown in Figure 2.9, ante-hoc approaches include a natural way of interpretation based on the incorporated interpretable module. Therefore, a unique model is employed for both predicting and interpreting.

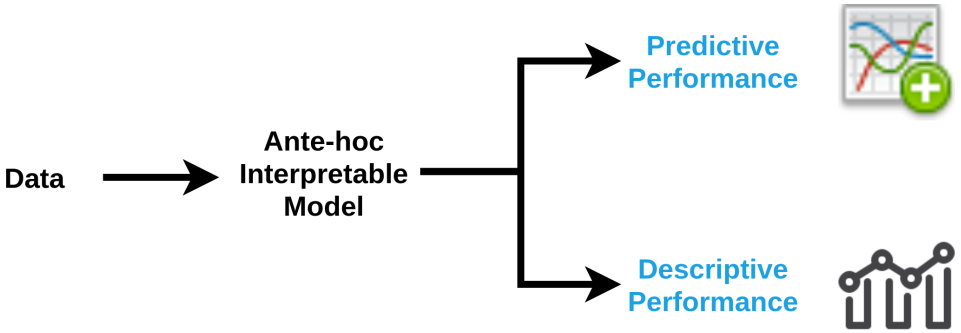


Figure 2.8: *Ante-hoc interpretable technique schema*

Ante-hoc approaches change the model's predictive performance, and it might lower the predictive performance as the interpretable module is incorporated with the predictive model but yields higher descriptive accuracy [33]. The main challenge of these approaches is to develop simple enough models to be easily understood by the user yet sophisticated enough to fit the underlying data properly. In the following, some of the ante-hoc interpretable approaches are introduced:

Ante-hoc approaches:

- *Attention-based approaches:* the Reversed Time Attention Model (RETAIN) is proposed by [34] to help doctors understand the deep learning predictions. The proposed architecture is composed of two RNNs

with an attention mechanism. The attention mechanism incorporated with RNNs helped explain which part the model was focusing on and which features influenced the model's choice.

- *Bayesian deep learning approaches: Bayesian network* is a graph composed of factors in interest and relationships. An expert can implement knowledge and especially dependency between factors as the graph structure. The relationships are expressed as simple probabilistic functions, although they can now be expressed by deep learning to obtain a deep generative model [35, 36]. We can mine the latent variables and obtain new knowledge on the target domain. The model is no longer a black-box but interpretable at a certain level [37].

Post-hoc interpretable approach

Post-hoc approaches allow models to be trained typically, with the interpretable module being incorporated at inference time. These approaches can be used on any machine learning model and usually work by analyzing feature input and output pairs. These interpretation methods take a trained model as input and extract information about what relationships the model has learned. post-hoc methods aim to keep a trained model unchanged and mimic or explain its behavior using an external module at testing time.

At this stage, the practitioner analyzes a trained model to provide insights into the learned relationships, mainly when the model's parameters do not clearly show what relationships the model has learned. Various post hoc interpretability methods have been developed to provide insight into

Figure 2.9: *Post-hoc interpretable technique schema*

what a trained model has learned without changing the underlying model. These methods are particularly crucial for settings where the collected data is high-dimensional and complex, such as image data. Once the information has been extracted from the fitted model, it can be analyzed using standards, exploratory data analysis techniques, such as scatter plots and histograms.

Since explanations from post-hoc models do not correspond to how the model predicts, there is skepticism regarding applying these models in scenarios that may require critical decision-making. Several post-hoc approaches have been developed recently; in the following, we further explore some of these approaches.

Post-hoc approaches:

- *Local Interpretable Model-Agnostic Explanations (LIME)*: LIME [38] attempts to understand the model by perturbing the input of data samples and understanding how the predictions change. It modifies a single data sample by tweaking the feature values and observing the output's resulting impact. Linear models are used to approximate local behavior. The linear model might not be powerful enough to explain the behavior of the original model. Non-linearity at local regions happens for those datasets that require complex, non-interpretable

models. Not being able to apply LIME in these scenarios is a significant pitfall.

- *SHapley Additive exPlanations (SHAP)*: SHAP [39] is a method to explain individual predictions. SHAP is based on the game theory. Shapley values indicate a fair distribution of predictions among the features. However, SHAP is computationally expensive and ignores feature dependence while computing the interpretability of a model.

- *Shapley Value Sampling (SVS)*: SVS is a perturbation-based method to compute variable attribution, which is based on sampling theory that can be used to estimate Shapley values [40]. The SVS produces variable ranking concerning each variable input and ranks the variables based on their predictive power.

- *Integrated Gradients (IG)*: IG [41] is a variation of computing the gradient on the prediction output concerning the input. IG computes the average gradient while the input varies along a linear path from a baseline $\tilde{x}$ to input x.

- *Guided Back-propagation (GB)*: GB [42] computes the gradient of the target output concerning the input. GB does not compute a true gradient but rather an imputed version of it.

Although post-hoc interpretability techniques are convenient, and in some ways, are ideal, they often rely on surrogate models or other approximations that can degrade the descriptive performance.

2.3.2 Scope

Interpretation techniques can be classified regarding the scope of interpretation, which refers to the portion of the prediction process they aim to explain. They can explain either a single prediction made by a model called a local scope or explain an entire model's behavior, which is called the global scope. The local scope is usually called the instance (sample) explanation, and the global scope is usually called the model explanation.

Global

In the global scope scenario, the whole system can be explained, and the logic can be followed from the input to every possible outcome. Additionally, global interpretations help us understand the inputs and their entire modeled relationship with the prediction target. This level of interpretability is about understanding how the model makes decisions based on a holistic view of its features and the learned components such as weights, parameters, and structures.

When a practitioner is interested in more general relationships learned by a model, e.g., relationships relevant to a particular class of responses or sub-population, they use global-level interpretations. At the global level, feature importance scores try to capture how much individual features contribute to a prediction across a dataset. These scores can provide insights into the features that the model has identified as important for outcomes and their relative importance. Varying methods have been developed to score individual features in many models including neural networks [43], random forests [44], and generic classifiers [45]. In addition to feature im-

portance, methods have been developed to extract important interactions between features. Interactions are important as ML models are often highly non-linear and learn complex interactions between features. Methods exist to extract interactions from a variety of ML models including random forests [46, 47] and neural networks [48, 49].

Local

In the local scope scenario, the algorithm can be explained solely for every single prediction. Additionally, local scope interpretations can sometimes be aggregated for the entire dataset to provide a global interpretability scope. The most popular approach to prediction-level interpretation has involved assigning importance scores to individual features. Intuitively, a variable with a significantly positive (negative) score made a profoundly positive (negative) contribution to a particular prediction. In the deep learning literature, several different approaches have been proposed to address this problem [41, 50]. These are often displayed in the form of a heat map highlighting the most influential features. Note that feature importance scores at the prediction level can offer much more information than feature importance scores at the global level. This results from heterogeneity in a non-linear model: the importance of a feature can vary for different examples due to interactions with other features. Because small sections of a machine-learned response function are more likely to be linear, monotonic, or otherwise well-behaved, local explanations can be more accurate than global explanations.

It is worthy to note that it is also very likely that the best explanations

of a machine learning model will come from combining the results of global and local interpretation techniques. In the next section, various aspects such as machine learning applications, challenges, and limitations faced by machine learning researchers in this field are investigated.

2.4 Machine learning in Critical care

In this section, varying applications of the machine in the critical care domain and the existing challenges and limitations are discussed.

2.4.1 Applications

The typical machine learning applications in the critical care domain are based on predictive modeling and clustering models.

Predictive models

Numerous predictive models in ICU exist, such as Acute Physiology and Chronic Health Evaluation score (APACHE), and Simplified Acute Physiology Score (SAPS). Currently, SAPS-III [51] and APACHE-IV [52] are the most used predictive models in predicting in-hospital ICU mortality.

Knaus et al. [53] developed APACHE, which is one of the well-known predictive models to estimate the risk of mortality in ICU patients based on logistic regression models. Similarly, Churpek et al. [54] developed a logistic regression model in a large dataset, including more than 250,000 patients who estimated the risk for ICU transfer, cardiac arrest, or mortality in ICU patients. In a follow-up study, Churpek et al. [55] showed that

more advanced machine learning methods, such as random forests (RF) and gradient boosted machines (GBM), could predict clinical deterioration better compared with classic logistic regression.

The first version of SAPS was developed by Le Gall et al. [56] in 1984. Moreover, Lee et.al [57] showed APACHE-IV outperforms SAPS-III in mortality prediction after liver transplantation.

Clustering models

Many clustering techniques in critical care have been used to discover the subgroups of patients with similar clinical characteristics, which may help anticipate diagnostic and treatment strategies. In this regard, Hyun et al. [58] applied the K-means clustering method, which identified three clusters with different mortality rates, hemodialysis rates, and blood product transfusion rates. In another study, Calfee et al. [59] assessed whether subphenotypes exist within acute respiratory distress syndrome (ARDS); in this context, they applied latent class modeling and identified two subphenotypes using clinical data from two randomized controlled trials of ARDS. The sub-phenotypes identified had distinct differences in inflammatory profiles, response to ventilator strategies, and clinical outcomes.

Furthermore, Luo et al. [60] analyzed multiple physiological variables of patients in the MIMIC dataset and applied non-negative matrix factorization to group-related trends, which were shown to predict 30-day mortality while maintaining model interpretability effectively. Finally, Vranas et al. [61] applied clustering analysis to discover and validate six clinically recognizable subgroups of ICU patients who differed significantly in all

baseline characteristics and clinical trajectories despite sharing common diagnoses.

2.4.2 Challenges and Limitations

The following sections explore the main challenges and limitations of machine learning in the critical care domain. Additionally, we consider the steps required to transfer these technologies from research to clinical practice.

Data accessibility and complexity

A significant challenge faced by researchers in critical care is the need for data openness and reproducibility with the demand for data privacy and security. In a recent study, Johnson et al. [62] showed an alarming lack of reproducibility in data science studies using the same ICU data, which suggests that algorithms, study procedures, source code, and even datasets should be publicly available to ensure reproducibility. However, this data openness must not result in poor data governance, lack of data security, or loss of confidentiality, all of which are necessary to perform ethical research and maintain public trust.

A common concern among clinicians is the loss of autonomy in the face of machine learning systems. This concern exists even though clinicians acknowledge that the complexity of medicine exceeds the unaided human mind's capacity and that perhaps these novel computational systems can help manage some of this complexity [63]. In other words, humans typically make decisions using fewer than six data points because anything

more than that becomes cognitively too expensive [64]. However, an ICU patient can generate thousands of data points in a single day. When fatigue and interruptions are added, it is not surprising that some clinical decisions end up being suboptimal [65]. Conversely, computers can explore continuously through tens of thousands of data points. They can quickly analyze complex nonlinear interactions between variables effortlessly.

Data incompleteness and Imbalanced data

The data missing from patient profiles in intensive care units (ICUs) are substantial and unavoidable [66, 67, 68]. However, this incompleteness is not always random or because of imperfections in the data collection process. Many studies have focused on resolving this issue [69]. Although many researchers treat missing data as a challenge [70], others continue to debate whether a lack of completeness also provides useful information [71]. Recently, Angiel et al. [71] demonstrated that the laboratory ordering time for some laboratory tests is more informative than the actual values in predicting 3-year survival. In general, two reasons are given for missing data in EHRs:

- No intention to collect: the clinical variable was never measured because there was no clinical indication to do so.
- Intention to collect: variables are missing, although the variables were measured [67].

Therefore, the health care process affects the recorded EHR and can cause incompleteness in data.

The imbalance property that is common to many real clinical datasets makes classification a challenging task. The imbalanced classification problem in the healthcare domain, where data are often highly skewed due to individual heterogeneity and diversity, affects issues such as cancer diagnostics [72], patient safety informatics [73], and disease risk prediction [74]. Most standard classifiers, such as logistic regression and the support vector machine, implicitly assume that both classes are equally distributed. Additionally, these methods are designed to maximize overall classification accuracy. As a result, they favor the majority class, resulting in low sensitivity toward the minority class [75].

Real-world applications

Like most emerging technologies, machine learning research products in critical care will undoubtedly go through a series of hype and disappointment before becoming accepted, proven assets in the study, and care of critically ill patients. One of the main challenges that machine learning faces in critical care are that, despite the increasing number of studies and publications in the field, there have been few examples of machine learning projects that have resulted in the successful implementation of data-driven systems in the ICU [76]. This lack of exposure in the clinical setting inevitably results in clinicians' mistrust of these data-driven systems.

Interpretability

Although clinicians are happy to use the machine learning model to investigate hidden patterns in the patient records, however, they are cautious of

the idea of sharing clinical decision-making responsibilities with machine learning algorithms, especially when they view them as "black boxes" [77]. Likely, only implementing well-designed, interpretable, and practical data-driven systems in the ICU will make clinicians start to gain trust in them. Furthermore, the implementation of these systems must be performed under the rigorous umbrella of well-controlled experimental studies, including (but not limited to) simulation testing and randomized controlled trials. The medical informatics literature has good examples of using scientifically rigorous approaches to implementing and testing digital solutions such as clinical decision support tools and can serve as a model to follow.

The evaluation metrics employed in this thesis are studied further in the following section.

2.5 Evaluation Metrics

To automatically evaluate predictive models' performance, metrics relying on gold standard outcomes are used. Here we give an overview of the automatic metrics that are used in our experimental settings, namely: Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), Area Under Receiver Operating Characteristic (AUROC), Area Under Precision-Recall Curve (AUPRC), Precision, Recall, Matthews correlation coefficient (MCC), R-squared (R^2), Mean Absolute Error (MAE).

Actual / Predicted	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Table 2.1: *Confusion Matrix for Binary Classification*

2.5.1 Classification

Sensitivity, Specificity, PPV, and NPV, collectively known as "test characteristics," are important ways to express the usefulness of diagnostic tests [78]. In this context, the Sensitivity metric is used to measure the fraction of positive cases that are correctly classified as positive, while the specificity metric is used to measure the fraction of negative cases that are correctly classified as negative, by referring to Table 2.1 formally, we have:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.10}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2.11}$$

PPV is the percentage of patients who are predicted as positive and have the disease; NPV is the percentage of patients who are predicted as negative and do not have the disease, formally:

$$PPV = \frac{TP}{TP + FP} \tag{2.12}$$

$$NPV = \frac{TN}{TN + FN} \tag{2.13}$$

Sensitivity and specificity are fixed for a particular type of prediction, while PPV and NPV depend directly upon the prevalence of a disease in a population. Therefore the PPV will increase with increasing prevalence,

and NPV decreases with an increase in prevalence. AUROC is a graphical plot that illustrates a binary classifier system's diagnostic ability as its discrimination threshold is varied. In other words, AUROC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

As far as in the clinical domain, one of the challenges is class imbalance; AUPRC is a highly informative metric in a imbalanced data setting. Similarly, MCC is a measure of the quality of binary classification problems because it takes into account the balance ratios of the four confusion matrix categories (true positives, true negatives, false positives, false negatives), formally:

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{2.14}$$

It is worth mentioning that in addition to AUROC and AUPRC, which are applicable in multi-class classification, there is an additional metric called F1-score that could be used to evaluate the model's performance. F1-score is calculated based on Precision and Recall. Formally:

$$F1 - score = 2 * \left(\frac{Precision * Recall}{Precision + Recall} \right) \tag{2.15}$$

The formula of F1-score can be interpreted as a weighted average between Precision and Recall, where F1-score reaches its best value at 1 and worst score at 0, the harmonic mean is useful to find the best trade-off between the two Precision and Recall [79].

2.5.2 Regression

Several statistical metrics have been used to evaluate the regression methods' predictive efficiency, as presented below. Among them, the coefficient of determination or R-squared (R^2) and Mean Absolute Error (MAE) are predominantly used to measure these methods' average error. R-squared provides a measure of how well-observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, formally:

$$R^2 = (\frac{1}{n}\frac{\sum_{i=1}^{n}[(x_i - \bar{x}) * (y_i - \bar{y})]}{\sigma_x * \sigma_y})^2 \tag{2.16}$$

Where n is the number of observations used to fit the model, x_i is the x value for observation i, $\bar{x}$ is the mean x value, y_i is the y value for observation i, $\bar{y}$ is the mean y value, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

MAE is an excellent metric to measure the average error and a good indicator of average model performance. Formally:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.17}$$

In the following Chapters, the main contributions of the thesis are described in detail. In Chapter 3 a detailed description of "Benchmarking machine learning models on eICU-CRD dataset" is provided, while in Chapter 4 a comprehensive study regarding "Delirium Prediction in Critically Ill Patients" is described, and in Chapter 5 we proposed and described an interpretable machine learning algorithm for the time-series EHRs data.

Chapter 3

Benchmarking Machine Learning Models in eICU-CRD Dataset

Progress of machine learning in critical care has been difficult to track, in part due to absence of public benchmarks. Other fields of research (such as computer vision and natural language processing) have established various competitions and public benchmarks. Recent availability of large clinical datasets has enabled the possibility of establishing public benchmarks. Taking advantage of this opportunity, we propose a public benchmark suite to address four areas of critical care, namely mortality prediction, estimation of length of stay, patient phenotyping and risk of decompensation. We define each task and compare the performance of both clinical models as well as baseline and deep learning models using eICU-CRD (Collaborative Research Database) of around 73,000 patients. This is the first public benchmark on a *multi-centre* critical care dataset, comparing the performance of clinical gold standard with our predictive model. We also investigate the impact of numerical variables as well as handling of cate-

gorical variables on each of the defined tasks. The source code, detailing our methods and experiments is publicly available such that anyone can replicate our results and build upon our work.

3.1 Introduction

Increasing availability of clinical data and advances in machine learning have addressed a wide range of healthcare problems, such as risk assessment and prediction in acute, chronic and critical care. Critical care is a particularly data-intensive field, since continuous monitoring of patients in Intensive Care Units (ICU) generates large streams of data that can be harnessed by machine learning algorithms. However, progress in harnessing digital health data faces several obstacles, including reproducibility of results and comparability between competing models. While, other areas of machine learning research, such as image and natural language processing have established a number of benchmarks and competitions (including ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [80] and National NLP Clinical Challenges (N2C2) [81], respectively), progress in machine learning for critical care has been difficult to measure, in part due to absence of public benchmarks. Availability of large clinical data sets, including Medical Information Mart for Intensive Care (MIMIC III) [4] and more recently, a multi-centre eICU-CRD (Collaborative Research Database) [5] are opening the possibility of establishing public benchmarks and consequently tracking the progress of machine learning models in critical care. Availing of this opportunity, we propose a public benchmark suite to ad-

dress four areas of critical care, namely mortality prediction, estimation of length of stay (LoS), patient phenotyping and risk of decompensation. We define each task and evaluate our algorithms on a multi-centre dataset of 73,718 patients (containing 4,564,844 clinical records) collected from 335 ICUs across 208 hospitals. While there has been work in this area that has focused on the single-center MIMIC III clinical dataset [82], our work is the first to focus on a multi-center critical care dataset, the eICU-CRD [5]. Evaluating models on a multi-center dataset typically results in the inclusion of a wider range of patient groups, larger number of patients, external validity and lower systematic bias in comparison to a single-center dataset, resulting in increased generalizability of the study [83, 84]. However building a predictive model on a multi-centre dataset is more challenging due to heterogeneity of the data. Nevertheless, the performance of our models (as measured by AUROC) compare favourably with the performance of the models using the single-center MIMIC-III dataset as reported in [82].

The main contributions of this work are as follows: i) we provide the baseline performance, (using either on clinical gold standard or Logistic/Linear Regression algorithm) and compare it against our benchmark result, achieved using a model based on bidirectional long short-term memory (BiLSTM); ii) investigate impact of categorical and numerical variables on all four benchmarking tasks; iii) evaluate entity embedding for categorical variables, versus one hot encoding; iv) show that for some tasks the number of variables can be reduced significantly without greatly impacting prediction performance; and v) report six evaluation metrics for each of the tasks, facilitating direct comparison with future results.

3.2 Related work

In this Section, we provide a brief review of the most relevant studies related to each of the tasks, mortality, length of stay, phenotyping, and physiologic decompensation. We briefly review the other benchmarking studies in critical care, related to our work.

Mortality prediction. Many clinical scoring systems have been developed for mortality prediction, including Acute Physiology and Chronic Health Evaluation (APACHE III [85], APACHE IV [52]) and Simplified Acute Physiology Score [86] (SAPS II, SAPS III). Most of these scoring systems use logistic regression to identify predictive variables to establish these scoring systems. Providing an accurate prediction of mortality risk for patients admitted to ICU using the first 24/48 hours of ICU data could serve as an input to clinical decision making and reduce the healthcare costs. In this regard, recent advances in deep learning have been shown to outperform the conventional machine learning methods as well as clinical prediction techniques such as APACHE and SAPS [82] [87] [88]. Mortality prediction has been a popular application for deep learning researchers in recent years, though model architecture and problem definition vary widely. Convolutional neural network and gradient boosted tree algorithm have been used by Darabi et al. [89], in order to predict long-term mortality risk (30 days) on a subset of MIMIC-III dataset. Similarly, Celi et al. [90] developed mortality prediction models based on a subset of MIMIC database using logistic regression, Bayesian network and artificial neural network.

Length of stay. Resource allocation and identifying patients with unexpected extended ICU stays would help decision-making systems to improve the quality of care and ICU resource allocation. Therefore forecasting the length of stay (LoS) in ICU would be significantly important in order to provide high-quality care to a patient, and it would avoid extra costs for care providers. In this regard, Sotoodeh et al. [91] applied hidden markov models to predict LoS by using the first 48 hours of physiological measurements. Ma et al. [92] defined LoS as a classification problem in which the objective was to create a personalized model for patients to forecast LoS. Previous studies [87] [82] have shown that deep learning models obtain good results on forecasting length of stay in ICU. In this regard, Tu et al. [93] applied neural network based methods on a Canadian private dataset, which includes patients with cardiac surgery. The developed model was able to detect the patient with low, intermediate, and high prolonged stay in ICU.

Phenotyping. Phenotyping has been a popular task in recent years [94] [95], although problem definition varies widely, from focusing on ICD based diagnosis [88] up to including clinical procedures and medications [96] [97]. Several works on phenotyping from clinical time series have focused on variations of tensor factorization and related models [94] [95] [98], and the most recently published studies on phenotyping are focused on deep learning methods. In this regard, Razavian et al. [99] and Lipton et al. [88] applied deep learning methods to predict diagnoses. While the first trained RNN LSTM and CNN for prediction of 133 diseases based on 18 laboratory tests on a private dataset including 298k patients, the latter applied an

RNN LSTM on a single-center, private pediatric intensive care unit (PICU) dataset in order to classify 128 diagnoses given 13 clinical measurements.

Physiologic decompensation. Early detection of physiologic decompensation could be used to avoid or delay the occurrence of decompensation. Recently machine learning researchers have started to apply various machine learning methods in order to predict the decompensation incident. Recent study by Ren et al. [100] applied gradient boosting models (GBM) to predict required intubation 3 hours ahead of time, in this work they used a cohort of 12,470 patients to predict unexpected respiratory decompensation. Differently, Xu et al. [101] proposed a deep learning model to predict the decompensation event. The proposed attention-based model was applied on MIMIC-III Waveform Database and it outperformed several machine learning and deep learning models.

Benchmark. Harutyunyan et al. [82] developed a deep learning model based on RNN LSTM called multi-task RNN, in order to predict a number of clinical tasks such as mortality prediction in hospital, physiologic decompensation, phenotyping, and length of stay in ICU unit. The proposed model was applied on MIMIC-III dataset. Similarly, Purushotham et al. [87] have provided a single-center benchmark of several machine learning and deep learning models trained on MIMIC-III for various tasks, showing that deep learning models consistently outperformed conventional machine learning models and clinical scoring systems. One common theme across the reviewed work is that the current literature focuses on single-center databases, while we did not find any work in this area that addressed multi-centre datasets, including the associated challenges.

3.3 Materials and methods

3.3.1 Ethics statement

The analysis using the eICU-CRD is exempt from institutional review board approval due to the retrospective design, lack of direct patient intervention, and the security schema, for which the re-identification risk was certified as meeting safe harbor standards by an independent privacy expert (Privacert, Cambridge, MA) (Health Insurance Portability and Accountability Act Certification no. 1031219-2).

3.3.2 eICU dataset description and cohort selection

The eICU-CRD [5] is a multi-center intensive care unit database with high granularity data for over 200,000 admissions to ICUs monitored by eICU-CRD programs across the United States. The eICU-CRD comprises 200,859 patient unit encounters for 139,367 unique patients admitted between 2014 and 2015 to 208 hospitals located throughout the US. We selected adult patients (age $>$ 18) that had an ICU admission with at least 15 records, leading to 73,718 unique patients with a median age of 62.41 years (IQR, 52-75), 45.5% female. Hospital mortality rate was 8.3% and average LoS in hospital and in unit were 5.29 days and 3.9 days respectively (further details are provided in Table 3.1). Cohort selection criteria are detailed in Figure 3.1.

The final patient cohort contained 4,564,844 clinical records where we grouped these records on 1 hour window, imputed the missing values based on the mean of that window and took the last valid record of that specific

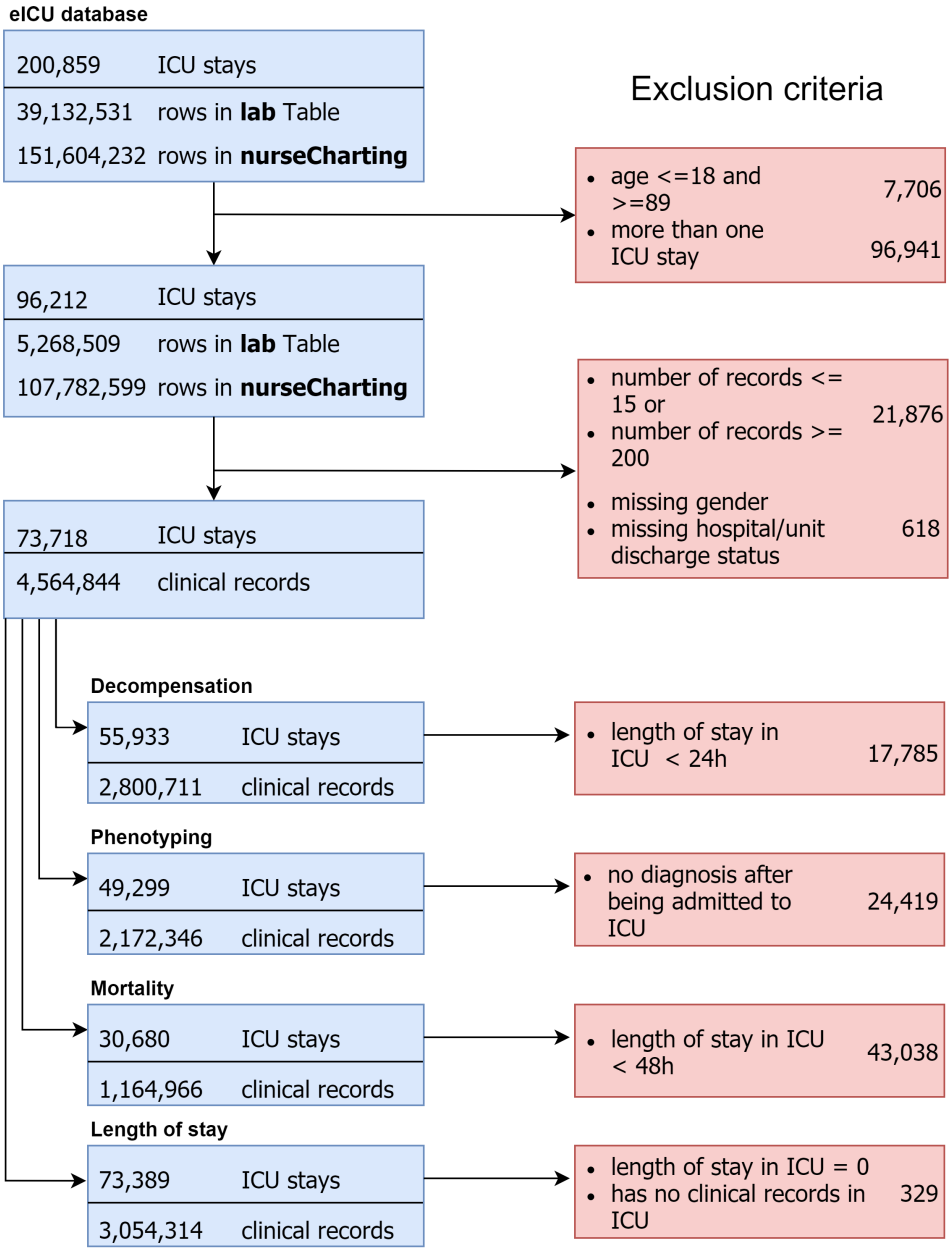


Figure 3.1: *Cohort selection criteria*

window. Out of 31 tables in the eICU-CRD (v1.0) we selected variables from the following tables: *patient* (administrative information and patient demographics), *lab* (Laboratory measurements collected during routine care), *nurse charting* (bedside documentation) and *diagnosis* based

	Overall	Dead at Hospital	Alive at Hospital
ICU Admissions	73,718	6,167	67,551
Age	62.41 [52-75]	68.12 [59-80]	61.8 [52-75]
Gender (F)	33,544 (45.5)	2,830 (45.8)	30,714 (45.4)
Ethnicity			
Caucasian	56,973 (77.2)	4,866 (78.9)	52,107 (77.1)
African American	7,982 (10.8)	582 (9.4)	7,400 (10.9)
Hispanic	2,937 (3.98)	226 (3.6)	2,711 (4)
Asian	1,174 (1.59)	97 (1.5)	1,077 (1.5)
Native American	413 (0.56)	42 (0.68)	371 (0.54)
Unknown	4,239 (5.7)	354 (5.7)	3,885 (5.7)
Outcomes			
Hospital LoS* (days)	5.29 [2.53-6.84]	3.9 [1.42-5.22]	5.41 [2.65-6.92]
ICU LoS* (days)	2.32 [1.01-2.91]	3.17 [1.19-4.43]	2.24 [1-2.83]
Hospital Death	6,167 (8.36)	6,167 (100)	-
ICU Death	4,575 (6.2)	4,575 (74.1)	-

Table 3.1: *Characteristics and mortality outcome measures. *LoS (Length of Stay). Continuous variables are presented as Median [Interquartile Range Q1–Q3]; binary or categorical variables as Count (%)*

on advice from a clinician as well as consistency with other similar tasks reported in the related work section. Selected variables are shown in Table 3.2 and are common across all the four tasks.

Variable	Data Type
Heart rate	Numerical
Mean arterial pressure	Numerical
Diastolic blood pressure	Numerical
Systolic blood pressure	Numerical
O2	Numerical
Respiratory rate	Numerical
Temperature	Numerical
Glucose	Numerical
FiO_2	Numerical
pH	Numerical
Height	Numerical
Weight	Numerical
Age	Numerical
Admission diagnosis	Categorical
Ethnicity	Categorical
Gender	Categorical
Glasgow Coma Score Total	Categorical
Glasgow Coma Score Eyes	Categorical
Glasgow Coma Score Motor	Categorical
Glasgow Coma Score Verbal	Categorical

Table 3.2: *Selected variables for all the four tasks*

3.3.3 Data Preprocessing

The clinical variables depending on the clinical needs and the nature of the clinical variable, are measured in different intervals; for instance, the vital signs are recorded more frequently than lab measurements. In this context, to address the sparsity of data, all variables were aggregated into hourly

intervals, where the last measured value was used as a candidate for that interval. In cases where the last value for each variable is not measured in the interval, the representative of that interval was computed by averaging the available measurements in the interval. Missing values that were collected hourly, like vital signs, were imputed by normal values. Categorical variables were converted into a vector to capture the semantics of each category at the model derivation phase. For all continuous variables, we utilized the recorded value in the database without any adaptation. To address the imbalanced data in the classification tasks, we employed an over-sampling method to have an equal number of samples across different classes while providing the data as input to the models.

3.3.4 Description of tasks

In this section, we define four different benchmark tasks, namely in-hospital mortality prediction, remaining LoS forecasting, patient phenotyping, and risk of physiologic decompensation. After applying selection criteria for each task, the resulting patient cohorts are outlined in Table 3.3

Task	No. of patients	Clinical records
In-hospital Mortality	30,680	1,164,966
Remaining LoS	73,389	3,054,314
Phenotyping	49,299	2,172,346
Physiologic Decompensation	55,933	2,800,711

Table 3.3: *Number of patients and records in four tasks*

Mortality prediction

In-hospital mortality is defined as the patient's outcome at the hospital discharge. This is a binary classification task, where each data sample spans a 1-hour window. The cohort for this task was selected based on the presence of hospital discharge status in patients' record and length of stay of at least 48 hours (we focus on prediction during the first 24 and 48 hours). This selection criteria resulted in 30,680 patients containing 1,164,966 records.

Length of stay prediction

Length of stay is one of the most important factors accounting for the overall hospital costs, as such its forecast could play an important role in healthcare management [102]. Length of stay is estimated through analysis of events occurring within a fixed time-window, once every hour from the initial ICU admission. This is a regression task, where we use 20 clinical variables described in Table 3.2. For this cohort we selected patients whose LoS was present in their records. These selection criteria resulted in 73,389 ICU stays, containing 3,054,314 records. The mean LoS was 1.86 days with standard deviation of 1.94 days, as shown in Table 3.1.

Phenotyping

Phenotyping is a classification problem where we classify whether a condition (ICD-9 code) is present in a particular ICU stay record. Since any given patient may have more than one ICD-9 code, this is defined as a multi-label classification problem.

While our definition is focused on diagnosis using ICD codes for this task, the definition of phenotyping may encompass other domains, such as procedures [96] [97] for example. However, expanding the definition of phenotyping beyond standardised ICD codes would have required development of non-standardised rules, as no common standard approach for defining and validating EHR phenotyping algorithms exists [103] [104]. Consequently, it would have been challenging to compare this work with the already published benchmarks. Furthermore, there is some concern regarding reproducibility of rule-based phenotyping as found in [104]. Considering these issues, as well as keeping consistent with previously published benchmarks, we settled on using ICD codes as the basis for the definition of this task. Accordingly, the dataset contains 767 unique ICD codes, which are grouped into 25 categories shown in Table 3.4. The cohort for this task, considering initial inclusion criteria as well as recorded diagnosis during the ICU stay, resulted in 49,299 patients.

Physiologic Decompensation

There are a number of ways to define decompensation, however in clinical setting majority of early warning systems, such as National Early Warning Score (NEWS) [105] are based on prediction of mortality within the next time window (such as 24 hours after the assessment). Following suit and keeping consistent with previously published benchmarks [82], we also define decompensation as a binary classification problem, where the target label indicates whether the patient dies within the next 24 hours. The cohort for this task results in 55,933 patients (2,800,711 records), where

Type	Phenotype
Acute	1- Respiratory failure; insufficiency; arrest
	2- Fluid and electrolyte disorders
	3- Septicemia
	4- Acute and unspecified renal failure
	5- Pneumonia
	6- Acute cerebrovascular disease
	7- Acute myocardial infarction
	8- Gastrointestinal hemorrhage
	9- Shock
	10- Pleurisy; pneumothorax; pulmonary collapse
	11- Other lower respiratory disease
	12- Complications of surgical
	13- Other upper respiratory disease
Chronic	14- Hypertension with complications
	15- Essential hypertension
	16- Chronic kidney disease
	17- Chronic obstructive pulmonary disease
	18- Disorders of lipid metabolism
	19- Coronary atherosclerosis and related
	20- Diabetes mellitus without complication
Mixed	21- Cardiac dysrhythmias
	22- Congestive heart failure; non hypertensive
	23- Diabetes mellitus with complications
	24- Other liver diseases
	25- Conduction disorders

Table 3.4: *Phenotype categories*

the decompensation rate is around 6.5% (3,664 patients).

3.3.5 Prediction algorithms

Baselines

We compare our model with two standard baseline approaches namely, logistic/linear regression (LR) and a 1-layer artificial neural network (ANN). The embeddings for these models are learned in the same way as for the proposed BiLSTM model as explained in the section that follows.

Deep Learning models

In this section, we describe the selected clinical variables, approaches to represent these variables as well as baseline and deep models used in this study. The architecture of this work consists of three modules, namely input module, encoder module and output module as shown in Fig. 3.2.

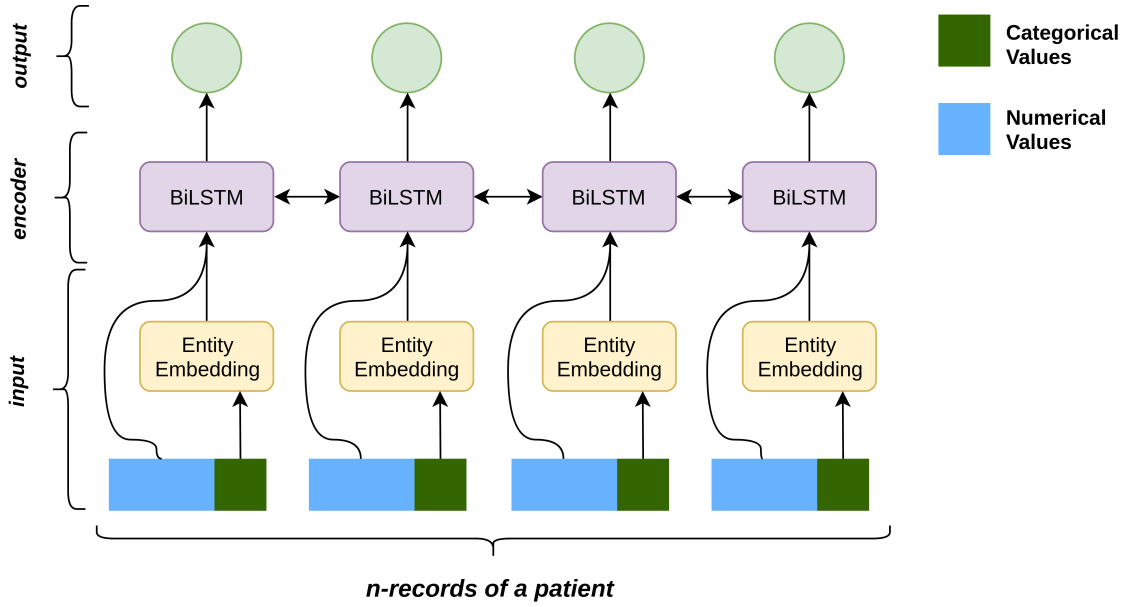


Figure 3.2: *Model architecture*

Input representation: We process and model both numerical and categorical variables separately, as shown in Table 3.2. Categorical variables are represented using either one-hot encoding (OHE) or entity embedding (EE). OHE is the baseline approach that converts the variables into binary representation. Using this approach for our 7 categorical variables results in 429 unique records, rendering a large sparse matrix. In response, we represent each variable as an embedding and compare the performance with the OHE approach. We use entity embedding [106], where each categorical

variable in the dataset is mapped to a vector and the corresponding embedding is added to the patient's record. This entity embedding is learned by the neural network during the training phase along with other parameters. As such, the final representation of the input at time t is as follows:

$$x_t = [Num_t; U(Cat_t)]$$

where Num_t is the numerical variable, Cat_t is the categorical variable at time t and U is the embedding matrix learned by the model.

Encoder: To capture sequential dependency in our data, we use Recurrent Neural Network (RNN) that resemble a chain of repeating modules to efficiently model sequential data [107]. They take sequential data $X = (x_1, x_2, ...x_n)$ as input and provide a hidden representation $H = (h_1, h_2, ...h_n)$ which captures the information at every time step in the input. Formally,

$$h_t = f(x_t + Wh_{t-1})$$

where x_t is the input at time t, W is the parameter of RNN learned during training and f is a non-linear operation such as sigmoid, tanh or ReLU.

A drawback of regular RNNs is that the input sequence is fed in one direction, normally from past to future. In order to capture both past and future context, we use a Bidirectional Long Short Term Memory (BiLSTM) [108] [26] for our model, which processes the input in both forward and backward direction. Using a BiLSTM the model is able to capture the context of a record not only by its preceding records but also with the following records, allowing the model to produce more informed predictions. The input at time t is represented by both its forward context $\overrightarrow{h_t}$

and backward context $\overleftarrow{h_t}$ as $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$. Similarly, the representation of the completed patient record is given by $h_T = [\overrightarrow{h_n}; \overleftarrow{h_1}]$.

Output: The choice of output layer is based on whether the benchmarking task is a regression or a classification task.

Remaining LoS prediction is a regression task, in which we predict the remaining LoS record-wise. That is, each patient record is fed to the model to predict the remaining LoS for that specific time step. This task is realized using a many to many architecture, where we assign a label to each patient record. The score for this task is obtained using:

$$\widehat{y_t} = ReLU(W \cdot h_t) \tag{3.1}$$

where y_t is the remaining LoS predicted and $ReLU$ is the non-linear activation function used as the prediction of remaining LoS cannot be negative.

In-hospital mortality and decompensation are binary classification tasks. For the in-hospital mortality the many to one architecture is applied and the classifier is as follows:

$$\widehat{y} = \sigma(W \cdot h_T) \tag{3.2}$$

For the decompensation task, a many to many architecture is applied. Prediction at each-time step is treated as a binary classification and the classifier is defined as:

$$\widehat{y_t} = \sigma(W \cdot h_t) \tag{3.3}$$

Phenotyping is defined as a multi-label task with 25 binary classifiers for each phenotype, and the score for the task is obtained using:

$$\widehat{y_t}^n = \sigma(W_n \cdot h_t) \tag{3.4}$$

where t is the time step and n is the phenotype being predicted and W_n is the model parameter.

3.3.6 Source Code

We made use of several open-source libraries based on Python to conduct our experiments; machine learning framework Scikit-learn [109] and DL framework Pytorch [110].

The source code for our experiments is publicly available at this GitHub repository, so that anyone with access to the public eICU-CRD can replicate our experiments and build upon our work.

3.4 Results

In this section, we report benchmarking results of methods and prediction algorithms, focusing on answering the following questions: (a) How does performance of our model compare to the performance of clinical scoring systems as well as baseline algorithms (logistic/linear regression in our case); and (b) What is the impact on prediction performance when using different feature sets, such as categorical and numerical variables, solely categorical and solely numerical variables? We evaluate our model through a 5-fold cross-validation using the following evaluation metrics: for the regression task we report coefficient of determination R^2, and Mean Absolute Error (MAE), while for the classification tasks we report AUROC (Area Under the Receiver Operating Characteristics), AUPRC (Area Under the Precision Recall Curve), Specificity and Sensitivity (set to 90% to facilitate

direct comparison of results), Positive Predictive Value (PPV) and Negative Predictive Value (NPV); all the numerical results are presented with 95% confidence interval (CI).

3.4.1 Mortality prediction

Results from this task indicate that the proposed approach of learning embeddings for categorical variables is more effective than OHE representation. This holds true for both baseline models (LR and ANN) as well as BiLSTM model, reflected in the prediction performance of each model. Furthermore, *BiLSTM* model result difference is statistically significant compared to all the other approaches in predicting mortality in both the 24 hour window and the 48 hour window considering AUROC, AUPRC, and Specificity metrics as shown in Table 3.5. It is interesting to note that using only categorical variables (reducing the number of variables from 20 to only 7) with embedding provides a better performance than using numerical variables only (AUROC 78.23% (95% CI, 77.08% - 79.43%) vs. 76.60% (95% CI, 76.03% - 77.24%) for the first 24h). These results suggest that EE of categorical features in vector space is more effective in the prediction of mortality.

Data	Model	Num.	Cat.	Repn.	AUROC% (95% CI)	AUPRC% (95% CI)	Spec.% (95% CI)	Sens.%	PPV% (95% CI)	NPV%(95% CI)
First 24 hours	APACHE	✓	✓	Not spec.	77.30	41.23	38.74	86	57.09	93.07
	LR	✓	✓	EMB	79.95 (79.10 - 80.85)†	40.56 (38.84 - 42.27)†	46.82 (43.34 - 50.30)†	90	64.73 (60.08 - 69.37)	90.10 (89.93 - 90.27)†
	ANN	✓	✓	EMB	82.46 (81.75 - 83.24)†	45.94 (44.77 - 47.11)†	50.70 (46.52 - 54.89)	90	**66.02 (64.01 - 68.03)**	90.68 (90.38 - 90.98)
	BiLSTM	✓	✓	EMB	**83.70 (83.07 - 84.40)**	**48.47 (46.53 - 50.40)**	**53.44 (50.54 - 56.34)**	90	65.91 (57.07 - 74.76)	91.16 (90.18 - 92.13)
	BiLSTM	✓	✓	OHE	82.89 (82.18 - 83.65)‡	46.71 (44.52 - 48.90)	50.78 (48.18 - 53.37)	90	64.36 (59.12 - 69.60)	**91.21 (90.67 - 91.75)**
	BiLSTM	×	✓	EMB	78.23 (77.08 - 79.43)†	39.89 (37.60 - 42.18)†	41.74 (39.58 - 43.91)†	90	65.19 (53.04 - 77.34)	90.42 (89.62 - 91.22)
	BiLSTM	✓	×	×	76.60 (76.03 - 77.24)†	38.61 (36.89 - 40.34)†	36.80 (34.40 - 39.20)†	90	63.95 (60.40 - 67.50)	90.18 (89.76 - 90.60)‡
First 48 hours	LR	✓	✓	EMB	82.31 (81.56 - 83.12)†	45.41 (44.01 - 46.80)†	50.96 (46.72 - 55.20)†	90	**68.54 (64.91 - 72.18)**	90.34 (90.11 - 90.57)†
	ANN	✓	✓	EMB	85.27 (84.69 - 85.90)†	52.34 (51.01 - 53.67)†	57.31 (55.03 - 59.58)	90	67.73 (64.27 - 71.19)	91.73 (91.33 - 92.13)
	BiLSTM	✓	✓	EMB	**86.55 (85.65 - 87.52)**	**54.98 (53.20 - 56.77)**	**59.70 (54.85 - 64.54)**†	90	67.16 (56.38 - 77.95)	**92.22 (90.82 - 93.63)**
	BiLSTM	✓	✓	OHE	85.28 (84.29 - 86.37)‡	52.38 (51.25 - 53.50)†	57.22 (53.17 - 61.26)	90	63.63 (56.36 - 70.90)	92.01 (91.20 - 92.82)
	BiLSTM	×	✓	EMB	80.33 (79.19 - 81.54)†	45.51 (42.88 - 48.14)	45.51 (43.00 - 48.02)†	90	63.00 (46.27 - 79.73)	91.33 (90.12 - 92.54)
	BiLSTM	✓	×	×	81.21 (79.75 - 82.74)†	45.78 (44.07 - 47.49)†	47.90 (43.48 - 52.31)†	90	66.30 (58.76 - 73.83)	90.85 (90.53 - 91.17)‡

Table 3.5: In-hospital mortality prediction during first 24 and 48 hours in ICU. (*Num.* and *Cat.* indicate presence of numerical and categorical variables respectively. *Repn.* indicates representation of categorical variables, either One Hot Encoding (OHE) or embedding (EMB)). If the differences between *BiLSTM* result and other models (LR, ANN, and BiLSTM with varying data representation) is statistically significant on a two-tailed t-test then it is indicated with †, ‡ († $p < 0.05$, ‡ $p < 0.1$). The best-performing metric values are represented in **bold** font.

3.4.2 Remaining length of stay in unit prediction

Predicting remaining LoS in the ICU with a limited number of clinical variables is a highly challenging task, as shown by [82]; As Indicated in Figure 3.1 to predict remaining LoS in ICU, we included patients who stayed in ICU, which resulted in a wide range of LoS in ICU from 0 to 51 days as shown in Figure 3.4 which makes the prediction of remaining LoS in the ICU more challenging.

Our approach is designed in a way that the model requires 12 hours (derivation window) of observation data to predict the remaining LOS at the 13th hour, with a 6-hour sliding window, as shown in Figure 3.3. By using the past 12 hours of stay in the ICU, the model creates a data representation and subsequently predicts the remaining LoS at the 13th

hour up to the ICU discharge.

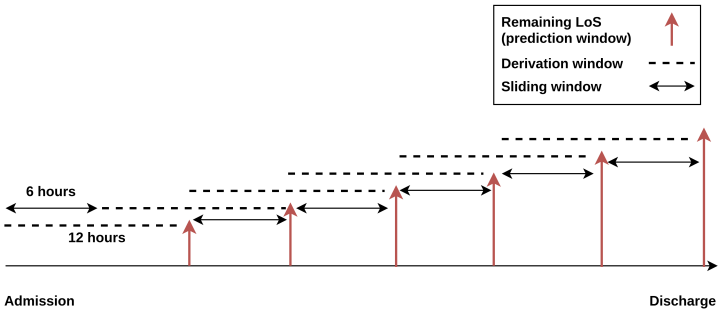


Figure 3.3: *Remaining LoS prediction schematic*

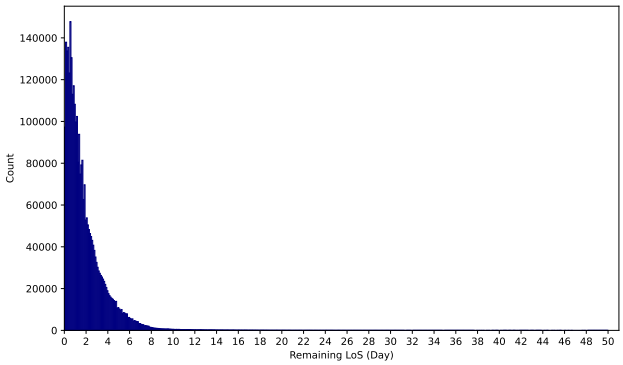


Figure 3.4: *Remaining LoS in ICU distribution*

Results from this task indicate that comparing *BiLSTM* with ANN, we achieved incremental improvements in the R^2 although the result differences were not statistically significant due to the low number of time steps provided to *BiLSTM*. Moreover, statistically significant differences were found comparing *BiLSTM* result to LR, BiLSTM with OHE representation, BiLSTM with solely numerical, and BiLSTM with solely categorical variables in predicting remaining LoS in the ICU as shown in Table 3.6. It

is noteworthy that using only numerical variables provides better performance than categorical variables only (R^2 0.046 (95% CI, 0.037 - 0.056) vs. -0.003 (95% CI, -0.042 - 0.035)). We focus the comparison of our results on R^2 as there is some evidence to suggest that R^2 metric is more informative than MAE in evaluation of regression analysis [111].

Data	Model	Num.	Cat.	Repn.	R^2(95% CI)	MAE [Day](95% CI)
In ICU unit	LR	✓	✓	EMB	0.042 (0.040 - 0.044)†	1.301 (1.288 - 1.314)
	ANN	✓	✓	EMB	0.066 (0.056 - 0.076)	**1.282 (1.246 - 1.319)**
	BiLSTM	✓	✓	EMB	**0.075 (0.066 - 0.083)**	1.292 (1.256 - 1.328)
	BiLSTM	✓	✓	OHE	0.054 (0.036 - 0.071)†	1.293 (1.262 - 1.322)
	BiLSTM	×	✓	EMB	-0.003 (-0.042 - 0.035)†	1.320 (1.286 - 1.354)
	BiLSTM	✓	×	×	0.046 (0.037 - 0.056)†	1.317(1.282 - 1.353)

Table 3.6: Remaining LoS prediction in rolling window manner.
If the differences between *BiLSTM* result and other models (LR, ANN, and BiLSTM with varying data representation) is statistically significant on a two-tailed t-test then it is indicated with †, ‡ († $p < 0.05$, ‡ $p < 0.1$). The best-performing metric values are represented in **bold** font.

3.4.3 Phenotyping

For the phenotyping task, we focus on comparing performance (AUROC with 95% CI) of the proposed model on different subset of features, namely numerical versus categorical variables. Results from this task indicate that using both numerical and categorical features we achieved statistically significant differences compare to results from employing solely numerical and solely categorical features. Moreover, using only the categorical features, modelled as entity embeddings shows a significantly higher performance 79.40% (95% CI, 77.54% - 81.26%) compared to using only the numeri-

cal features 66.25% (95% CI, 63.68% - 68.60%) as outlined in Table 3.7. Clearly categorical features are more effective in representing patients' phenotype, since integrating both of the subsets does not significantly improve the result of 79.40% from 78.12%.

In this task there is a wide difference between performance of the model on individual diseases, varying from 60.81% (95% CI, 54.86% - 66.76%) (diabetes mellitus without complications) to 94.34% (95% CI, 93.64% - 95.04%) (acute cerebrovascular disease). As a general trend prediction performance on acute diseases is higher (82.06% (95% CI, 80.58% - 83.54%)) than that on chronic diseases (73.03% (95% CI, 70.60% - 75.46%)). This may be due to the slow-progressing nature of chronic diseases, where recorded ICU data is relatively short and thus unable to fully capture events related to chronic diseases.

3.4.4 Decompensation prediction

As mentioned in Section Physiologic Decompensation, mortality prediction and decompensation are related, with the difference that in decompensation we predict whether the patient survives in the next 24 hours, given the previous 12 hours. Similar to the remaining LoS task, as shown in Figure 3.5 the model, using the previous 12 hours of data, predicts the decompensation state in a 6-hour rolling-window.

As demonstrated in Table 3.8 the *BiLSTM* outperforms ANN although the result differences were not statistically significant due to the low number of time steps provided to *BiLSTM*. Moreover, the results differences comparing *BiLSTM* to LR, BiLSTM with solely categorical variables, and

Phenotype	Prevalence	Type	Variable		
			Num & cat	Num.	Cat.
			AUROC% (95% CI)	AUROC% (95% CI)	AUROC% (95% CI)
Respiratory failure; insufficiency; arrest	0.241	acute	82.69 (81.94 - 83.44)	72.56 (71.48 - 73.64)†	81.45 (81.20 - 81.70)†
Fluid and electrolyte disorders	0.156	acute	70.92 (69.57 - 72.26)	60.14 (59.02 - 61.27)†	71.89 (71.07 - 72.72)
Septicemia	0.145	acute	91.16 (90.56 - 91.76)	70.56 (68.85 - 72.28)†	91.14 (90.49 - 91.80)
Acute and unspecified renal failure	0.142	acute	75.46 (74.70 - 76.22)	65.20 (64.16 - 66.25)†	74.16 (73.09 - 75.23)‡
Pneumonia	0.120	acute	88.49 (87.18 - 89.79)	69.46 (67.95 - 70.97)†	88.93 (88.19 - 89.67)
Acute cerebrovascular disease	0.108	acute	94.34 (93.64 - 95.04)	73.88 (72.34 - 75.42)†	94.16 (93.82 - 94.50)
Acute myocardial infarction	0.090	acute	91.16 (89.53 - 92.78)	69.55 (67.75 - 71.35)†	91.27 (90.14 - 92.40)
Gastrointestinal hemorrhage	0.079	acute	90.43 (89.18 - 91.67)	60.22 (58.95 - 61.48)†	91.20 (90.43 - 91.96)
Shock	0.068	acute	85.23 (84.62 - 85.83)	76.61 (75.30 - 77.93)†	83.06 (82.00 - 84.11)†
Pleurisy; pneumothorax; pulmonary collapse	0.039	acute	69.34 (67.04 - 71.63)	59.96 (57.59 - 62.33)†	70.78 (68.34 - 73.21)
Other lower respiratory disease	0.030	acute	80.14 (78.97 - 81.31)	57.24 (55.86 - 58.61)†	80.74 (78.98 - 82.50)
Complications of surgical	0.011	acute	68.12 (64.25 - 71.97)	52.27 (47.70 - 56.83)†	69.03 (65.80 - 72.26)
Other upper respiratory disease	0.007	acute	79.34 (76.36 - 82.31)	52.94 (42.95 - 62.94)†	76.23 (68.17 - 84.29)
Macro-average (acute diseases)	—	—	82.06 (80.58 - 83.54)	64.66 (62.30 - 67.02)†	81.85 (80.13 - 83.56)
Hypertension with complications	0.019	chronic	84.92 (82.03 - 87.80)	79.51 (78.07 - 80.94)†	80.79 (78.25 - 83.32)†
Essential hypertension	0.203	chronic	71.00 (70.44 - 71.55)	66.40 (65.28 - 67.51)†	68.33 (67.07 - 69.59)†
Chronic kidney disease	0.104	chronic	65.83 (63.70 - 67.96)	61.88 (60.25 - 63.50)†	60.53 (58.10 - 62.96)†
Chronic obstructive pulmonary disease	0.093	chronic	76.12 (74.35 - 77.88)	63.49 (61.78 - 65.20)†	74.03 (71.16 - 76.90)
Disorders of lipid metabolism	0.054	chronic	72.22 (69.88 - 74.56)	62.94 (61.81 - 64.06)†	71.74 (70.47 - 73.01)
Coronary atherosclerosis and related	0.041	chronic	80.34 (78.97 - 81.71)	64.46 (61.11 - 67.81)†	79.55 (77.84 - 81.26)
Diabetes mellitus without complication	0.006	chronic	60.81 (54.86 - 66.76)	58.16 (50.89 - 65.43)†	57.89 (51.47 - 64.31)
Macro-average (chronic diseases)	—	—	73.03 (70.60 - 75.46)	65.26 (62.74 - 67.77)†	70.40 (67.76 - 73.05)
Cardiac dysrhythmias	0.165	mixed	74.85 (72.12 - 77.57)	65.68 (65.11 - 66.25)†	71.35 (70.70 - 72.00)†
Congestive heart failure; non hypertensive	0.106	mixed	78.98 (78.16 - 79.79)	65.72 (64.94 - 66.51)†	76.72 (74.13 - 79.31)‡
Diabetes mellitus with complications	0.047	mixed	92.96 (92.41 - 93.50)	89.03 (87.08 - 90.97)†	89.36 (88.05 - 90.66)†
Other liver diseases	0.039	mixed	76.48 (73.34 - 79.62)	68.29 (66.21 - 70.37)†	76.15 (73.40 - 78.90)
Conduction disorders	0.013	mixed	83.80 (80.76 - 86.84)	67.37 (59.55 - 75.19)†	82.54 (78.14 - 86.93)
Macro-average (mixed diseases)	—	—	81.41 (79.36 - 83.46)	71.81 (68.58 - 73.86)†	79.22 (76.88 - 81.56)
Macro-average (all diseases)	—	—	79.40 (77.54 - 81.26)	66.25 (63.68 - 68.60)†	78.12 (76.01 - 80.22)

Table 3.7: Phenotyping task on eICU-CRD (reported scores are AUROC with 95% CI) If the differences between *BiLSTM* result and other models If the differences between the proposed BiLSTM model result using Num & cat variables is statistically significant than only Num variables or Cat variables on a two-tailed t-test then it is indicated with †, ‡ († $p < 0.05$, ‡ $p < 0.1$).

BiLSTM with solely numerical variables are statistically significant considering AUROC, AUPRC, and NPV metrics. Furthermore, unlike the remaining LoS task, the prediction of decompensation using the categori-

cal variables solely outperforms employing the numerical variables solely.

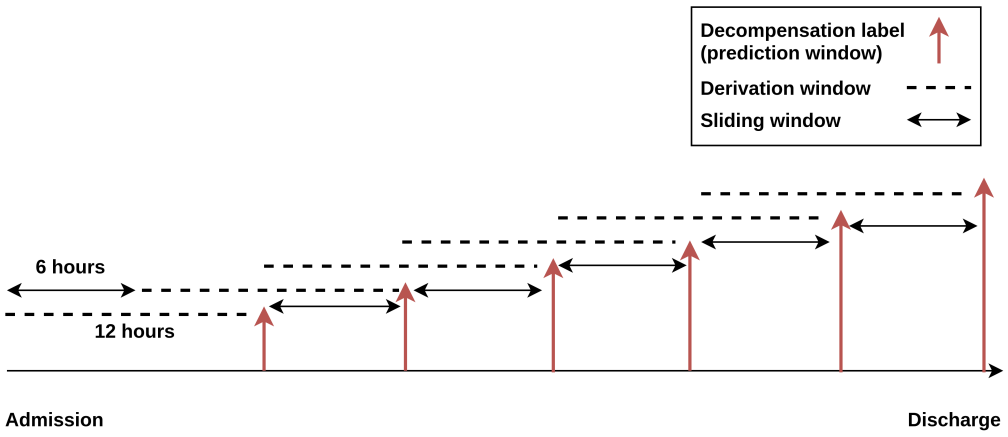


Figure 3.5: *Decompensation prediction schematic*

Data	Model	Num.	Cat.	Repn.	AUROC% (95% CI)	AUPRC% (95% CI)	Spec.% (95% CI)	Sens.%	PPV% (95% CI)	NPV% (95% CI)
In ICU unit	LR	✓	✓	EMB	78.99 (77.06 - 80.95)†	22.78 (19.88 - 25.67)†	43.92 (40.02 - 47.82)†	90.00	49.44 (35.70 - 63.19)	95.08 (94.64 - 95.52)
	ANN	✓	✓	EMB	83.57 (82.31 - 84.87)	27.86 (24.95 - 30.78)	**54.08 (49.75 - 58.41)**	90.00	48.55 (42.14 - 54.97)	95.01 (94.71 - 95.32)
	BiLSTM	✓	✓	EMB	**83.78 (81.86 - 85.73)**	**30.79 (28.09 - 33.49)**	53.36 (45.92 - 60.80)	90.00	51.83 (43.80 - 59.86)	**95.15 (94.83 - 95.46)**
	BiLSTM	×	✓	EMB	78.17 (75.80 - 80.56)†	20.87 (19.45 - 22.23)†	42.27 (37.33 - 47.21)†	90.00	43.55 (38.21 - 48.89)‡	95.08 (94.63 - 95.52)
	BiLSTM	✓	×	×	72.96 (71.05 - 74.92)†	20.00 (18.91 - 21.10)†	29.05 (23.02 - 35.08)†	90.00	**53.58 (38.43 - 68.72)**	94.93 (94.61 - 95.25)

Table 3.8: Decompensation risk prediction in eICU-CRD in a rolling-window manner
If the differences between *BiLSTM* result and other models (LR, ANN, and BiLSTM with varying data representation) is statistically significant on a two-tailed t-test then it is indicated with †, ‡ († $p < 0.05$, ‡ $p < 0.1$). The best-performing metric values are represented in **bold** font.

3.5 Discussion

In this study we have described four standardised benchmarks in machine learning for critical care research. Our definition of benchmark tasks is consistent with previously published benchmarks to facilitate comparison with already published results. However, in this work we focus on the more

recent eICU-CRD, where clinical data has been collected from 335 ICUs across 208 hospitals across the United States. Our dataset contains a larger number of patients and a wider range of patient groups, in comparison to benchmarks published using a single center dataset, which should result in lower systematic bias and increased generalisability of the study.

3.6 Conclusion

We provided a set of baselines for our benchmarks and show that BiLSTM model outperforms clinical gold standard as well as the baseline models. Of note is the impact of entity embedding of categorical variables in further improving the performance of our LSTM-based model. Clearly, interpretability remains a significant challenge of models based on deep neural networks, including our BiLSTM model. However, there has been significant progress in "opening the black box" [112] as demonstrated by a recently updated review of interpretability methods [113], bringing these models one step closer to clinical practice. As our work is meant to track the progress of machine learning in critical care, interpretability is certainly an important aspect of this progress. We believe that our work will provide a solid basis to further improve critical care decision making and we provide the source code for other researchers that wish to replicate our experiments and build upon our results.

In the next chapter, we provide a comprehensive study related to the prediction of delirium in critically ill patients in ICU which employs two publicly available datasets such as eICU-CRD and MIMIC-III.

Chapter 4

Delirium Prediction in the ICU

Delirium adversely affects both short and long-term patient outcomes. Current methods of identifying patients at risk of delirium are based on questionnaires with moderate accuracy or more advanced predictive models require hundreds of variables. This chapter proposes developing a delirium prediction model using machine learning models by using a limited number of commonly available variables. Studying the linear or non-linear relationship between these variables over temporal windows using machine learning models, we aim to predict the risk of delirium in the varying observation and prediction windows. Moreover, we rank the variables based on the variables' predictive influence to help the caregivers with interpretability, and this should help to facilitate for implementation of a screening tool to help caregivers at the bedside.

Delirium occurrence is common but preventive strategies are resource intensive. Screening tools can prioritize patients at risk. Using machine learning we can capture time and treatment effects that pose a challenge to delirium prediction. We aim to develop a delirium prediction model that

can be used as a screening tool. We retrospectively studied patients from the eICU Collaborative Research Database (eICU-CRD) and the Medical Information Mart for Intensive Care version III (MIMIC-III) database. The eICU-CRD contains 200,859 intensive care unit (ICU) admissions collected from 335 ICUs across 208 hospitals in the US from 2014 to 2015. The MIMIC-III database contains 53,432 adult ICU admissions between 2001-2012.

Patients with at least one Confusion Assessment Method -Intensive Care Unit (CAM-ICU) value and ICU length of stay greater than 24 hours were included in our study. We compared 21 quantitative clinical parameters between CAM-ICU-positive and CAM-ICU-negative patients to validate our predictive model. We assessed performance of models built using multiple machine learning methods against multiple observation and prediction windows.

We evaluated 16546 and 6294 patients from eICU-CRD and MIMIC-III databases respectively. Performance was best in BiLSTM models where, precision and recall changed from 37.52% (95% CI 36.00%-39.05%) to 17.45 (95%CI 15.83%-19.08%) and 86.1% (95% CI 82.49%-89.71%) to 75.58%(95% CI 68.33%-82.83%) respectively as prediction window increased from 12 to 96 hours. After optimizing for higher recall, precision and recall changed from 26.96% (95% CI 24.99%-28.94%) to 11.34% (95% CI 10.71%-11.98%) and 93.73% (95% CI 93.1%-94.37%) to 92.57% (95% CI 88.19%-96.95%) respectively. Similar results were obtained in the MIMIC-III cohort. Our model performed comparably to contemporary models using fewer variables. Using techniques like sliding windows, modification of threshold

to augment recall and feature ranking for interpretability helped address shortcomings of current models and build a model suitable for a screening tool.

4.1 Introduction

The diagnosis of delirium is common in critically ill patients and depending on the patient population its incidence can be up to 80% [114]. Delirium leads to increased hospital length of stay and need for prolonged institutionalization for critically ill patients [115, 116, 117]. Delirium drives up healthcare costs, and its impact often persists beyond the intensive care unit (ICU) including risk for functional decline in daily living activities, and long-term cognitive impairment [118, 119, 120, 121, 122]. Treatment and prevention of delirium is dependent on identifying the complex interplay of multiple triggers in the ICU [123]. A multimodal strategy of evidence-based best-practice recommendations aimed at coordinating multidisciplinary care to reduce delirium risk and expedite ICU discharge commonly referred to as the ABCDEF bundle is effective in both preventing and treating delirium [124, 125]. Unfortunately this bundle of interventions requires education of caregivers, coordination between a multidisciplinary team, is labor and resource intensive, and therefore not consistently implemented across all ICU patients and all health care settings [124, 126].

A screening tool to prioritize ABCDEF implementation to those who are most vulnerable can be an invaluable tool to maximize the benefit of the resource-intensive preventive measures. Current assessment tools, such as

the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU), only diagnoses delirium after its onset [127]. Although certain patient characteristics, such as age, illness severity, and certain medications, are considered high risk for development of delirium or while elevations in inflammatory bio-markers possibly associated with severe disease, these risk factors have been inconsistent in their ability to predict the onset of delirium [128, 129, 130].

Previous prediction models trained on small patient cohorts lacked adequate power to capture the complex relationships between delirium and the time-varying predictor variables [131, 132]. To improve accuracy larger administrative datasets were used to develop prediction models, using several hundred variables, but these models lack interpretability, and are almost impossible to adopt in day-to-day practice (PMID 30646095) [133]. Additionally, most of these models are not specific to the critically ill population and cannot be extrapolated to the ICU [132]. We propose to build an ABCDEF screening tool by developing and fine tuning a delirium prediction model that requires fewer variables than existing models and is able to predict the risk of delirium in a continuous fashion using a sliding window. Using both conventional machine learning methods and deep learning algorithms we will evaluate performance of our model across various observation and prediction windows to address the issues of variability across time and treatment effects. In addition, we will rank the independent variables in order of their predictive importance to help with interpretability. These attributes should help pave the way for implementation of a screening tool to help caregivers at the bedside.

4.2 Methodology

4.2.1 Ethical Review

The data in the MIMIC-III is de-identified, and the institutional review boards of the Massachusetts Institute of Technology (No. 0403000206) and Beth Israel Deaconess Medical Center (2001-P-001699/14) both approved the use of the database for research. The analysis using the eICU-CRD is exempt from institutional review board approval due to the retrospective design, lack of direct patient intervention, and the security schema, for which the re-identification risk was certified as meeting safe harbor standards by an independent privacy expert (Privacert, Cambridge, MA) (Health Insurance Portability and Accountability Act Certification no. 1031219-2).

4.2.2 Study Population

The eICU-CRD is a freely available multicenter database comprising 200,859 patient unit encounters for 139,367 unique patients admitted between 2014 and 2015 in over 200 hospitals located throughout the US [134]. The MIMIC-III database is an open-access single-center ICU database including 53,423 distinct hospital admissions for 46,476 unique patients admitted from 2001 to 2012 [135]. Both datasets comprise data on patient demographics, vitals, clinical flowsheets, laboratory values, medications, interventions, and outcomes.

Any patient admitted to the ICU for 24 hours or more and with at least one CAM-ICU assessment was included in our study population Figure

4.1.

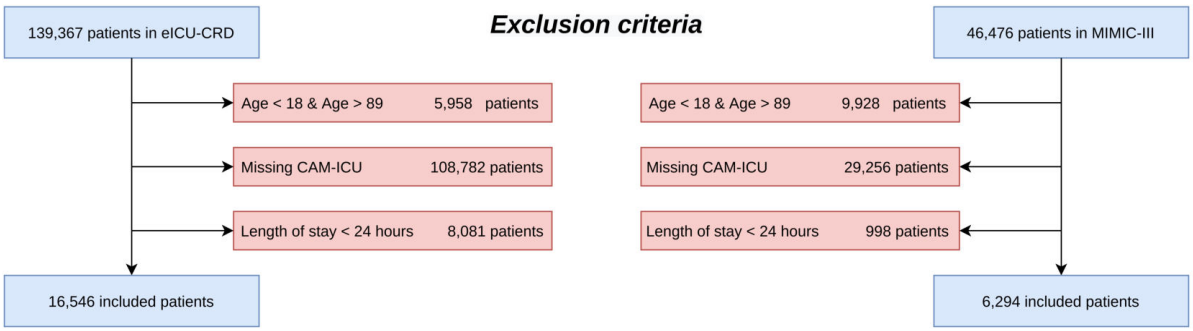


Figure 4.1: *Cohort selection criteria*

4.2.3 Delirium Assessment

We evaluated the ability of our model to predict delirium at 12 hours, 24 hours, 48 hours, 72 hours and 96 hours after observing for 12 hours, 24 hours or 48 hours. The diagnosis of delirium was made when at least one CAM-ICU value was positive [127]. In instances with multiple CAM-ICU assessments, onset of delirium was determined from the time of the first positive CAM-ICU.

4.2.4 Variable Selection

The rationale for selection of independent variables was based on their ability to predict delirium in prior literature, availability in our databases, ease of extracting and monitoring in a real-time environment. We identified 21 categorical or numerical variables classified into demographic data, vital signs, laboratory values, and vasopressor dose that fulfilled above criteria [136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147]. We also cal-

culated daily sequential organ failure assessment (SOFA) scores to provide overall patient status. Since admission diagnoses or past medical history were not consistently available in the applied datasets, we excluded them. Downstream variables such as outcomes would not be available in real-time and similarly excluded. Initiation of delirium therapies like antipsychotic drugs could be a reaction to onset of delirium, and hence excluded to avoid confounding. Table 5.2 lists all the variables used.

Variable group	**Variable name**
Demographic data	age, gender, height, weight
Vita signs	oxygen saturation (SpO_2), heart rate (HR), temperature
Other Measurements	sofa, sofa without GCS, Ventilation
Laboratory Measurements	white blood cell count (WBC), sodium (Na), blood urea nitrogen (BUN), glucose, hemoglobin, platelets, potassium, chloride, bicarbonate, creatinine
Medications as continuous drips	Dopamine, epinephrine, norepinephrine, phenylephrine (all calculated as norepinephrine equivalent)

Table 4.1: *Variables included in the prediction models*

4.2.5 Data Pre-processing

All variables were aggregated into hourly intervals, where the last recorded value was used as a candidate for that interval. In cases where the last value for each variable is not measured in the interval, the representative of that interval was computed by averaging the available measurements in the interval. Missing values that were collected hourly like vital signs were imputed by forward and backward imputation. Categorical variables were

converted into a vector in order to capture the semantics of each category at the model derivation phase. For all continuous variables, we utilized the recorded value in the database without any adaptation. As depicted in Table 4.2, heat-map further details the set of variables, including linear correlations between each variable.

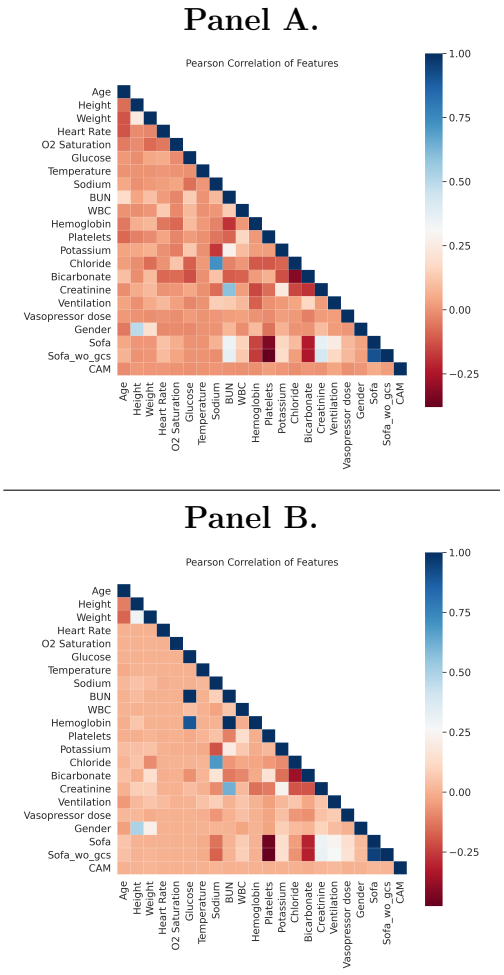


Table 4.2: *Heat-map showing correlation between variables. Blue shows strong positive correlation, Red shows strong negative correlation. Panel A: eICU-CRD, Panel B: MIMIC-III*

4.2.6 Model Derivation and Validation

We evaluated the results based on 5-fold stratified cross-validation. This method divides data into 5 folds where in each fold there exist 5 subsets of data. Four of these subsets are considered a derivation set and one subset is considered as a mutually exclusive validation set. Typically, metrics calculated based on the k-fold stratified cross-validation can effectively assess overfitting and has lower variance [148].

We used 3 sets of algorithms to evaluate delirium prediction, namely Logistic Regression (LR), Random Forest (RF), and Bidirectional Long Short-Term Memory (BiLSTM). BiLSTM represents an evolution from recurrent neural network-based LSTM, and with a backward input, preserves information from both past and future. This produces more accurate predictions [149, 150].

Considering that both LR and RF are unable to process time series variables efficiently, we pre-processed the clinical variables and all time steps and corresponding variables were flattened into a single record. This was done to ensure that both LR and RF have access to the same data about the changes in patient state as BiLSTM, to ensure a fair performance comparison.

4.2.7 Statistical Analysis

The classification results for delirium prediction are reported using the Area Under Receiver Operating Characteristic (AUROC), Area Under Precision Recall Curve (AUPRC), Precision and Recall.

4.2.8 Model Interpretability

Although there are many definitions of interpretability, we focused on how the model ranks each input variable with respect to outcome prediction. LR and RF have been successfully employed in the clinical domain due to their ease of interpretation, but they perform poorly on large, high-dimensional, longitudinal and irregular EHR datasets [151].

In this context, we employed the Shapley Value Sampling (SVS) method to probe the Bi-LSTM model [152]. SVS is a perturbation-based method to compute variable attribution, which is based on sampling theory that can be used to estimate Shapley values [40]. The SVS produces feature ranking with respect to each feature input, allowing us to rank these variables based on their predictive power. Given that interpretability of neural networks is still an open research question, especially for temporal neural networks, we also provide results from two other methods, namely Integrated Gradient (IG) and Guided Backpropagation (GB), to ensure that the variable importance results are consistent across the three methods [153, 154, 42].

4.2.9 Source Code

We made use of several open-source libraries based on Python to conduct our experiments; machine learning framework Scikit-learn [109] and DL framework Pytorch [110]. The entire source code detailing our methods and experiments is publicly available at this GitHub repository such that anyone can replicate our results and build upon our work.

4.3 Results

4.3.1 Patient characteristics

The eICU-CRD cohort consisted of 16,546 patients, with a mean age of 62.84 (±16.02) years and 46.53% were female. The incidence of delirium was 19.06% as shown in Table 5.1. In the first 48 hours of admission, 59.30% of patients presented with delirium. The MIMIC-III cohort consisted of 6,294 patients, with a mean age of 63.58 (±15.79) years and 43.82% were female. The incidence of delirium was 20.15% and 66.34% of patients presented within the first 48 hours of ICU admission. For vital signs and laboratory values that were generated hourly, an average of approximately 7% were missing values.

Variables	eICU			MIMIC		
	CAM-ICU +	CAM-ICU -	p value	CAM-ICU +	CAM-ICU -	p value
Number of patients	3153	13393	—	1268	5026	—
Age, mean (SD), years	65.53 (15.14)	62.20 (16.16)	< 0.05	64.81 (15.62)	63.27 (15.82)	< 0.05
Female (%)	1405 (44)	6295 (47)	—	545 (43)	2211 (44)	—
Height, mean (SD), m	168.47 (18.23)	169.25 (15.90)	< 0.05	170.06 (14.22)	168.88 (14.87)	0.054
Weight, mean (SD), kg	83.06 (29.88)	85.00 (25.58)	< 0.05	82.68 (30.25)	81.53 (24.89)	0.15
Heart Rate, mean (SD), bpm	88.22 (18.06)	85.09 (17.73)	< 0.05	88.60 (17.53)	85.12 (17.29)	< 0.05
Oxygen Saturation, mean (SD), %	97.16 (2.72)	96.80 (2.79)	< 0.05	97.17 (2.71)	96.58 (4.50)	< 0.05
Glucose, mean (SD), mg/dL	140.32 (45.97)	146.46 (56.31)	< 0.05	144.51 (58.70)	141.25 (51.43)	< 0.05
Temperature, mean (SD), °C	37.01 (0.69)	36.97 (2.65)	< 0.05	37.06 (0.76)	36.88 (0.76)	< 0.05
Serum Sodium, mean (SD), mEq/L	140.32 (5.80)	138.57 (5.04)	< 0.05	139.39 (5.48)	138.32 (4.89)	< 0.05
BUN, mean (SD), mg/dL	31.93 (22.10)	25.88 (18.64)	< 0.05	33.96 (24.46)	28.10 (20.77)	< 0.05
WBC, mean (SD), per microliter	13.01 (6.47)	11.08 (5.51)	< 0.05	12.13 (7.73)	10.74 (6.29)	< 0.05
Hemoglobin, mean (SD), g/dL	9.73 (1.89)	10.00 (2.08)	< 0.05	9.76 (1.68)	10.27 (1.76)	< 0.05
Platelets, mean (SD), per microliter	201.34 (122.76)	210.23 (108.70)	< 0.05	202.59 (137.23)	199.53 (114.33)	< 0.05
Serum Potassium, mean (SD), mEq/L	3.98 (0.59)	4.00 (0.57)	0.1431	4.03 (0.57)	4.07 (0.56)	< 0.05
Chloride, mean (SD), mEq/L	105.54 (6.86)	103.24 (6.29)	< 0.05	104.57 (6.69)	104.36 (6.37)	< 0.05
Serum Bicarbonate, mean (SD), mEq/L	35.23 (5.02)	25.52 (5.02)	< 0.05	25.16 (5.21)	24.88 (4.95)	< 0.05
Serum creatinine, mean (SD), mg/dL	1.45 (1.16)	1.37 (1.21)	< 0.05	1.63 (1.28)	1.37 (1.05)	< 0.05
Ventilation, mean (SD)	0.87 (0.34)	0.71 (0.45)	< 0.05	0.56 (0.50)	0.33 (0.47)	< 0.05
Total norepinephrine dose (SD), mcg/kg/min	0.02 (0.31)	0.01 (0.28)	< 0.05	0.08 (0.63)	0.06 (0.57)	< 0.05
SOFA, mean (SD)	4.9 (3.3)	3.42 (2.84)	< 0.05	6.46 (3.77)	6.67 (3.34)	< 0.05
SOFA without GCS, mean (SD)	3.27 (2.83)	2.58 (2.33)	< 0.05	5.42 (3.65)	4.99 (3.13)	< 0.05

Table 4.3: *Characteristics of the included patients divided by the CAM-ICU status*

4.3.2 Performance of Machine Learning Models

The BiLSTM algorithm was noted to have had the highest AUROC and AUPRC values for most of the observation-prediction combinations. With 24 hour observation of the eICU-CRD cohort and 48 hour prediction, the AUROC of BiLSTM model was 84.87% (95% CI, 83.32%-86.41%), LR 82.57% (95% CI, 79.64%-85.47%) and RF 83.24% (95% CI, 81.83%-84.67%), and AUPRC of 34.97% (95% CI,32.22%-37.27%), 31.07% (95% CI, 27.62%-33.81%) and 32.82% (95% CI, 28.89%-36.75%) respectively (Figure 4.4).

As shown in Figure 4.4, BiLSTM outperforms LR and RF by a small margin when the observation window is small (12 hours); however, the observation window has a direct effect on this difference margin between BiLSTM, LR, and RF. This margin increases as the observation window increases, meaning by providing more data sequences to BiLSTM, the performance margin increases among BiLSTM, LR, and RF.

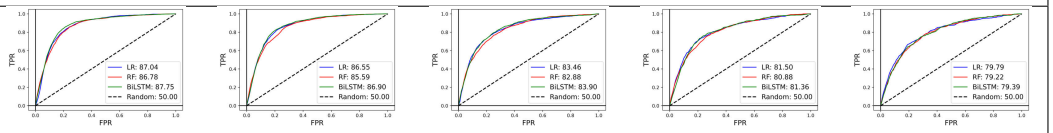
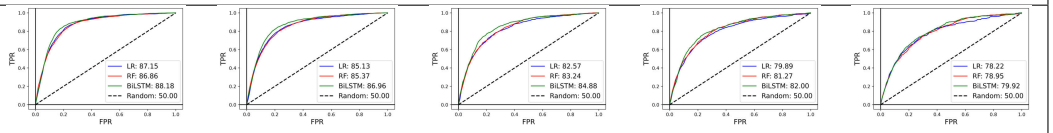
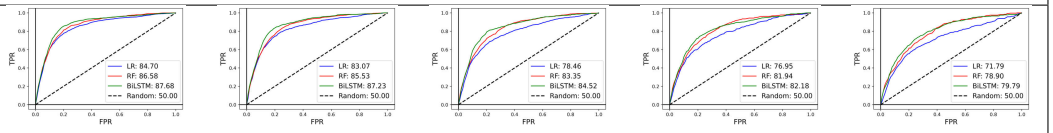
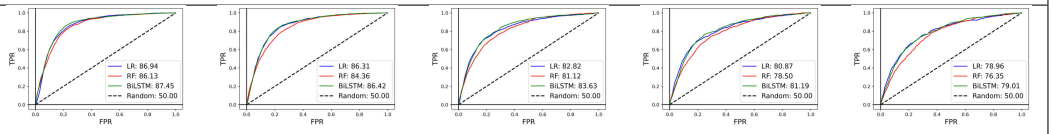
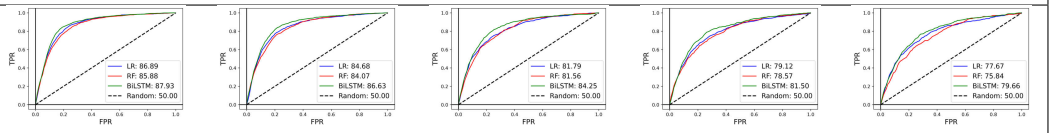
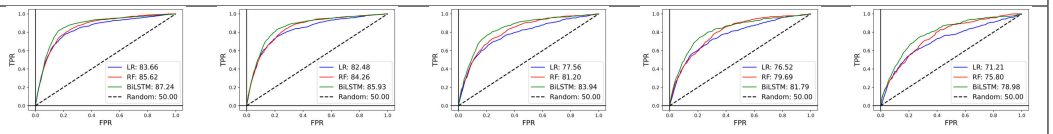
Obs. \ Pred.	12h	24h	48h	72h	96h
Panel A.	**unmodified thresholds**				
12h					
24h					
48h					
Panel B.	**thresholds optimized for higher recall**				
12h					
24h					
48h					

Table 4.4: *AUROC Graphs for Machine Learning Models on eICU-CRD - Model derived and validated using Cross-validation. Panel A: unmodified thresholds, Panel B: thresholds optimized for higher recall*

Since BiLSTM had the best AUROCs and AUPRCs, we calculated the precision and recall values in each observation-prediction window using BiLSTM. In the eICU-CRD derivation cohort, for the 12 hour observation window, the precision and recall decreased from 37.52% (95% CI,36.00%-39.05%) to 28.68% (95% CI, 24.88%-32.49%) and from 86.1% (95% CI, 82.49%-89.71%) to 63.49% (95% CI, 52.91%-74.08%) respectively when the prediction window changed from 12 hour to 96 hours (Table 4.5 and Table 4.6). When increasing the observation window for 48 hour prediction, the precision and recall changed from 32.82% (95% CI, 29.6%-36 .04%) to 17.9% (95% CI, 15.37%-20.44%) and from 82.22% (95% CI, 78.16%-86.27%) to 73.95% (95% CI, 64.8%-83.11%).

As we were interested in making our model more sensitive for screening, we changed thresholds to have higher recall at the expense of precision. For a 12 hour observation window, while recall changed slightly from 93.73% (95% CI, 93.1% - 94.37%) to 92.57% (95% CI, 88.19%-96.95%) as the prediction window changed from 12 hour to 96 hours, the precision decreased from 26.96% (95% CI, 24.99%-28.94%) to 11.34% (95% CI, 10.71%-11.98%) (Table 4.6). For the 48 hour prediction window as we increased the observation window from 12 hours to 48 hours, the precision and recall changed from 16.82% (95% CI, 15.61%-18.02%) to 15.64% (95% CI, 13.96%-17.42%) and 92.15% (95% CI, 88.47%-95.82) to 91.13% (95% CI, 89.57%- 92.69%) respectively. Similar results for the MIMIC-III cohort are presented in (Table 4.9, Figure 4.7 and Figure 4.8). A heat map demonstrating correlation among features is presented in Table 4.2 for the eICU-CRD and MIMIC-III populations.

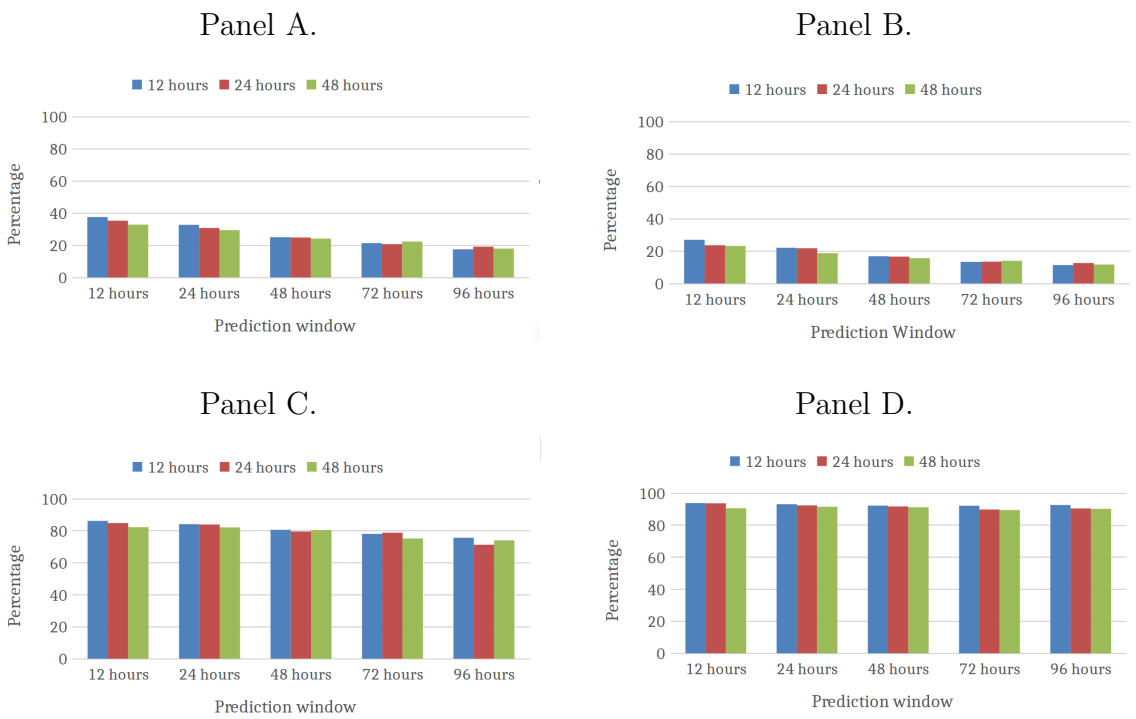


Table 4.5: *Comparison of Precision and Recall in different observation and prediction windows in the eICU cohort. Panel A. Precision with unmodified thresholds, Panel B: Precision with thresholds adjusted for higher recall, Panel C. Recall with unmodified thresholds, Panel D: Recall with thresholds adjusted for higher recall*

Pred. / Obs.	12 hours	24 hours	48 hours	72 hours	96 hours
A.	**Unmodified thresholds**				
	AUROC				
12 hours	87.82 (87.17-88.30)	86.82 (85.15-88.64)	84.00 (81.68-86.13)	81.45 (78.61-84.10)	79.03 (76.69-82.11)
24 hours	88.39 (86.41-89.96)	86.68 (85.79-88.15)	84.87 (83.32-86.41)	81.99 (80.66-83.38)	79.93 (76.57-83.34)
48 hours	88.00 (75.78-89.59)	87.23 (86.30-88.20)	84.51 (82.14-86.92)	82.19 (80.99-83.41)	79.78 (75.37-84.25)
	AUPRC				
12 hours	46.86 (42.52-50.85)	40.92 (37.03-44.46)	34.04 (28.99-38.24)	26.78 (25.24-27.71)	24.90 (18.48-30.22)
24 hours	44.62 (39.11-50.02)	40.85 (38.38-43.10)	34.97 (32.22-37.27)	28.68 (23.78-33.02)	26.37 (21.00-31.28)
48 hours	41.67 (37.52-45.62)	39.64 (37.00-42.07)	33.35 (27.58-38.88)	29.75 (26.06-32.90)	26.43 (19.65-32.71)
	Precision				
12 hours	37.52 (36.00-39.05)	32.68 (29.09-36.28)	25.01 (22.73-27.28)	21.30 (20.09-22.49)	17.45 (15.83-19.08)
24 hours	35.27 (33.51-37.03)	30.69 (28.71-32.66)	24.84 (23.35-26.32)	20.69 (18.24-23.15)	19.08 (17.85-20.31)
48 hours	32.82 (29.60-36.04)	29.37 (25.18-33.56)	24.17 (21.68-26.67)	22.25 (18.85-25.66)	17.90 (15.37-20.44)
	Recall				
12 hours	86.10 (82.49-89.71)	84.09 (81.81-86.37)	80.53 (76.76-84.30)	77.96 (69.87-86.05)	75.58 (68.33-82.83)
24 hours	84.74 (81.57-87.90)	83.87 (81.24-86.50)	79.44 (75.53-83.35)	78.73 (72.41-85.05)	71.20 (61.95-80.45)
48 hours	82.22 (78.16-86.27)	82.06 (78.55-85.56)	80.38 (75.53-85.24)	75.11 (67.85-82.37)	73.95 (64.80-83.11)
B.	**Threshold optimized favoring a higher recall**				
	AUROC				
12 hours	87.45 (86.87-88.03)	86.41 (84.12-88.71)	83.63 (81.43-85.83)	81.19 (78.49-83.89)	79.01 (76.10-81.92)
24 hours	87.93 (86.39-89.48)	86.63 (85.41-87.86)	84.25 (82.92-85.62)	81.50 (80.13-82.90)	79.66 (76.61-82.72)
48 hours	87.24 (85.34-89.15)	85.93 (84.29-87.60)	83.94 (81.72-85.90)	81.76 (81.03-82.61)	78.99 (74.90-83.01)
	AUPRC				
12 hours	46.63 (42.17-50.93)	39.52 (34.88-43.89)	33.21 (28.84-36.83)	26.55 (23.55-28.97)	24.30 (18.52-29.21)
24 hours	44.55 (39.40-49.02)	39.95 (38.25-41.47)	33.70 (30.96-36.07)	27.49 (22.86-31.46)	26.11 (22.02-29.85)
48 hours	40.96 (36.55-44.72)	36.98 (32.38-41.33)	32.12 (26.50-37.45)	29.55 (25.26-33.35)	24.65 (15.60-33.20)
	Precision				
12 hours	26.96 (24.99-28.94)	22.04 (20.66-23.42)	16.82 (15.61-18.02)	13.33 (13.03-13.60)	11.34 (10.71-11.98)
24 hours	23.61 (22.55-24.66)	21.73 (20.63-22.83)	16.57 (15.74-17.38)	13.46 (12.29-14.62)	12.60 (11.81-13.39)
48 hours	23.18 (20.49-25.87)	18.70 (14.49-22.87)	15.64 (13.96-17.42)	14.02 (12.06-16.04)	11.69 (10.75-12.73)
	Recall				
12 hours	93.73 (93.10-94.37)	93.08 (90.42-95.75)	92.15 (88.47-95.82)	92.08 (90.25-93.91)	92.57 (88.19-96.95)
24 hours	93.59 (91.69-95.48)	92.29 (88.83-95.76)	91.65 (89.07-94.23)	89.72 (86.74-92.69)	90.40 (88.58-92.23)
48 hours	90.49 (86.48-94.50)	91.46 (89.97-92.95)	91.13 (89.57-92.69)	89.37 (84.87-93.41)	90.20 (82.79-97.61)

Table 4.6: *Performance metrics of derived model in eICU-CRD cohort, metrics are reported in percentage with (95 %CI). Panel A. Unmodified thresholds. Panel B. After thresholds were optimized favoring higher recall.*

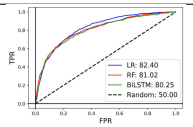
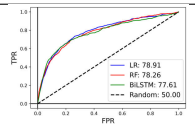
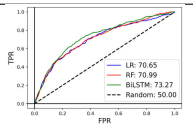
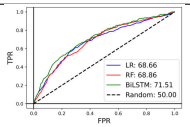
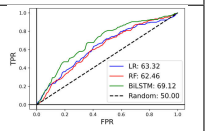
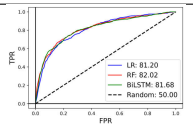
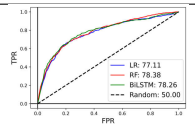
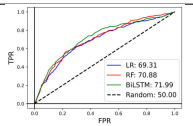
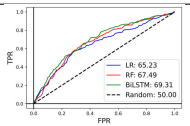
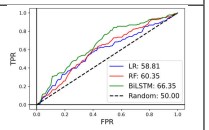
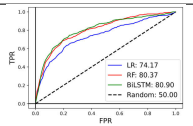
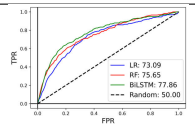
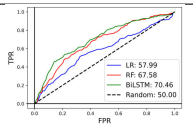
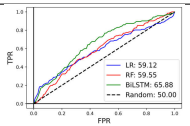
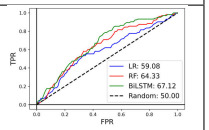
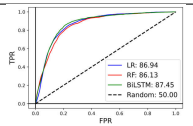
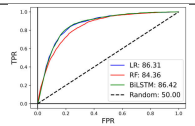
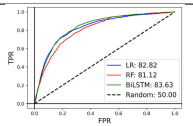
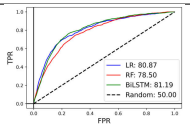
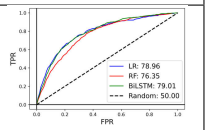
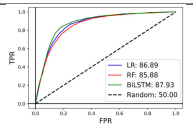
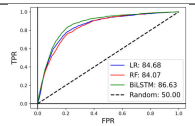
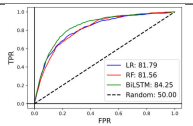
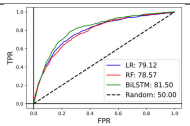
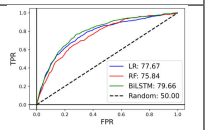
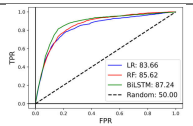
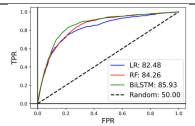
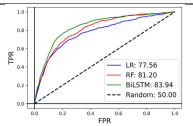
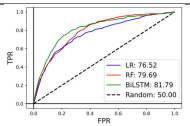
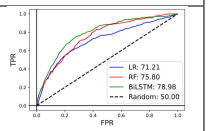
Pred. / Obs.	12h	24h	48h	72h	96h
Panel A.	**unmodified thresholds**				
12h					
24h					
48h					
Panel B.	**thresholds optimized for higher recall**				
12h					
24h					
48h					

Table 4.7: *AUROC Graphs for Machine Learning Models on MIMIC-III - Model derived and validated using Cross-validation. Panel A: unmodified thresholds, Panel B: thresholds optimized for higher recall*

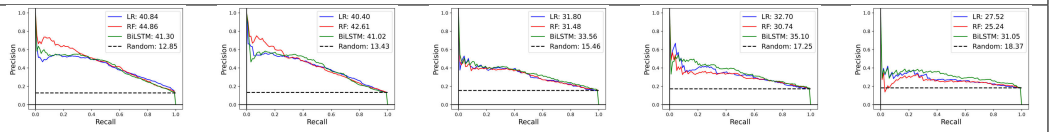
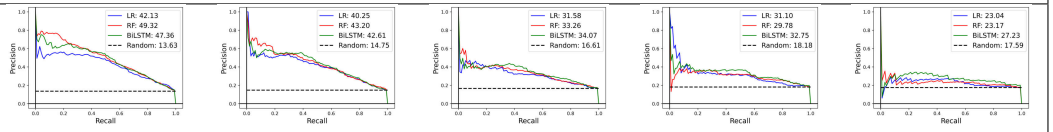
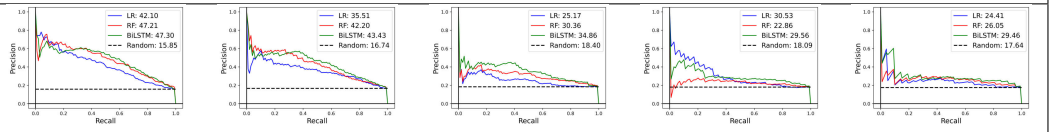
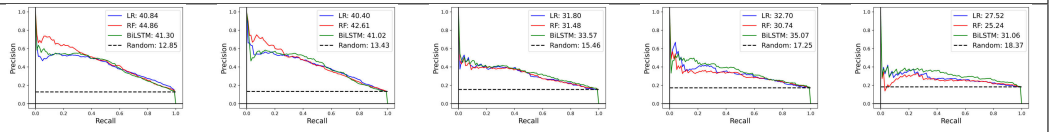
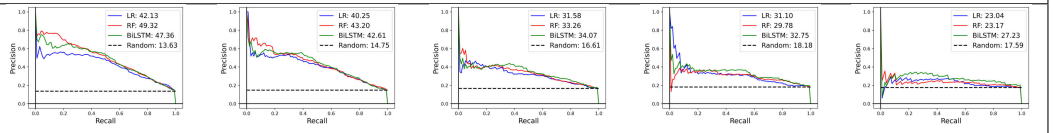
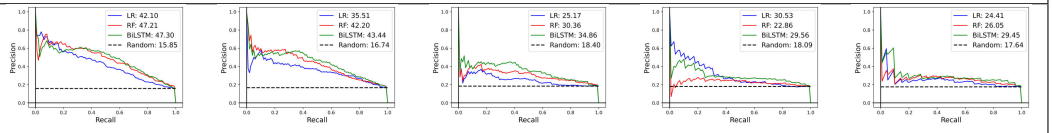
Obs. \ Pred.	12h	24h	48h	72h	96h
Panel A.	**unmodified thresholds**				
12h					
24h					
48h					
Panel B.	**thresholds optimized for higher recall**				
12h					
24h					
48h					

Table 4.8: *AUPRC Graphs for Machine Learning Models on MIMIC-III - Model derived and validated using Cross-validation. Panel A: unmodified thresholds, Panel B: thresholds optimized for higher recall*

4.3.3 Interpretability

As discussed earlier, we employed SVS, IG, and GB interpretable methods to probe the Bi-LSTM model [152, 153, 154, 42]. Table 4.10 depicts how each of the interpretable methods ranks the features that have contributed to delirium prediction according to their relative importance in the eICU-CRD derived model. In this regards, clinical features such as ventilation, heart rate, age, white blood cell count, SOFA score and vasopressor use are the highest ranked features across different prediction windows. Most of these features are also the highest ranked features when assessing interpretability in the MIMIC III cohort as depicted in Table 4.11.

4.4 Discussion

Our study shows that a machine learning model using only a few routine clinical variables replicated the performance of previously reported models that were developed using hundreds of variables. Our study successfully demonstrated that we can modify the performance of a model to fit our clinical needs as an effective screening tool. We took the following steps that helped us achieve our goal: 1.) we studied the peak delirium onset time in our population and optimized the model to maximize predictive accuracy in that time frame; 2.) we incorporated sliding windows in our model for continuous prediction across time and address drop in performance associated with predictions further ahead; and 3.) We adjusted our thresholds to favor a high recall to ensure the model detects all patients at risk of delirium. Furthermore, we demonstrated that performance across

different datasets diminish in accuracy and needs to be individualized to the population. Our features when ranked suggest older and more critically ill patients are at greater risk of delirium, especially in combination with mechanical ventilation and vasopressor therapy. Our model's ranking of features is consistent with what we already know as high-risk features. We have shared our code for replication and for adjustments based on specific needs.

Screening tools like CAM-ICU describe a snapshot in time and do not give an idea of the patient's progress nor are predictive. Strategies based on established best practices such as ABCDEF are resource intensive and challenging to implement universally [124]. Despite effective prevention strategies, delirium is still commonplace in the ICU highlighting a need for a screening tool that prioritizes patients at risk. Few models exist that can both accurately predict and be easy to implement. Most models use several hundred variables or use only a snapshot of features that can vary with time. Also, many of these models were trained on small datasets and use inconsistent approaches for collecting and/or stratifying data into training and validation cohorts limiting generalizability [131, 133]. The Pre-Deliric and e-Pre-Deliric, were built with a handful of predictor variables from a large patient cohort, and been externally validated [155, 156]. However, they employ data from admission variables that change and lose predictive power with time. Recent machine learning based algorithms were able to predict delirium accurately, but using over 700 predictor variables [133]. Importantly, these models have not investigated how their performance changes with different prediction windows, optimal time of observation,

capture the evolution of a patient's state through time and unable to adjust delirium risk estimation temporally. In our knowledge, our report is one of the first instances of delirium prediction, where we have not only tried to predict accurately across different scenarios but also addressed the aforementioned issues with prior prediction models. Notably, we have iteratively developed our model to address the challenges that are posed by low incidence of delirium, temporal progression of disease and different patient populations. Additionally, we have ventured into the realm of explaining how our features contribute, something that is rare in models using deep learning.

The BiLSTM-based model, which has the advantage of capturing temporal dependencies, performed the best of the 3 models evaluated, suggesting that the trajectory of predictive features is more informative than a single value. A longer observation window gained little in terms of model performance. A 48 hour observation window even led to a drop in accuracy but this is likely due to a decrease in the size of the training cohort. Another possibility is that factors contributing to delirium are proximal to its onset, further justifying the use of continuous prediction using a sliding window. The decay in performance of the algorithm as it predicts delirium with longer lead time is similar in both MIMIC-III and eICU-CRD.

A screening tool needs to be highly sensitive. This is best addressed by a model with a high recall. We adjusted thresholds favoring a high recall while sacrificing precision (Table 4.6) to achieve this purpose. Also it is desirable to have prediction algorithms that have short observation duration and predict the furthest ahead. In our case, since most (59% in

eICU-CRD, 66% in MIMIC-III) delirium cases occurred within 48 hours of ICU admission as shown in Table 4.12 ,hence we targeted performance for a 48 hour prediction window with a 12 or 24 hour observation window. We also demonstrated that as the prediction window moved beyond 48 hours the model maintained recall, but with a precipitous drop in precision. Non-trivial tuning of hyper-parameters is required when algorithms are ported across populations. We suggest the performance of different observation and prediction times be studied on the local dataset and depending on the objective of the algorithm the optimal windows are determined.

Pred. / Obs.	12 hours	24 hours	48 hours	72 hours	96 hours
A.	**Unmodified thresholds**				
	AUROC				
12 hours	80.34 (78.31-82.21)	77.64 (75.92-79.28)	73.38 (69.43-77.15)	71.47 (66.24-76.77)	69.21 (63.95-74.41)
24 hours	81.72 (78.09-85.36)	78.25 (75.97-80.63)	72.14 (64.37-79.61)	69.06 (61.33-77.29)	66.26 (56.31-76.38)
48 hours	81.15 (79.46-82.30)	77.90 (74.96-80.84)	70.38 (64.35-76.59)	65.87 (58.36-73.44)	67.20 (61.93-72.46)
	AUPRC				
12 hours	41.61 (36.14-46.56)	40.97 (34.96-46.05)	33.52 (30.06-37.08)	34.93 (29.39-39.89)	31.16 (26.65-35.69)
24 hours	48.00 (43.11-52.94)	42.54 (36.27-48.58)	34.19 (27.39-40.66)	32.76 (24.17-41.03)	27.29 (19.35-34.66)
48 hours	48.08 (42.59-53.32)	43.48 (36.68-50.22)	34.15 (29.67-38.03)	28.02 (22.75-32.56)	29.33 (24.80-33.66)
	Precision				
12 hours	30.14 (26.54-33.74)	35.12 (31.85-38.39)	30.99 (27.91-34.07)	30.86 (26.90-34.82)	28.68 (24.88-32.49)
24 hours	34.07 (31.36-36.79)	33.35 (29.82-36.88)	30.21 (27.00-33.41)	28.36 (23.08-33.65)	24.71 (19.89-29.52)
48 hours	36.05 (32.37-39.74)	34.27 (32.22-36.32)	30.61 (28.27-32.95)	26.69 (21.06-32.32)	26.92 (22.57-31.26)
	Recall				
12 hours	71.75 (68.75-74.74)	64.80 (57.11-72.49)	65.36 (62.42-68.29)	62.91 (58.26-67.57)	63.49 (52.91-74.08)
24 hours	73.93 (67.53-80.32)	69.23 (66.58-71.89)	65.38 (59.77-70.99)	60.35 (49.13-71.57)	60.42 (46.04-74.80)
48 hours	74.35 (67.09-81.61)	70.00 (66.75-73.25)	64.04 (53.35-74.74)	60.69 (48.72-72.66)	64.00 (49.96-78.04)
A.	**Threshold optimized favoring a higher recall**				
	AUROC				
12 hours	80.25 (78.31-82.21)	77.61 (75.92-79.28)	73.27 (69.43-77.15)	71.51 (66.25-76.76)	69.12 (63.96-74.38)
24 hours	81.67 (78.09-85.36)	78.26 (75.97-80.63)	71.99 (64.37-79.62)	69.31 (61.33-77.28)	66.35 (56.31-76.38)
48 hours	80.89 (79.46-82.30)	77.87 (74.96-80.83)	70.47 (64.35-76.60)	65.86 (58.35-73.44)	67.09 (61.94-72.40)
	AUPRC				
12 hours	41.30 (36.12-46.04)	41.02 (35.64-45.60)	33.56 (29.61-37.58)	35.07 (29.90-39.51)	31.07 (26.63-35.49)
24 hours	47.35 (43.17-51.63)	42.61 (36.53-48.48)	34.07 (26.91-40.81)	32.74 (24.69-40.53)	27.24 (19.51-34.35)
48 hours	47.30 (43.15-51.42)	43.50 (36.80-50.05)	34.86 (30.14-39.06)	29.53 (21.64-36.90)	29.44 (24.71-33.94)
	Precision				
12 hours	20.98 (19.31-22.64)	23.78 (20.54-27.01)	21.67 (19.08-24.26)	23.27 (21.34-25.21)	23.30 (21.45-25.14)
24 hours	25.67 (24.48-26.86)	25.41 (21.16-29.65)	23.09 (21.71-24.47)	23.35 (21.12-25.57)	20.90 (18.66-23.15)
48 hours	28.08 (24.45-31.75)	26.67 (25.20-28.14)	24.57 (23.20-25.93)	22.51 (19.89-25.05)	23.70 (22.20-25.20)
	Recall				
12 hours	86.63 (83.32-90.01)	76.95 (73.08-80.82)	81.46 (71.94-90.98)	84.47 (77.72-91.22)	87.38 (73.19-99.05)
24 hours	82.22 (76.40-88.05)	81.14 (79.88-82.40)	87.36 (74.76-92.75)	84.11 (74.56-93.66)	86.14 (73.04-99.24)
48 hours	83.18 (76.24-90.13)	83.79 (76.92-90.66)	82.24 (71.18-93.30)	83.20 (70.13-96.27)	87.38 (78.40-96.36)

Table 4.9: *Performance metrics of derived model in MIMIC-III cohort, metrics are reported in percentage with (95 %CI). Panel A. Unmodified thresholds. Panel B. After thresholds were optimized favoring higher recall.*

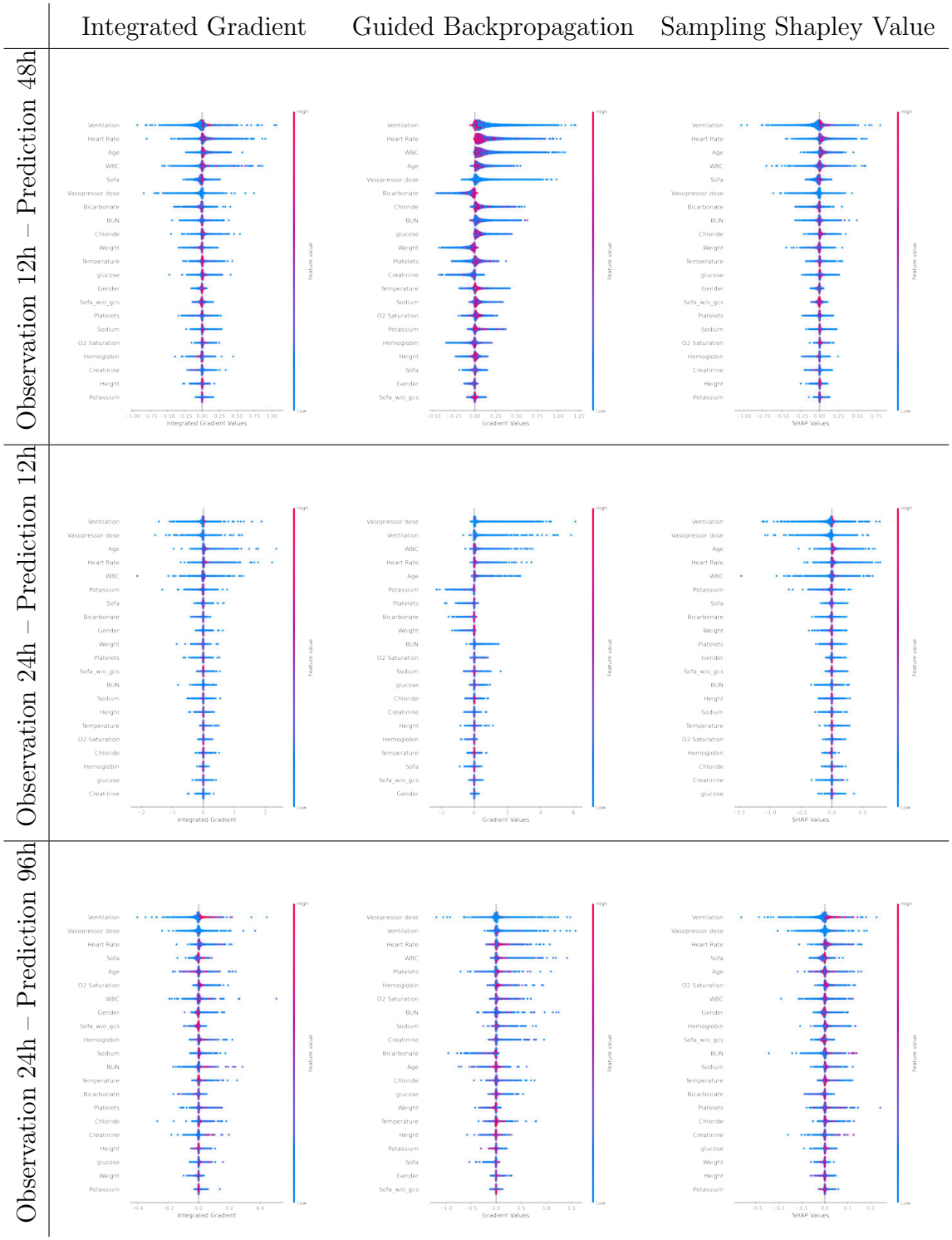


Table 4.10: *Interpreting features. Features ranked according to their importance in descending order in long short term memory model in eICU-CRD. Color shows whether ranked variable value is high (red) or low (blue) for that observation*

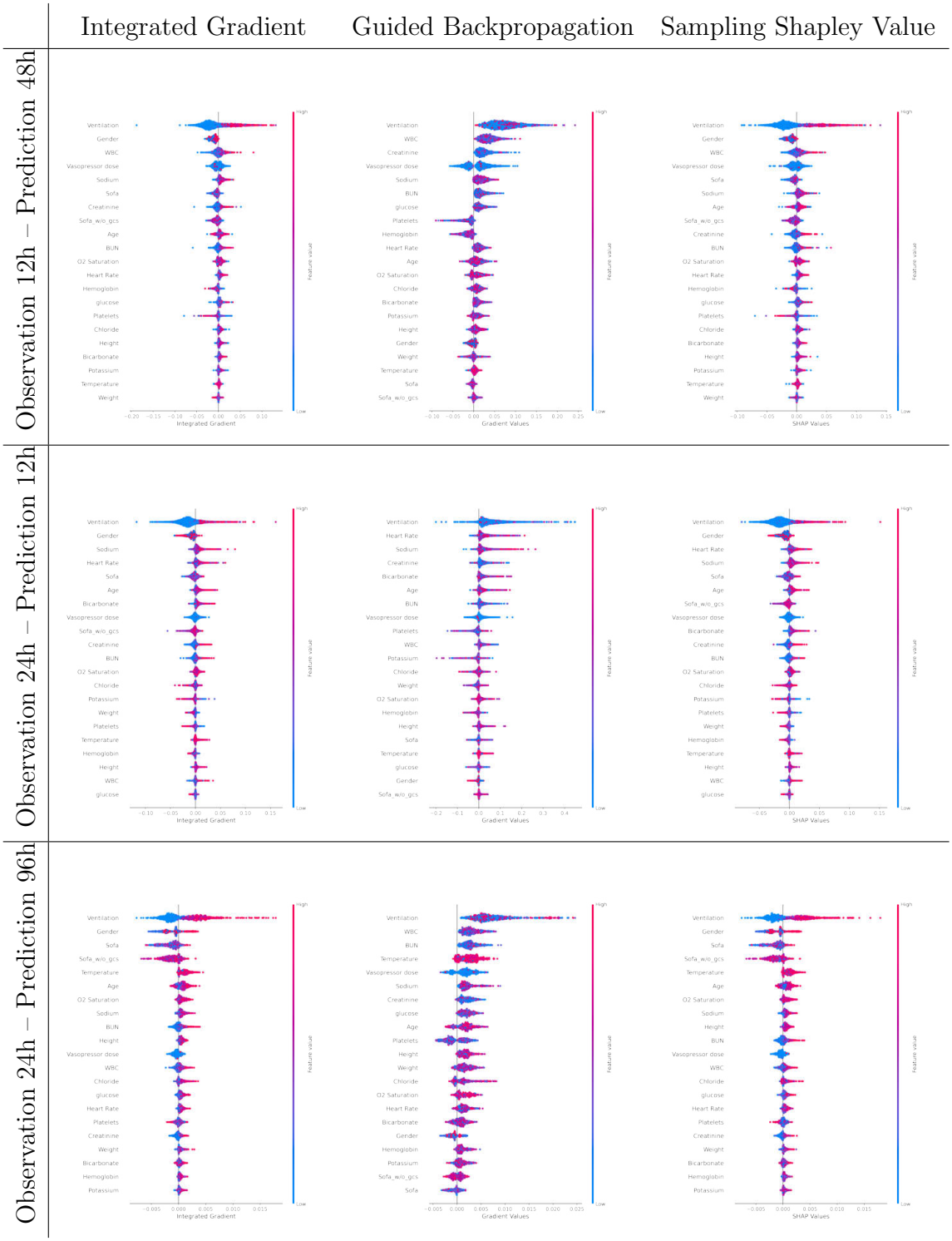


Table 4.11: *Interpreting features. Features ranked according to their importance in descending order in long short term memory model in MIMIC-III. Color shows whether ranked variable value is high (red) or low (blue) for that observation*

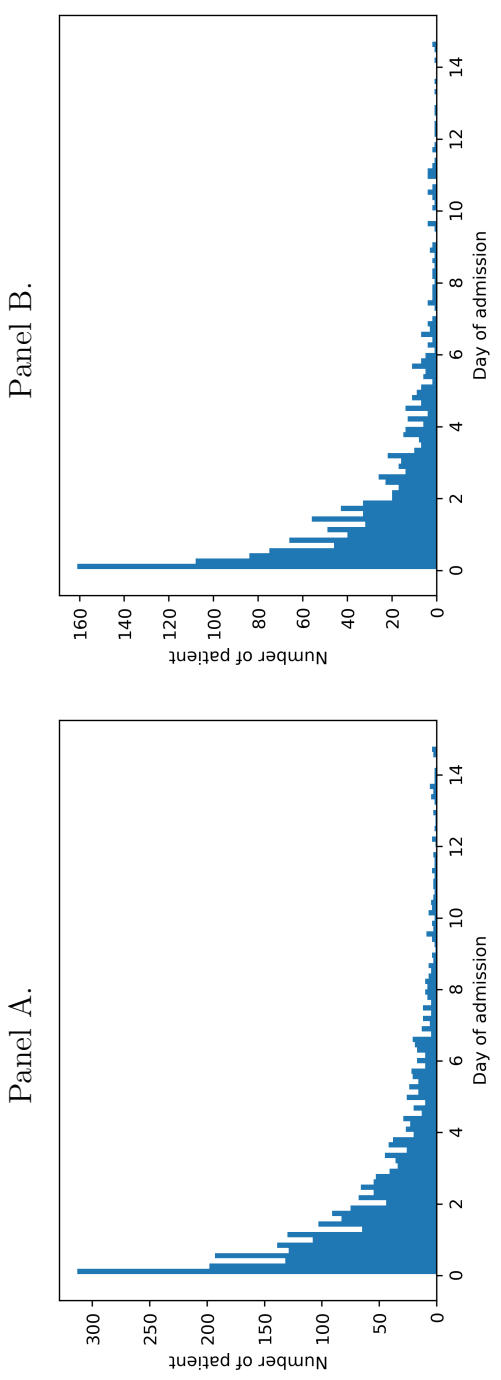


Table 4.12: *Delirium incidence by day. Panel A: eICU-CRD, Panel B: MIMIC-III*

Delirium is precipitated through many factors, some that are unique to the ICU. Our variables were chosen a priori based on literature review. We only included variables that can be easily extracted in real time. Instead of using static values, we employed a sliding window for prediction and incorporated the trajectory of each variable over time. Our results indicate that this strategy predicts delirium more accurately than values captured at a moment in time and eliminates the need for long term prediction.

Since we conducted a retrospective study, causality between the features and delirium cannot be established. Other limitations include selection bias (we excluded observations with missing CAM-ICU values) and interpreter bias (the data recorded in the databases might have been collected after the onset of delirium, given the noncontinuous nature of CAM-ICU measurement). Additionally CAM-ICU was scored by different nurses at different times and in different units, potentially resulting in inter-operator variability.

4.5 Conclusion

We successfully designed a delirium prediction model as a potential screening tool for ABCDEF bundle implementation. Using a few clinically relevant predictor variables we were able to achieve comparable performance to contemporary and well reported models. We were able to tackle the challenge presented by evolving temporal and treatment effects by using methods that captured temporal trends in data rather than static values and sliding observation windows, threshold adjustments to ensure consis-

tently high recall. Additionally we peeked at interpreting the model and shared our code online for reproducibility. We believe our model will help with identifying patients at risk of delirium early and will allow us to target preventive therapies, which is often time consuming and resource-intensive, to the patients who are most likely to benefit.

In the next chapter, an interpretable attention-based deep learning model for time-series EHRs is proposed and implemented to facilitate the interpretability and rank the most influential variables in predicting the delirium outcomes.

Chapter 5

An Interpretable Deep Learning Model for Time-Series Electronic Health Records: Delirium Prediction in ICU

The adoption of deep learning (DL) models has received increasing attention in the clinical domain, particularly in the intensive care unit (ICU). In this context, the interpretability of the outcomes predicted by DL models becomes an essential step towards the applicability of DL models in clinical practice. To address this challenge, we propose two ante-hoc attention-based neural network models to interpret the DL models outcome. The proposed models employ two attention-based mechanisms, including self-attention and effective-attention, to capture influential variables on the predicted outcome. We evaluated our proposed models on two real-world clinical datasets comprising 15726 patients, to predict the incidence of delirium 12 hours or 48 hours before the delirium onset.

Besides, the proposed models are compared to three post-hoc interpretable algorithms considering descriptive performance. In this context, the proposed models cover most of the top-10 variables ranked by the other three post-hoc interpretable algorithms, with the advantage of taking into account the dependencies among variables and additionally the dependencies between varying time steps. The experimental results demonstrate that the proposed models can improve predictive and descriptive performances by employing the BiLSTM model with the effective-attention mechanism.

5.1 Introduction

Deep learning (DL) methods and specifically recurrent neural networks (RNNs) are revolutionizing many fields such as natural language processing [157], machine translation [158], and as well clinical domain [159]. In this regard, the use of DL models has demonstrated an upward trend in the clinical field for the past years [160]. These models can capture the non-linear relationships in clinical data and significantly outperform the conventional machine learning (ML) models. However, DL models show a limited degree of interpretability and are considered black-boxes [161]. Therefore, we need to probe these models better to extract a degree of interpretability from them to make these models more reliable for clinicians.

The conventional machine learning models have been used in the intensive care unit (ICU), which are interpretable [162] but cannot capture the non-linear relationship in data. This is because the data in ICU is

recorded in a time-series manner and conventional ML models do not have the ability to deal with time-series inputs. More advanced models can deal with time-series data and consider the patients' status evolution frequently such as RNNs but are not inherently interpretable.

The interpretability of DL models remains one of the biggest challenges in the ML domain. In this context, interpretability and explainability concepts are often used interchangeably within the general Artificial Intelligence (AI) community [163]. Interpretable models are categorized into post-hoc and ante-hoc models. Post-hoc models incorporate the interpretable module only at inference and as such, they aim to keep a trained model unchanged, while explaining their behavior externally. Examples of post-hoc methods include Shapley Value Sampling (SVS)[164], Integrated Gradients (IG) [41], and Guided Back-propagation (GB) [165].

In contrast, the ante-hoc models incorporate the interpretable module during the training. As a consequence, a single model is employed for both prediction and interpretation. Attention-based models, such as [11] belong to ante-hoc interpretable models.

Self-attention is an attention mechanism that relates different positions of a single sequence in order to compute a representation of the sequence. The self-attention mechanism has been employed successfully in a variety of tasks, including machine translation [29], abstractive summarization [30], and textual entailment [31].

Recently developed attention models offer the promise of providing interpretability while retaining the flexibility and versatility of DL models. The attention-based models were initially employed to predict outpatient

disease progression [11]. The Attend and Diagnose model of Song et al. [166] used a self-attention mechanism to improve an RNN's predictive accuracy for four clinical tasks but did not explore interpretability. While important time points were easily extracted from this model, identifying important variables at a given point in time required additional calculation which is not considered in the proposed method. Choi et al. [34] proposed RETAIN model, which uses two separate RNN layers integrated with an attention layer over both variables and time using embedded variables. In contrast to our model, Choi et al. do not consider the dependencies between time-steps and dependencies between different variables and use two separate RNN networks, which could be computationally expensive; their model is trained and validated on EHR data to predict heart failure. The attention-based model of Kaji et al. [11], which is applied to three clinical tasks, focused on variable-level interpretability. However, it does not consider the time-level importance, and dependencies among variables and times-steps are not considered. Similar to [11], a possible way of interpreting the structured data is to employ an attention-layer straight after the input layer, which computes the coefficient of each variable before being fed into RNN. In the study done by Zhang et al. [167] an LSTM-based model with event embedding and time encoding is leveraged to model clinical time series for early prediction of sepsis in the emergency department. Additionally, an attention mechanism and global max pooling techniques are employed to enable interpretation for the LSTM-based model. Unlike [167] that converted numerical values into categorical values and created an embedding out of them, in our study, we deal with actual numerical val-

ues and for the categorical variables we converted them into embeddings. Additionally, we employed double self-attention architecture to provide a meaningful interpretability.

However, the above-mentioned ante-hoc interpretable models have three limitations as follows:

1. The dependencies among clinical variables and time-steps are not captured.

2. Time-step importance is not considered as the attention is applied on variable-level.

3. The predictive performance is worsened compared to the BiLSTM (Bidirectional long short-term memory) by 3%.

To address the limitations as mentioned earlier, we propose a Double Self-attention Architecture (DSA), which employs a self-attention [32] mechanism at a variable-level and another self-attention mechanism at the time-step level. Additionally, an effective-attention model is used to interpret the model's outcome as it was found to be more effective than self-attention in [168]. Effective-attention is computed from a matrix decomposition of self-attention mechanism which is explained comprehensively in the explanation module section.

For brevity and concerning the use of self-attention in both algorithms, we term both double self-attention and double effective-attention architectures as *DSA* in the rest of this article. In summary, the contributions of this work are as follows:

- DSA simultaneously attends over the variable level and the time-step level.
- DSA takes into account the dependencies between different time-steps and as well considers the correlation among clinical variables while computing the importance of each variable and time-step.
- DSA outperforms the ante-hoc interpretable models while providing a meaningful interpretability.
- Comparison to the other post-hoc interpretable models verifies the soundness of variable ranking provided by DSA.

We developed and validated an interpretable DL model to provide a variable ranking based on prediction of the onset of delirium in critically ill-patients to prioritize the patients at risk. This is a clinically important case study, because delirium occurrence is common in the ICU. At the same time its etiology is not well understood, while the preventive strategies, such as ABCDEF bundle, are very resource intensive [169]. Our model allows for (1) an interpretable DL model, (2) variable ranking by considering varying aspects such as variable inter-dependence and time-step dependencies, and (3) an interpretable screening tool that can prioritize patients at risk, thus reducing the burden on care providers.

We developed and validated an interpretable DL model to provide a variable ranking based on prediction of the onset of delirium in critically ill-patients to prioritize the patients at risk. This is important because delirium occurrence is common in the ICU, with a not fully understood etiology, while the preventive strategies are resource intensive [169]. Such

a model allows for (1) an interpretable DL model, (2) variable ranking by considering varying aspects such as variable inter-dependence and time-step dependencies, and (3) an interpretable screening tool that can prioritize patients at risk, thus reducing the burden on care providers.

5.2 Materials and methods

5.2.1 Data description, cohort selection and outcome definition

The eICU-CRD is a freely available multi-center database comprising 200,859 patient unit encounters for 139,367 unique patients admitted between 2014 and 2015 in over 200 hospitals located throughout the US [5]. The MIMIC-III database is an open-access single-center ICU database including 53,423 distinct hospital admissions for 46,476 unique patients admitted from 2001 to 2012 [4]. Both datasets comprise patient demographics, vitals, clinical flowsheets, laboratory values, medications, interventions, and outcomes. Any patient admitted to the ICU for 24 hours or more and with at least one CAM (Confusion Assessment Method) was included in our study population. In the patient records, in the case of multiple positive CAM-ICU records, the first CAM-ICU was considered as the incidence of delirium. The patients older than 18 and younger than 89 are included in the study, resulting in 22840 patients (16546 patients from eICU-CRD and 6294 patients from MIMIC-III). The patients characteristics for both datasets are demonstrated in the table 5.1

Variables	eICU			MIMIC		
	CAM-ICU +	CAM-ICU -	p value	CAM-ICU +	CAM-ICU -	p value
Number of patients	3153	13393	—	1268	5026	—
Age, mean (SD), years	65.53 (15.14)	62.20 (16.16)	< 0.05	64.81 (15.62)	63.27 (15.82)	< 0.05
Female (%)	1405 (44)	6295 (47)	—	545 (43)	2211 (44)	—
Height, mean (SD), m	168.47 (18.23)	169.25 (15.90)	< 0.05	170.06 (14.22)	168.88 (14.87)	0.054
Weight, mean (SD), kg	83.06 (29.88)	85.00 (25.58)	< 0.05	82.68 (30.25)	81.53 (24.89)	0.15
Heart Rate, mean (SD), bpm	88.22 (18.06)	85.09 (17.73)	< 0.05	88.60 (17.53)	85.12 (17.29)	< 0.05
Oxygen Saturation, mean (SD), %	97.16 (2.72)	96.80 (2.79)	< 0.05	97.17 (2.71)	96.58 (4.50)	< 0.05
Glucose, mean (SD), mg/dL	140.32 (45.97)	146.46 (56.31)	< 0.05	144.51 (58.70)	141.25 (51.43)	< 0.05
Temperature, mean (SD), °C	37.01 (0.69)	36.97 (2.65)	< 0.05	37.06 (0.76)	36.88 (0.76)	< 0.05
Serum Sodium, mean (SD), mEq/L	140.32 (5.80)	138.57 (5.04)	< 0.05	139.39 (5.48)	138.32 (4.89)	< 0.05
BUN, mean (SD), mg/dL	31.93 (22.10)	25.88 (18.64)	< 0.05	33.96 (24.46)	28.10 (20.77)	< 0.05
WBC, mean (SD), per microliter	13.01 (6.47)	11.08 (5.51)	< 0.05	12.13 (7.73)	10.74 (6.29)	< 0.05
Hemoglobin, mean (SD), g/dL	9.73 (1.89)	10.00 (2.08)	< 0.05	9.76 (1.68)	10.27 (1.76)	< 0.05
Platelets, mean (SD), per microliter	201.34 (122.76)	210.23 (108.70)	< 0.05	202.59 (137.23)	199.53 (114.33)	< 0.05
Serum Potassium, mean (SD), mEq/L	3.98 (0.59)	4.00 (0.57)	0.1431	4.03 (0.57)	4.07 (0.56)	< 0.05
Chloride, mean (SD), mEq/L	105.54 (6.86)	103.24 (6.29)	< 0.05	104.57 (6.69)	104.36 (6.37)	< 0.05
Serum Bicarbonate, mean (SD), mEq/L	35.23 (5.02)	25.52 (5.02)	< 0.05	25.16 (5.21)	24.88 (4.95)	< 0.05
Serum creatinine, mean (SD), mg/dL	1.45 (1.16)	1.37 (1.21)	< 0.05	1.63 (1.28)	1.37 (1.05)	< 0.05
Ventilation, mean (SD)	0.87 (0.34)	0.71 (0.45)	< 0.05	0.56 (0.50)	0.33 (0.47)	< 0.05
Total norepinephrine dose (SD), mcg/kg/min	0.02 (0.31)	0.01 (0.28)	< 0.05	0.08 (0.63)	0.06 (0.57)	< 0.05
SOFA, mean (SD)	4.9 (3.3)	3.42 (2.84)	< 0.05	6.46 (3.77)	6.67 (3.34)	< 0.05
SOFA without GCS, mean (SD)	3.27 (2.83)	2.58 (2.33)	< 0.05	5.42 (3.65)	4.99 (3.13)	< 0.05

Table 5.1: *Characteristics of the included patients divided by the CAM-ICU status*

5.2.2 Variable selection

We compiled 21 clinical variables identified by critical care clinicians as relevant to delirium prediction, commonly used in the literature, and available in both data-sets, including demographics, vital signs, laboratory measurements, and medication data. A more detailed list of the included clinical variables in this study is depicted in Table 5.2.

Variable group	**Variable name**
Demographic data	age, gender, height, weight
Vita signs	oxygen saturation (SpO_2), heart rate (HR), temperature
Other Measurements	sofa, sofa without GCS, Ventilation
Laboratory Measurements	white blood cell count (WBC), sodium (Na), blood urea nitrogen (BUN), glucose, hemoglobin, platelets, potassium, chloride, bicarbonate, creatinine
Medications as continuous drips	Dopamine, epinephrine, norepinephrine, phenylephrine (all calculated as norepinephrine equivalent)

Table 5.2: *Variables included in the prediction models*

5.2.3 Outcome assessment

In this study, we evaluated the ability of the proposed model to provide a meaningful variable ranking in the case of delirium prediction in different settings, such as varying observation window (12h and 24h) and different prediction window (12h and 48h) illustrated in Figure 5.1. In this work, the derivation windows of 12h or 24h are chosen based on the Intensive Care Delirium Screening Checklist (ICDSC) [170] which is an 8-24 hours window [171] to predict the incidence of delirium in the following 12h or 48h. Prediction of delirium incidence is based on the ability of our model using a multi-variable sequence of clinical variables in the observation win-

dow to infer the presence or absence of delirium in the next 12 hours or 48 hours based on the defined settings.

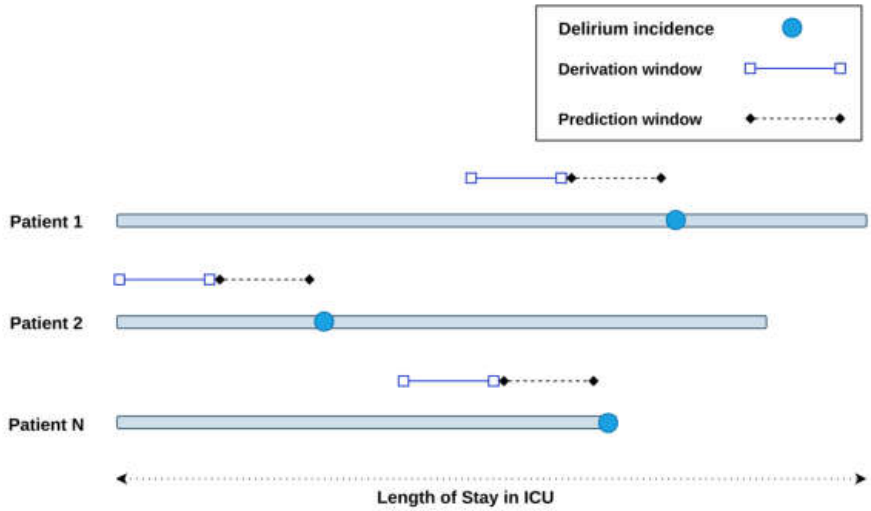


Figure 5.1: *Delirium prediction schema; derivation window represents the collected data for each study (12h, 24h), and the prediction window represents time to predict delirium prior to its incidence (12h, 48h)*

5.2.4 Model development

Each patient can be viewed as a sequence of medical records (vital signs, laboratory measurements, and demographics) ordered by time, and each record contains a set of clinical variables. A three dimensional data with patient ICU stays (n = 15,726), time steps (n = 24 or 12), and variables (n = 21) serves as input to the model. As is demonstrated in Figure 5.2, the proposed model is divided into three modules, namely input preparation, explainable module, and prediction module.

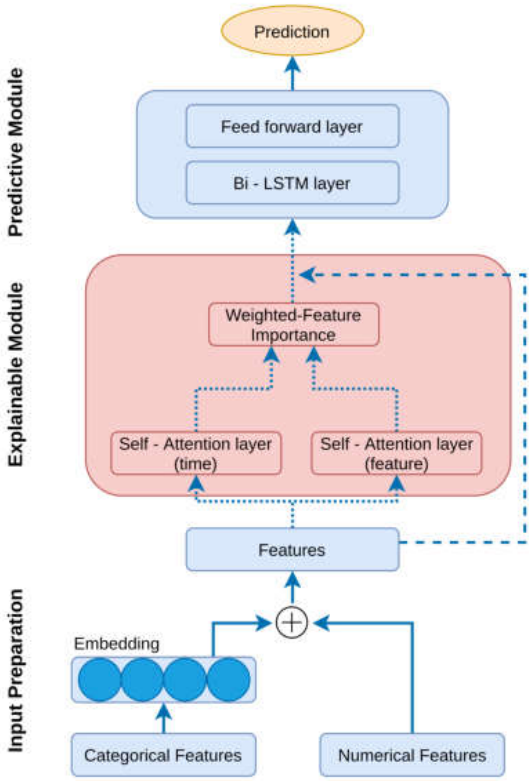


Figure 5.2: *Proposed architecture*

In the input preparation phase, similar to [172] the embedding layer converts categorical variables into vectors. The model in the training phase learns the vectors related to each categorical variable. The vectors of the categorical variable are concatenated with the numerical variables to be fed into the model. The input representation at time t is as follows:

$$x_t = Concat[(Numerical_t, U(Categorical_t)] \tag{5.1}$$

$Numerical_t$ stands for the numerical variable, $Categorical_t$ stands for the categorical variable at time t, and U is the embedding matrix.

In the explanation module, the input x_t is fed into two different self-attention layers, as shown in Figure 5.2. The self-attention mechanism on

the right side is applied on the variables to compute the variable importance, namely α_v. The self-attention mechanism on the left side is applied on the time-steps to compute the time-step importance named as α_t. The coefficient of the contribution is computed via both α_v and α_t using a dot-product applied as follows:

$$c(x_t) = \underbrace{\alpha_v}_{\text{self-attention on variable}} \odot \underbrace{\alpha_t}_{\text{self-attention on time}} \tag{5.2}$$

The input data is weighted with the computed attention using a residual connection as shown in the equation 5.3

$$wi_t = x_t \odot \underbrace{c(x_t)}_{\text{contribution coefficient}} \tag{5.3}$$

where wi_t is the weighted input at time t, x_t is the input which was computed in equation 5.1, and c is the computed coefficient of the contribution in equation 5.2.

In the prediction module, similar to [11] the weighted input is fed to a masking layer to filter the time steps where patients have less than 12 or 24 hours of data available and are fed into a BiLSTM layer to get the data representation for each patient. Formally:

$$h_t = f(wi_t \quad + \quad Wh_{t-1}) \tag{5.4}$$

where wi_t is the weighted input at time t, W is the parameter of RNN learned during training, and f is $tanh$ non-linear operation.

The BiLSTM layer with 128 units is connected to a hyperbolic tangent activation function. The output layer of our network consists of one dense

neuron with a softmax activation to output the probability of a given event over ICU stays.

5.2.5 Model training and evaluation

We evaluated both the descriptive and the predictive performance of our model. The predictive performance of our model Was compared to the model proposed by Kaji et al. [11] and to BiLSTM. While all the prediction scenarios employed the same architecture, hyper-parameter optimization was manually done based on a better convergence of the model and higher predictive performance. BiLSTMs were trained with Adam optimizer with a learning rate of $7.5 * 10^{-4}$, and decay of $1 * 10^{-6}$ was used in all models. Batch sizes of 128 were used for all models. The models, as mentioned earlier, were trained for 50 epochs, and cross-entropy defined the loss function. We evaluated the results based on 5-fold stratified cross-validation. Typically, metrics computed based on the k-fold stratified cross-validation can assess overfitting and have lower variance [173]. We report the predictive performance using the Area Under Receiver Operating Characteristic (AUROC), Area Under Precision-Recall Curve (AUPRC), Precision and, Recall with the Confidence Interval (CI) of 95%.

5.2.6 Explanation module

Understanding how the model predicts a patient's delirium is an essential step in validating its use. DL techniques are typically considered black boxes where it is difficult to determine how a predictive model generates a prediction. A model should provide meaningful explanations related to the

clinical variables and which can be utilized by clinicians during a day-to-day routine. Recent advances in ML techniques like attention mechanisms have produced a better way to probe interpretability. Attention-based [11] models give importance to the classification associated with each input variable given to the model, and this allows us to identify the most predictive variables that contribute to the severity of the diagnosis. In this section, we employ two attention-based models to understand what has been learned by our models. In detail, we can observe which time-steps and variables the model relies on assigning a degree of significance to the time-steps and variables.

As it is shown in Figure 5.3.a, the data of each patient is fed as input into attention-based (self-attention or effective-attention) layers. We employ two 3-head attention mechanisms in order to compute time-step importance and variable importance. As it is shown in Figure 5.3.b, to compute the importance of a single time-step, we need to score each time-step of the input sequence against this single time-step. The score is computed as it is shown in equation 5.6 and determines focus that needed to be placed on other time-steps as we encode a time-step at a specific position. In this way, we can capture the dependencies between time steps while computing the importance of each time step. As is depicted in Figure 5.3.c, to compute the importance of a single variable, we need to score each variable of the input sequence against this single variable, and the score determines focus that needed to be placed on other variables as we encode a variable. While computing the importance of each variable, the self-attention enables capturing dependencies among variables, including

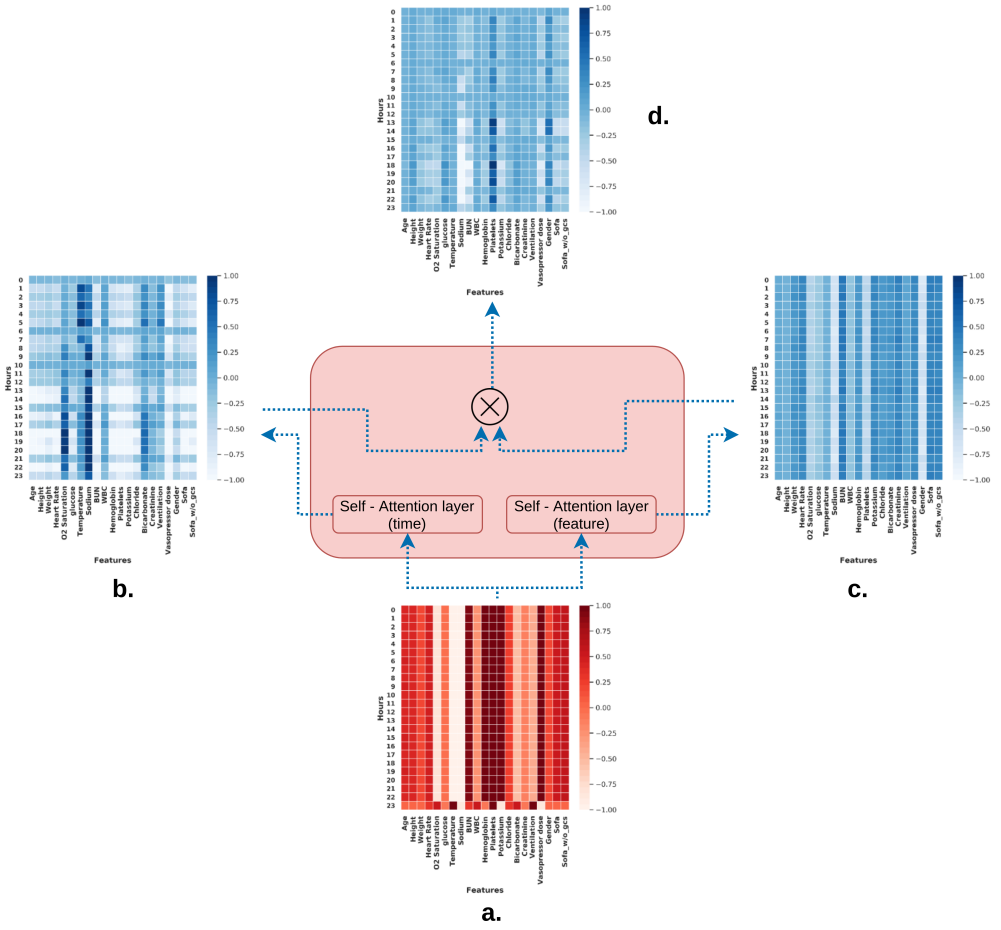


Figure 5.3: *Explanation module: a.Input data; b.Time importance; c.Variable importance; d.Variable importance by considering time importance*

the temporal dimension.

In the following sections, we provide a detailed description of the varying version of attention mechanisms which provide interpretable outputs, namely self-attention and effective-attention.

Self-attention

The implications of time steps and clinical variables vary depending on the context. To capture this contextual information, we applied two self-attention layers to which one self-attention layer attends over time-steps and the other self-attention layer attends over variables.

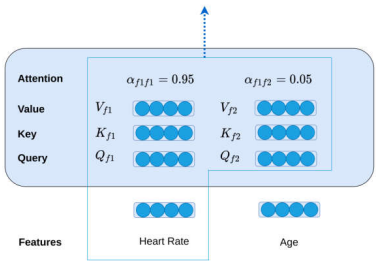


Figure 5.4: *The architecture of self-attention*

As it is demonstrated in Figure 5.4, Self-attention is an attention mechanism that relates different positions of a single sequence in order to compute a representation of the sequence. The self-attention mechanism has been employed successfully in a variety of tasks, including machine translation [29], abstractive summarization [30], and textual entailment [31]. Formally,

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{5.5}$$

Where Q, K, V are computed by multiplying input with the learned matrices W_Q, W_k, W_V during training.

DSA employs multi-head self-attention, which projects queries, keys, and values h times with different, learned linear projections. The scores are computed in parallel and are concatenated to get one matrix score, Formally:

$$Multihead(Q, K, V) = Concat(head_1, head_2)W^O \tag{5.6}$$

where $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$

Where parameters matrices such as W_i^Q, W_i^K, W_i^V, and W^O are the projections [32].

Effective-attention

As it is demonstrated in [174], self-attention can be decomposed into two matrices: i) the component in the left nullspace of V which is indicated with ($A^{\|}$) and ii) the component orthogonal to the nullspace ($A^{\perp}$). The matrix $A^{\|}$ is irrelevant for the model output because its product with the value matrix is equal to zero. The matrix $A^{\perp}$ contributes to the model output, which is so-called *effective-attention*.

Additionally, as Sun et al.[168] noted effective-attention is associated less with the variables related to the language modeling pretraining such as separator [SEP], and it has the potential to illustrate linguistic variables much better than self-attention. Equivalent to our study, we believe that effective-attention is less associated with with less important clinical variables and it can capture better than self-attention the most relevant clinical variables to predict the outcome, namely delirum onset.

$$AV = (A^{\|} + A^{\perp})V = \vec{0} + A^{\perp}V = A^{\perp}V \tag{5.7}$$

As it is illustrated in equation 5.7, $A^{\|}V$ is equal to zero, therefore the effective-attention matrix is equal to $A^{\perp}$.

The effective-attention matrix $A^{\perp}$, is computed as the following[174]:

- We first compute the singular value decomposition (SVD) of the value

matrix V which is $V = U \sum WT$

- The rows of U that correspond to singular values equal to zero span LN(V):
  $LN(V) = span\{u_1, ..., u_k\}$,
  Where k is the number of singular values that equal zero.

- We project each row a_i of the attention matrix A to LN(V) to construct a projection of the matrix A to LN(V):
  $P_{LN(V)}(a_i) = \sum_{j=1}^{k} \langle a_i, u_j \rangle \; u_j, \; \forall i \in \{1, ..., d_s\}$,
  $P_{LN(V)}(A) = [P_{LN(V)}(a_1), ..., P_{LN(V)}(a_{ds})]$
  where $\langle ., . \rangle$ denotes the dot product.

- effective-attention is equal to
  $A^{\perp} := A - P_{LN(V)}(A)$

It is worth mentioning that similar to [168], for analyzing effective-attention, we replace self-attention with effective-attention at the model test phase.

5.2.7 Data and code availability

We made use of several open-source libraries based on Python to conduct our experiments; machine learning framework Scikit-learn [109] and DL framework Pytorch [110]. We have made the experiments and implementation details publicly available at this GitHub repository such that anyone can implement the proposed method in this study.

5.3 Results

In this section we report both, the descriptive evaluation of the model as well as predictive performance. The descriptive performance is evaluated against the well known algorithms, including Shapley Value Sampling, Integrated Gradients and Guided Back-propagation, while the predictive performance is based on evaluation metrics such as AUROC, AUPRC, precision and recall with 95% CI.

5.3.1 Descriptive performance

As mentioned earlier the importance of interpretable deep learning models in the clinical domain, in this section we explore further interpretability by providing the most important clinical variables for delirium-onset prediction task. In this regard, we compute variable importance by considering the importance of one variable over other variables across the while cohort as shown in the Equation 5.2. Although there are many definitions of interpretability, we focused on how the model ranks each input variable with respect to outcome prediction. Given that interpretability of neural networks is still an open research question, especially for temporal neural networks [175], we also provide results from three other post-hoc models to compare with our proposed model. In this context, we employed as the benchmark the Shapley Value Sampling (SVS) [164], Integrated Gradient (IG) [41], and Guided Backpropagation (GB) [165], to ensure that the variable importance results computed by DSA are consistent across the three benchmark models. The top-10 influential variables ranked for MIMIC-III

and eICU-CRD are reported in Table 5.3 and Table 5.4 respectively. The variable ranking is reported using three different post-hoc interpretable algorithms, namely IG, SVS, and GB, compared to two proposed ante-hoc attention-based interpretable models, namely self-attention and effective-attention.

The most influential variables that have contributed to delirium prediction according to their relative importance in the eICU-CRD dataset as reported in Table 5.4 are heart rate, Ventilation, age, white blood cell count, and vasopressor dose according to the five algorithms. Most of these variables are also ranked in the top-10 in the MIMIC-III dataset as depicted in Table 5.3. Both proposed attention-based interpretable models (DSA) captured most of the important variables ranked by IG, SVS, and GB and in both datasets, validating the soundness of the proposed model, with the additional advantage of providing also inter-variable dependencies as well as temporal importance.

It is interesting to note that, effective-attention which was previously used in [168] shows a slightly higher number of variables in common with the other three post-hoc algorithms that is an extra point for effective-attention to be studied further in the case of clinical time-series data.

5.3.2 Predictive performance

We evaluated 12014 (24h derivation – 12h prediction) and 9481 (12h derivation – 48h prediction) from eICU-CRD and 3712 (24h derivation – 12h prediction) and 2128 patients (12h derivation – 48h prediction) from MIMIC-III databases. Considering AUPRC, Precision, and Recall, DSA outper-

	Observation window 12h – Prediction window 48h				
Algorithm / **Variable ranking**	**IG**	**SVS**	**GB**	**DSA (self-attention)**	**DSA (effective-attention)**
1	ventilation	ventilation	ventilation	weight	*heart rate*
2	WBC	gender	WBC	*hemoglobin*	*platelets*
3	creatinine	WBC	creatinine	*heart rate*	*BUN*
4	vasopressor dose	vasopressor dose	vasopressor dose	*sodium*	*hemoglobin*
5	sodium	sofa	sodium	*BUN*	SpO_2
6	BUN	sodium	BUN	*WBC*	potassium
7	glucose	age	glucose	potassium	*WBC*
8	platelets	sofa w/o GCS	platelets	*vasopressor dose*	height
9	hemoglobin	creatinine	hemoglobin	*glucose*	*sofa w/o GCS*
10	heart rate	BUN	heart rate	bicarbonate	*creatinine*
	Observation window 24h – Prediction window 12h				
1	ventilation	ventilation	ventilation	temperature	*WBC*
2	gender	gender	heart rate	*ventilation*	weight
3	sodium	heart rate	sodium	*sodium*	*age*
4	heart rate	sodium	creatinine	SpO_2	*ventilation*
5	sofa	sofa	bicarbonate	*platelets*	*bicarbonate*
6	age	age	age	*BUN*	SpO_2
7	bicarbonate	sofa w/o GCS	BUN	*age*	height
8	vasopressor dose	vasopressor dose	vasopressor dose	chloride	*gender*
9	sofa w/o GCS	bicarbonate	platelets	weight	potassium
10	creatinine	creatinine	WBC	*vasopressor dose*	*vasopressor dose*

Table 5.3: Variable ranking presented by different algorithms versus DSA (top-10 variables) on MIMIC-III dataset.

.

forms the proposed model by Kaji et al. in two different scenarios as shown in Table 5.5 for MIMIC-III and for eICU-CRD datasets as depicted in Table 5.6 . Another point to mention, although the predictive performance of the DSA is better than Kaji's model, its predictive performance is slightly worse than BiLSTM. This is due to the nature of ante-hoc interpretable models in which there is a trade-off between predictive performance power and descriptive performance [11].

	Observation window 12h – Prediction window 48h				
Variable ranking \ **Algorithm**	**IG**	**SVS**	**GB**	**DSA (self-attention)**	**DSA (effective-attention)**
1	ventilation	ventilation	ventilation	*heart rate*	sodium
2	heart rate	heart rate	heart rate	sodium	*heart rate*
3	age	age	WBC	platelets	hemoglobin
4	WBC	WBC	age	height	*WBC*
5	sofa	sofa	vasopressor dose	*age*	sofa w/o GCS
6	vasopressor dose	vasopressor dose	bicarbonate	*chloride*	*ventilation*
7	bicarbonate	bicarbonate	chloride	*weight*	*glucose*
8	BUN	BUN	BUN	*WBC*	*age*
9	chloride	chloride	glucose	sofa w/o GCS	*weight*
10	weight	weight	weight	SpO_2	height
	Observation window 24h – Prediction window 12h				
1	ventilation	ventilation	vasopressor dose	*potassium*	*age*
2	vasopressor dose	vasopressor dose	ventilation	temperature	*WBC*
3	age	age	WBC	creatinine	*vasopressor dose*
4	heart rate	heart rate	heart rate	*vasopressor dose*	*potassium*
5	WBC	WBC	age	SpO_2	*weight*
6	potassium	potassium	potassium	*weight*	sodium
7	sofa	sofa	platelets	*heart rate*	hemoglobin
8	bicarbonate	bicarbonate	bicarbonate	*age*	creatinine
9	gender	weight	weight	*platelets*	*gender*
10	weight	platelets	BUN	*BUN*	*ventilation*

Table 5.4: Variable ranking presented by varying algorithms vs. proposed model (top-10 variables) on eICU-CRD dataset.

5.4 Discussion

Our study shows that the proposed models outperform the state-of-the-art interpretable model proposed by [11], while being interpretable and comparable to a handful number of post-hoc interpretable algorithms such as IG, SVS, and GB. This demonstrates the strength of DSA in both the predictive performance and the associated interpretable matrix.

Improving the prediction of delirium is a critical step towards improving ICU outcomes, and costs [176]. This study found that incorporating two self-attention layers with a BiLSTM layer can achieve informative AUROCs in modeling delirium-related tasks. The AUROC, AUPRC, Precision, and

	Observation window 12h – Prediction window 48h			
Model	AUROC% (95% CI)	AUPRC% (95% CI)	Precision% (95% CI)	Recall% (95% CI)
BiLSTM	71.37 (67.99 - 74.72)	29.81 (27.33 - 31.88)	28.45 (25.74 - 31.17)	65.98 (59.82 - 72.13)
DSA	68.66 (64.99 - 72.33)	28.58 (23.64 - 33.20)	26.70 (22.56 - 30.85)	59.85 (51.93 - 67.77)
Kaji model	67.56 (64.91 - 70.22)	27.90 (25.68 - 30.46)	24.31 (22.01 - 26.60)	58.07 (52.48 - 63.66)
	Observation window 24h – Prediction window 12h			
Model	AUROC% (95% CI)	AUPRC% (95% CI)	Precision% (95% CI)	Recall% (95% CI)
BiLSTM	81.24 (76.44 - 86.11)	44.45 (40.35 - 48.81)	35.03 (30.73 - 39.33)	71.16 (64.54 - 77.77)
DSA	80.50 (77.23 - 83.85)	44.90 (41.28 - 49.40)	35.58 (33.71 - 37.44)	68.98 (62.69 - 75.27)
Kaji model	78.33 (76.11 - 80.62)	41.69 (38.55 - 45.15)	32.39 (28.52 - 36.25)	65.63 (57.78 - 73.48)

Table 5.5: Predictive performance on MIMIC-III dataset

	Observation window 12h – Prediction window 48h			
Model	AUROC% (95% CI)	AUPRC% (95% CI)	Precision% (95% CI)	Recall% (95% CI)
BiLSTM	84.20 (82.52 - 85.86)	33.24 (29.41 - 36.54)	28.37 (27.15 - 29.58)	74.44 (70.91 - 77.98)
DSA	82.51 (80.33 - 84.67)	31.21 (28.01 - 33.70)	24.92 (24.22 - 25.61)	75.99 (72.20 - 79.78)
Kaji model	81.64 (80.05 - 83.27)	30.19 (27.41 - 32.27)	24.60 (23.23 - 25.97)	75.00 (70.56 - 79.44)
	Observation window 24h – Prediction window 12h			
Model	AUROC% (95% CI)	AUPRC% (95% CI)	Precision% (95% CI)	Recall% (95% CI)
BiLSTM	88.02 (86.31 - 89.75)	42.69 (38.71 - 46.1)	38.19 (36.78 - 39.6)	80.39 (76.62 - 84.16)
DSA	87.10 (85.15 - 89.03)	42.20 (38.78 - 45.69)	34.17 (32.88 - 35.46)	82.89 (78.54 - 87.25)
Kaji model	85.85 (84.16 - 87.57)	38.03 (34.64 - 41.02)	35.82 (34.48 - 37.15)	76.63 (72.84 - 80.42)

Table 5.6: Predictive performance on eICU-CRD dataset

Recall for the delirium prediction suggest that employing self-attention performs comparably to the BiLSTM while providing a meaningful variable ranking.

In this study, we demonstrated how to get a level of interpretability for a DL model for clinical events in ICU by incorporating self-attention mechanisms. As [177] noted, the interpretability of DL can facilitate the process of understanding the inferential process of a neural network and improve the model in terms of descriptive and predictive performance.

While many BiLSTM based models to predict clinical outcomes have incorporated attention, we are aware of a handful of them that used attention to identify the variables driving the prediction [34, 11, 166, 167, 178, 179, 180]. Several of the studies as mentioned earlier employed attention; however, none of them compared the proposed model with other interpretable models.

We demonstrated how self-attention could be applied to the input variables to provide a degree of interpretability by capturing variable dependencies and time-step dependencies. Our study has several limitations. Many variables that clinicians wanted to incorporate into this study were not available in eICU-CRD and MIMIC-III datasets or had a very high rate of missing data. We note that the proposed self-attention model can underline the importance of each variable but cannot identify whether a variable increase or decreases the probability of an event without additional analysis.

In conclusion, we believe that self-attention mechanisms could create interpretable decision support systems for intensivists. This study demonstrated that such an approach could learn informative models for predicting delirium and has shown how the individual variables underlying these predictions can be explored using self-attention mechanisms. We demonstrated that the explainability module proposed in this paper could be used effectively to visualize the variable importance to help understand the input variables.

We believe that the proposed interpretable model can aid in clinical decision making by helping clinicians focus on particular variables that

the model has deemed important at time points of interest of the disease trajectory.

In the next chapter, we conclude the thesis and a discussion regarding the future studies are provided.

Chapter 6

Conclusion and Future Directions

6.1 Conclusions

In this thesis, we presented our approaches to some limitations in applying machine learning in ICU. We focused on several clinical problems, namely mortality prediction, length of stay in ICU, phenotyping, decompensation, and delirium prediction in critically ill patients. Moreover, the interpretability of deep learning models applied to time series data is investigated. Our solutions presented in Chapter 3, 4, and 5 are built on top of RNN BiLSTM and leverage entity embedding and interpretability techniques.

The lack of publicly available benchmarks is an obstacle to accelerate the progress of machine learning in ICU. In this context, in Chapter 3 we proposed a benchmark on the application of machine learning models in multi-center ICU, which allows researchers to build on top of our work by improving the models or addressing other existing challenges in the application of machine learning in ICU. Our findings confirmed the advantages

of using entity embedding over one-hot encoding, RNN BiLSTM to capture temporal dependencies than conventional machine learning better. In some tasks, the ability to use fewer clinical variables while achieving the same predictive performance.

Delirium adversely affects both short-term and long-term patient outcomes. In Chapter 4, we proposed a delirium prediction model as a potential screening tool for ABCDEF bundle implementation to provide a higher quality of care for the patients and optimize the resources in ICU by using a few clinical variables. The proposed method, which employed RNN BiLSTM with entity embedding, tackled the challenge presented by evolving temporal and treatment effects in a sliding window manner. In addition to that, we attempted to make transparent the opacity of the proposed method (black-box) by applying three different interpretable methods. In this context, we provided variable ranking according to the variable influence in the prediction model's outcome. We believe our model will help with identifying patients at risk of delirium early and will allow us to target preventive therapies, which are often time-consuming and resource-intensive, to the patients who are most likely to benefit.

Building on these results, in Chapter 5, we proposed a novel ante-hoc interpretable deep learning model for time series EHRs. The proposed method utilizes self-attention or effective-attention mechanisms. The approach consists of two 3-head self/effective-attention mechanisms, one mechanism for capturing the variable dependencies and the other mechanism to capture the time dependencies to capture the most influential variables. Experimental results show that our solution can outperform

other ante-hoc interpretable methods considering predictive performance and performs comparably with the other three post-hoc interpretable methods. Our findings show that self-attention and effective-attention can provide some degree of interpretability while providing an acceptable predictive performance compare to the baselines. The soundness of the proposed method is verified by intensivists and additionally by comparison to the other three post-hoc interpretable methods.

6.2 Future Directions

To further advance the predictive and descriptive performances of machine learning models in ICU, we outline the following aspects as future work:

Healthcare data are usually difficult to access and share across research institutions and hospitals because of their sensitive nature, and this makes a big obstacle for developing generalizable and effective analytical approaches. Federated learning could be applied in this scenario to address the data accessibility issue. In this regard, federated learning provides the excellent potential to make a generalized and effective model while preserving data privacy [181].

The availability of the pre-trained models for some of the common clinical tasks could significantly improve the model performance and its generalizability as it has been done in varying fields such as image analysis [182, 183], natural language processing [184], and machine translation [185]. In this context, the transfer-learning approach could be applied to learn a better representation of the data and associated parameters of

pre-trained models for a new dataset or a new task [186].

Towards interpretability of the deep learning prediction outcome, interpretable deep learning models are gradually dominating the black-box deep learning models. However, the possible loss of the predictive performance for gaining descriptive performance is often unavoidable. This loss puts researchers in the difficulty of choosing between high predictive performance (black-box) and high descriptive performance (interpretable) models. In this regard, a hybrid interpretable model is desired. A hybrid interpretable model consists of varying modules for interpretability and prediction. This model can address the challenge of choosing between high predictive performance and high descriptive performance by providing high predictive and descriptive performances [187]

To conclude, recent studies show that the number of publicly available datasets in ICU is increasing gradually [188, 6]. Hence, even with many variants of deep learning modeling and a few publicly available datasets, the dependency on the generalizability and applicability of a model in ICU remains a primary challenge for machine learning research in ICU. With the increasing digitization of activities in the healthcare domain, working machine learning models for the majority of the clinical tasks creates a higher quality of healthcare services in the day-to-day lives of billions of people. In conclusion, we would like to highlight the importance of improving the predictive performance, generalizability, and interpretability of machine learning in the different clinical tasks, improving patients' quality of care, and optimizing the resources in the clinical domain to prioritize the patients at higher risk.

Bibliography

[1] Zeina Rayan, Marco Alfonse, and Abdel-Badeeh M Salem. Intensive care unit (icu) data analytics using machine learning techniques.

[2] Alistair EW Johnson, Mohammad M Ghassemi, Shamim Nemati, Katherine E Niehaus, David A Clifton, and Gari D Clifford. Machine learning and decision support in critical care. *Proceedings of the IEEE*, 104(2):444–466, 2016.

[3] L Nelson Sanchez-Pinto, Yuan Luo, and Matthew M Churpek. Big data and data science in critical care. *Chest*, 154(5):1239–1248, 2018.

[4] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[5] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5:180178, 2018.

[6] Patrick J Thoral, Jan M Peppink, Ronald H Driessen, Eric JG

Sijbrands, Erwin JO Kompanje, Lewis Kaplan, Heatherlee Bailey, Jozef Kesecioglu, Maurizio Cecconi, Matthew Churpek, et al. Sharing icu patient data responsibly under the society of critical care medicine/european society of intensive care medicine joint data science collaboration: The amsterdam university medical centers database (amsterdamumcdb) example. *Critical care medicine*, 49(6):e563, 2021.

[7] Yajuan Wang, Kenney Ng, Roy J Byrd, Jianying Hu, Shahram Ebadollahi, Zahra Daar, Christopher deFilippi, Steven R Steinhubl, and Walter F Stewart. Early detection of heart failure with varying prediction windows by structured and unstructured data in electronic health records. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2530–2533. IEEE, 2015.

[8] Jionglin Wu, Jason Roy, and Walter F Stewart. Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, pages S106–S113, 2010.

[9] Benjamin Goehry. Random forests for time-dependent processes. *ESAIM: Probability and Statistics*, 24:801–826, 2020.

[10] Benjamin Goehry, Hui Yan, Yannig Goude, Pascal Massart, and Jean-Michel Poggi. Random forests for time series: Accepted - november 2021. *REVSTAT-Statistical Journal*, Nov. 2021.

[11] Deepak A Kaji, John R Zech, Jun S Kim, Samuel K Cho, Neha S

Dangayach, Anthony B Costa, and Eric K Oermann. An attention based deep learning model of clinical events in the intensive care unit. *PloS one*, 14(2):e0211057, 2019.

[12] Neil A Halpern and Stephen M Pastores. Critical care medicine in the united states 2000–2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. *Critical care medicine*, 38(1):65–71, 2010.

[13] **Seyedmostafa Sheikhalishahi**, Vevake Balaraman, and Venet Osmani. Benchmarking machine learning models on multi-centre eicu critical care dataset. *PloS one*, 15(7):e0235424, 2020.

[14] Anirban Bhattacharyya, **Seyedmostafa Sheikhalishahi**, Siddharth Dugar, Sudhir Krishnan, Abhijit Duggal, and Venet Osmani. 400: Predicting delirium risk for the following 24 hours in critically ill patients using deep learning. *Critical Care Medicine*, 48(1):182, 2020.

[15] Anirban Bhattacharyya, **Seyedmostafa Sheikhalishahi**
, Heather Torbic, Wesley Yeung, Tiffany Wang, Jennifer Birst, Abhijit Duggal, Leo Anthony Celi, and Venet Osmani. Delirium prediction in the icu - designing a screening tool for preventive interventions (under review). *Journal of the American Medical Informatics Association Open.*

[16] **Seyedmostafa Sheikhalishahi**, Anirban
Bhattacharyya, Leo Anthony Celi, and Venet Osmani. An interpretable deep learning model for time-series electronic health records:

Case study of delirium prediction in critical care (under review). *Artificial Intelligence In Medicine*.

[17] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[18] David Freedman. *Statistical Models : Theory and Practice*. Cambridge University Press, August 2005.

[19] Shikha Chourasia. Survey paper on improved methods of id3 decision tree classification. *International Journal of Scientific and Research Publications*, 3(12):1–2, 2013.

[20] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[21] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[22] Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

[23] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[24] Yoshua Bengio, Patrice Simard, Paolo Frasconi, et al. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[25] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318, 2013.

[26] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[27] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. 1999.

[28] Jean-Benoit Delbrouck. Transformer for emotion recognition. *arXiv preprint arXiv:1805.02489*, 2018.

[29] Kaitao Song, Xu Tan, Furong Peng, and Jianfeng Lu. Hybrid self-attention network for machine translation. *arXiv preprint arXiv:1811.00253*, 2018.

[30] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization, 2017.

[31] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding, 2017.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010, 2017.

[33] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*, 2019.

[34] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pages 3504–3512, 2016.

[35] Andreas Lehrmann and Leonid Sigal. Non-parametric structured output networks. In *Advances in neural information processing systems*, pages 4214–4224, 2017.

[36] Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther. Auxiliary deep generative models. *arXiv preprint arXiv:1602.05473*, 2016.

[37] Takashi Matsubara. Bayesian deep learning: A model-based interpretable approach. *Nonlinear Theory and Its Applications, IEICE*, 11(1):16–35, 2020.

[38] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[39] Scott M Lundberg and Su-In Lee. A unified approach to interpret-

ing model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[40] Erik Strumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11:1–18, 2010.

[41] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017.

[42] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[43] Julian D Olden, Michael K Joy, and Russell G Death. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological modelling*, 178(3-4):389–397, 2004.

[44] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):307, 2008.

[45] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[46] Sumanta Basu, Karl Kumbier, James B Brown, and Bin Yu. Iterative random forests to discover predictive and stable high-order interac-

tions. *Proceedings of the National Academy of Sciences*, 115(8):1943–1948, 2018.

[47] Karl Kumbier, Sumanta Basu, James B Brown, Susan Celniker, and Bin Yu. Refining interaction search through signed iterative random forests. *arXiv preprint arXiv:1810.07287*, 2018.

[48] Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. *arXiv preprint arXiv:1705.04977*, 2017.

[49] Reza Abbasi-Asl and Bin Yu. Structural compression of convolutional neural networks based on greedy filter pruning. *arXiv preprint arXiv:1705.07356*, 21, 2017.

[50] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[51] Rui P Moreno, Philipp GH Metnitz, Eduardo Almeida, Barbara Jordan, Peter Bauer, Ricardo Abizanda Campos, Gaetano Iapichino, David Edbrooke, Maurizia Capuzzo, Jean-Roger Le Gall, et al. Saps 3—from evaluation of the patient to evaluation of the intensive care unit. part 2: Development of a prognostic model for hospital mortality at icu admission. *Intensive care medicine*, 31(10):1345–1355, 2005.

[52] Jack E Zimmerman, Andrew A Kramer, Douglas S McNair, and Fern M Malila. Acute physiology and chronic health evaluation (apache) iv: hospital mortality assessment for today's critically ill patients. *Critical care medicine*, 34(5):1297–1310, 2006.

[53] William A Knaus, Elizabeth A Draper, Douglas P Wagner, and Jack E Zimmerman. Apache ii: a severity of disease classification system. *Critical care medicine*, 13(10):818–829, 1985.

[54] 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*, 190(6):649–655, 2014.

[55] 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. *Critical care medicine*, 44(2):368, 2016.

[56] Jean-Roger Le Gall, Philippe Loirat, Annick Alperovitch, Paul Glaser, Claude Granthil, Daniel Mathieu, Philippe Mercier, Remi Thomas, and Daniel Villers. A simplified acute physiology score for icu patients. *Critical care medicine*, 12(11):975–977, 1984.

[57] Hannah Lee, Susie Yoon, Seung-Young Oh, Jungho Shin, Jeongsoo Kim, Chul-Woo Jung, and Ho Geol Ryu. Comparison of apache iv

with apache ii, saps 3, meld, meld-na, and ctp scores in predicting mortality after liver transplantation. *Scientific reports*, 7(1):1–10, 2017.

[58] Sookyung Hyun, Pacharmon Kaewprag, Cheryl Cooper, Brenda Hixon, and Susan Moffatt-Bruce. Exploration of critical care data by using unsupervised machine learning. *Computer Methods and Programs in Biomedicine*, page 105507, 2020.

[59] Carolyn S Calfee, Kevin Delucchi, Polly E Parsons, B Taylor Thompson, Lorraine B Ware, Michael A Matthay, Nhlbi Ards Network, et al. Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. *The Lancet Respiratory Medicine*, 2(8):611–620, 2014.

[60] Yuan Luo, Yu Xin, Rohit Joshi, Leo A Celi, and Peter Szolovits. Predicting icu mortality risk by grouping temporal trends from a multivariate panel of physiologic measurements. In *AAAI*, pages 42–50, 2016.

[61] Kelly C Vranas, Jeffrey K Jopling, Timothy E Sweeney, Meghan C Ramsey, Arnold S Milstein, Christopher G Slatore, Gabriel J Escobar, and Vincent X Liu. Identifying distinct subgroups of intensive care unit patients: a machine learning approach. *Critical care medicine*, 45(10):1607, 2017.

[62] Alistair EW Johnson, Tom J Pollard, and Roger G Mark. Repro-

ducibility in critical care: a mortality prediction case study. In *Machine Learning for Healthcare Conference*, pages 361–376, 2017.

[63] Ziad Obermeyer and Thomas H Lee. Lost in thought: the limits of the human mind and the future of medicine. *The New England journal of medicine*, 377(13):1209, 2017.

[64] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[65] Antoine Neuraz, Claude Guérin, Cécile Payet, Stéphanie Polazzi, Frédéric Aubrun, Frédéric Dailler, Jean-Jacques Lehot, Vincent Piriou, Jean Neidecker, Thomas Rimmelé, et al. Patient mortality is associated with staff resources and workload in the icu: a multicenter observational study. *Critical care medicine*, 43(8):1587–1594, 2015.

[66] Nicole G Weiskopf, George Hripcsak, Sushmita Swaminathan, and Chunhua Weng. Defining and measuring completeness of electronic health records for secondary use. *Journal of biomedical informatics*, 46(5):830–836, 2013.

[67] Brian J Wells, Kevin M Chagin, Amy S Nowacki, and Michael W Kattan. Strategies for handling missing data in electronic health record derived data. *Egems*, 1(3), 2013.

[68] Roderick J Little, Ralph D'Agostino, Michael L Cohen, Kay Dickersin, Scott S Emerson, John T Farrar, Constantine Frangakis, Joseph W Hogan, Geert Molenberghs, Susan A Murphy, et al. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367(14):1355–1360, 2012.

[69] Jonathan AC Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338, 2009.

[70] Anis Sharafoddini, Joel A Dubin, and Joon Lee. Patient similarity in prediction models based on health data: a scoping review. *JMIR medical informatics*, 5(1):e7, 2017.

[71] Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj*, 361, 2018.

[72] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 2011.

[73] Zoie Shui Yee Wong. Statistical classification of drug incidents due to look-alike sound-alike mix-ups. *Health informatics journal*, 22(2):276–292, 2016.

[74] Mohammed Khalilia, Sounak Chakraborty, and Mihail Popescu. Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, 11(1):51, 2011.

[75] Robert C Holte, Liane Acker, Bruce W Porter, et al. Concept learning

and the problem of small disjuncts. In *IJCAI*, volume 89, pages 813–818. Citeseer, 1989.

[76] Rahul C Deo. Machine learning in medicine. *Circulation*, 132(20):1920–1930, 2015.

[77] Abraham Verghese, Nigam H Shah, and Robert A Harrington. What this computer needs is a physician: humanism and artificial intelligence. *Jama*, 319(1):19–20, 2018.

[78] Rajul Parikh, Annie Mathai, Shefali Parikh, G Chandra Sekhar, and Ravi Thomas. Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, 56(1):45, 2008.

[79] Yutaka Sasaki. The truth of the f-measure. *Teach Tutor Mater*, 01 2007.

[80] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[81] Amber Stubbs, Michele Filannino, Ergin Soysal, Samuel Henry, and Özlem Uzuner. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *Journal of the American Medical Informatics Association*, 26(11):1163–1171, 2019.

[82] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg

Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

[83] Rinaldo Bellomo, Stephen J Warrillow, and Michael C Reade. Why we should be wary of single-center trials. *Critical care medicine*, 37(12):3114–3119, 2009.

[84] N Youssef, K Reinhart, and Y Sakr. The pros and cons of multicentre studies. *Neth J Crit Care*, 12(3), 2008.

[85] William A Knaus, Douglas P Wagner, Elizabeth A Draper, Jack E Zimmerman, Marilyn Bergner, Paulo G Bastos, Carl A Sirio, Donald J Murphy, Ted Lotring, Anne Damiano, et al. The apache iii prognostic system: risk prediction of hospital mortality for critically iii hospitalized adults. *Chest*, 100(6):1619–1636, 1991.

[86] Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24):2957–2963, 1993.

[87] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of Biomedical Informatics*, 83(April):112–134, 2018.

[88] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

[89] Hamid R. Darabi, Daniel Tsinis, Kevin Zecchini, Winthrop F. Whitcomb, and Alexander Liss. Forecasting mortality risk for patients admitted to intensive care units using machine learning. *Procedia Computer Science*, 140:306 – 313, 2018. Cyber Physical Systems and Deep Learning Chicago, Illinois November 5-7, 2018.

[90] Leo Anthony Celi, Sean Galvin, Guido Davidzon, Joon Lee, Daniel Scott, and Roger Mark. A database-driven decision support system: customized mortality prediction. *Journal of personalized medicine*, 2(4):138–148, 2012.

[91] Mani Sotoodeh and Joyce C Ho. Improving length of stay prediction using a hidden markov model. *AMIA Summits on Translational Science Proceedings*, 2019:425, 2019.

[92] Xin Ma, Yabin Si, Zifan Wang, and Youqing Wang. Length of stay prediction for icu patients using individualized single classification algorithm. *Computer methods and programs in biomedicine*, 186:105224, 2020.

[93] Jack V. Tu and Michael R.J. Guerriere. Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. *Computers and Biomedical Research*, 26(3):220 – 229, 1993.

[94] Joyce C Ho, Joydeep Ghosh, and Jimeng Sun. Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the 20th ACM*

SIGKDD international conference on Knowledge discovery and data mining, pages 115–124, 2014.

[95] Jiayu Zhou, Fei Wang, Jianying Hu, and Jieping Ye. From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 135–144, 2014.

[96] Tala B Shahin, Baran Balkan, Jarrod Mosier, and Vignesh Subbian. The connected intensive care unit patient: Exploratory analyses and cohort discovery from a critical care telemedicine database. *JMIR medical informatics*, 7(1):e13006, 2019.

[97] Jarrod Mosier and Vignesh Subbian. Rule-based cohort definitions for acute respiratory failure: Electronic phenotyping algorithm. *JMIR Medical Informatics*, 8(4):e18402, 2020.

[98] Yejin Kim, Robert El-Kareh, Jimeng Sun, Hwanjo Yu, and Xiaoqian Jiang. Discriminative and distinct phenotyping by constrained tensor factorization. *Scientific reports*, 7(1):1–12, 2017.

[99] Narges Razavian, Jake Marcus, and David Sontag. Multi-task prediction of disease onsets from longitudinal laboratory tests. In *Machine Learning for Healthcare Conference*, pages 73–100, 2016.

[100] Oliver Ren, Alistair EW Johnson, Eric P Lehman, Matthieu Komorowski, Jerome Aboab, Fengyi Tang, Zach Shahn, Daby Sow,

Roger Mark, and Li-wei Lehman. Predicting and understanding unexpected respiratory decompensation in critical care using sparse and heterogeneous clinical data. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 144–151. IEEE, 2018.

[101] Yanbo Xu, Siddharth Biswal, Shriprasad R Deshpande, Kevin O Maher, and Jimeng Sun. Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2565–2573. ACM, 2018.

[102] Mehmet Kılıç, Nureddin Yüzkat, Celaleddin Soyalp, and Nurçin Gülhaş. Cost analysis on intensive care unit costs based on the length of stay. *Turkish journal of anaesthesiology and reanimation*, 47(2):142, 2019.

[103] Spiros Denaxas, Arturo Gonzalez-Izquierdo, Natalie Fitzpatrick, Kenan Direk, and Harry Hemingway. Phenotyping uk electronic health records from 15 million individuals for precision medicine: The caliber resource. *Studies in health technology and informatics*, 262:220–223, 2019.

[104] Spiros Denaxas, Helen Parkinson, Natalie Fitzpatrick, Cathie Sudlow, and Harry Hemingway. Analyzing the heterogeneity of rule-based ehr phenotyping algorithms in caliber and the uk biobank. *BioRxiv*, page 685156, 2019.

[105] Ann McGinley and Rupert M Pearse. A national early warning score for acutely ill patients, 2012.

[106] Cheng Guo and Felix Berkhahn. Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*, 2016.

[107] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.

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

[109] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[110] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc,

E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[111] Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, 7:e623, 2021.

[112] Zhongheng Zhang, Marcus W Beck, David A Winkler, Bin Huang, Wilbert Sibanda, Hemant Goyal, et al. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Annals of translational medicine*, 6(11), 2018.

[113] Christoph Molnar. *Interpretable Machine Learning*. 2019. `https://christophm.github.io/interpretable-ml-book/`.

[114] Karla D Krewulak, Henry T Stelfox, Jeanna Parsons Leigh, E Wesley Ely, and Kirsten M Fiest. Incidence and prevalence of delirium subtypes in an adult icu: a systematic review and meta-analysis. *Critical care medicine*, 46(12):2029–2035, 2018.

[115] Monidipa Dasgupta and Chris Brymer. Poor functional recovery after delirium is associated with other geriatric syndromes and additional illnesses. *International psychogeriatrics*, 27(5):793, 2015.

[116] Biren B Kamdar, Michael P Combs, Elizabeth Colantuoni, Lauren M King, Timothy Niessen, Karin J Neufeld, Nancy A Collop, and Dale M Needham. The association of sleep quality, delirium,

and sedation status with daily participation in physical therapy in the icu. *Critical Care*, 20(1):1–9, 2016.

[117] Amra Sakusic, John C O'Horo, Mikhail Dziadzko, Dziadzko Volha, Rashid Ali, Tarun D Singh, Rahul Kashyap, Ann M Farrell, John D Fryer, Ronald Petersen, et al. Potentially modifiable risk factors for long-term cognitive impairment after critical illness: a systematic review. In *Mayo Clinic Proceedings*, volume 93, pages 68–82. Elsevier, 2018.

[118] Maria Schubert, Roger Schürch, Soenke Boettger, David Garcia Nuñez, Urs Schwarz, Dominique Bettex, Josef Jenewein, Jasmina Bogdanovic, Marina Lynne Staehli, Rebecca Spirig, et al. A hospital-wide evaluation of delirium prevalence and outcomes in acute care patients-a cohort study. *BMC health services research*, 18(1):1–12, 2018.

[119] E Wesley Ely, Ayumi Shintani, Brenda Truman, Theodore Speroff, Sharon M Gordon, Frank E Harrell Jr, Sharon K Inouye, Gordon R Bernard, and Robert S Dittus. Delirium as a predictor of mortality in mechanically ventilated patients in the intensive care unit. *Jama*, 291(14):1753–1762, 2004.

[120] Pratik Pandharipande, Bryan A Cotton, Ayumi Shintani, Jennifer Thompson, Brenda Truman Pun, John A Morris Jr, Robert Dittus, and E Wesley Ely. Prevalence and risk factors for development of delirium in surgical and trauma icu patients. *The Journal of trauma*, 65(1):34, 2008.

[121] John A McPherson, Chad E Wagner, Leanne M Boehm, J David Hall, Daniel C Johnson, Leanna R Miller, Kathleen M Burns, Jennifer L Thompson, Ayumi K Shintani, E Wesley Ely, et al. Delirium in the cardiovascular intensive care unit: exploring modifiable risk factors. *Critical care medicine*, 41(2):405, 2013.

[122] Pratik P Pandharipande, Timothy D Girard, James C Jackson, Alessandro Morandi, Jennifer L Thompson, Brenda T Pun, Nathan E Brummel, Christopher G Hughes, Eduard E Vasilevskis, Ayumi K Shintani, et al. Long-term cognitive impairment after critical illness. *New England Journal of Medicine*, 369(14):1306–1316, 2013.

[123] Tamara G Fong, Samir R Tulebaev, and Sharon K Inouye. Delirium in elderly adults: diagnosis, prevention and treatment. *Nature Reviews Neurology*, 5(4):210, 2009.

[124] Annachiara Marra, E Wesley Ely, Pratik P Pandharipande, and Mayur B Patel. The abcdef bundle in critical care. *Critical care clinics*, 33(2):225–243, 2017.

[125] John W Devlin, Yoanna Skrobik, Céline Gélinas, Dale M Needham, Arjen JC Slooter, Pratik P Pandharipande, Paula L Watson, Gerald L Weinhouse, Mark E Nunnally, Bram Rochwerg, et al. Clinical practice guidelines for the prevention and management of pain, agitation/sedation, delirium, immobility, and sleep disruption in adult patients in the icu. *Critical care medicine*, 46(9):e825–e873, 2018.

[126] S Jean Hsieh, Olufisayo Otusanya, Hayley B Gershengorn, Aluko A

Hope, Christopher Dayton, Daniela Levi, Melba Garcia, David Prince, Michele Mills, Dan Fein, et al. Staged implementation of awakening and breathing, coordination, delirium monitoring and management, and early mobilization bundle improves patient outcomes and reduces hospital costs. *Read Online: Critical Care Medicine— Society of Critical Care Medicine*, 47(7):885–893, 2019.

[127] E Wesley Ely, Richard Margolin, Joseph Francis, Lisa May, Brenda Truman, Robert Dittus, Theodore Speroff, Shiva Gautam, Gordon R Bernard, and Sharon K Inouye. Evaluation of delirium in critically ill patients: validation of the confusion assessment method for the intensive care unit (cam-icu). *Critical care medicine*, 29(7):1370–1379, 2001.

[128] Timothy D Girard, Pratik P Pandharipande, and E Wesley Ely. Delirium in the intensive care unit. *Critical care*, 12(3):1–9, 2008.

[129] Nathan E Brummel, Eduard E Vasilevskis, Jin Ho Han, Leanne Boehm, Brenda T Pun, and E Wesley Ely. Implementing delirium screening in the intensive care unit: secrets to success. *Critical care medicine*, 41(9):2196, 2013.

[130] Babar A Khan, Anthony J Perkins, Nagendra K Prasad, Anantha Shekhar, Noll L Campbell, Sujuan Gao, Sophia Wang, Sikandar H Khan, Edward R Marcantonio, Homer L Twigg, et al. Biomarkers of delirium duration and delirium severity in the icu. *Critical care medicine*, 48(3):353–361, 2020.

[131] Matthew M Ruppert, Jessica Lipori, Sandip Patel, Elizabeth Ingersent, Julie Cupka, Tezcan Ozrazgat-Baslanti, Tyler Loftus, Parisa Rashidi, and Azra Bihorac. Icu delirium-prediction models: A systematic review. *Critical care explorations*, 2(12), 2020.

[132] Heidi Lindroth, Lisa Bratzke, Suzanne Purvis, Roger Brown, Mark Coburn, Marko Mrkobrada, Matthew TV Chan, Daniel HJ Davis, Pratik Pandharipande, Cynthia M Carlsson, et al. Systematic review of prediction models for delirium in the older adult inpatient. *BMJ open*, 8(4):e019223, 2018.

[133] Andrew Wong, Albert T Young, April S Liang, Ralph Gonzales, Vanja C Douglas, and Dexter Hadley. Development and validation of an electronic health record–based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. *JAMA network open*, 1(4):e181018–e181018, 2018.

[134] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

[135] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[136] Sharon K Inouye, Catherine M Viscoli, Ralph I Horwitz, Leslie D Hurst, and Mary E Tinetti. A predictive model for delirium in hospitalized elderly medical patients based on admission characteristics. *Annals of internal medicine*, 119(6):474–481, 1993.

[137] Peter Pompei, Marquis Foreman, Mark A Rudberg, Sharon K Inouye, Victoria Braund, and Christine K Cassel. Delirium in hospitalized older persons: outcomes and predictors. *Journal of the American Geriatrics Society*, 42(8):809–815, 1994.

[138] Min Young Kim, Ui Jun Park, Hyoung Tae Kim, and Won Hyun Cho. Delirium prediction based on hospital information (delphi) in general surgery patients. *Medicine*, 95(12), 2016.

[139] Sarah T Pendlebury, Nicola G Lovett, Sarah C Smith, Rose Wharton, and Peter M Rothwell. Delirium risk stratification in consecutive unselected admissions to acute medicine: validation of a susceptibility score based on factors identified externally in pooled data for use at entry to the acute care pathway. *Age and ageing*, 46(2):226–231, 2017.

[140] James L Rudolph, Richard N Jones, Sue E Levkoff, Christopher Rockett, Sharon K Inouye, Frank W Sellke, Shukri F Khuri, Lewis A Lipsitz, Basel Ramlawi, Sidney Levitsky, et al. Derivation and validation of a preoperative prediction rule for delirium after cardiac surgery. *Circulation*, 119(2):229, 2009.

[141] Marcela P Carrasco, Luis Villarroel, Maricarmen Andrade, Jorge

Calderón, and Matías González. Development and validation of a delirium predictive score in older people. *Age and ageing*, 43(3):346–351, 2014.

[142] Jacqueline M Leung, Laura P Sands, Eunjung Lim, Tiffany L Tsai, and Sakura Kinjo. Does preoperative risk for delirium moderate the effects of postoperative pain and opiate use on postoperative delirium? *The American Journal of Geriatric Psychiatry*, 21(10):946–956, 2013.

[143] ST O'keeffe and JN Lavan. Predicting delirium in elderly patients: development and validation of a risk-stratification model. *Age and Ageing*, 25(4):317–321, 1996.

[144] Kees J Kalisvaart, Ralph Vreeswijk, Jos FM De Jonghe, Tjeerd Van Der Ploeg, Willem A Van Gool, and Piet Eikelenboom. Risk factors and prediction of postoperative delirium in elderly hip-surgery patients: Implementation and validation of a medical risk factor model. *Journal of the American Geriatrics Society*, 54(5):817–822, 2006.

[145] Soyeon Jang, Kwang-Ik Jung, Woo-Kyoung Yoo, Myung Hun Jung, and Suk Hoon Ohn. Risk factors for delirium during acute and subacute stages of various disorders in patients admitted to rehabilitation units. *Annals of rehabilitation medicine*, 40(6):1082, 2016.

[146] James L Rudolph, Kelly Doherty, Brittany Kelly, Jane A Driver, and Elizabeth Archambault. Validation of a delirium risk assessment

using electronic medical record information. *Journal of the American Medical Directors Association*, 17(3):244–248, 2016.

[147] James L Rudolph, Mary Beth Harrington, Michelle A Lucatorto, Jennifer G Chester, Joseph Francis, Kenneth J Shay, Veterans Affairs, and Delirium Working Group. Validation of a medical record-based delirium risk assessment. *Journal of the American Geriatrics Society*, 59:S289–S294, 2011.

[148] Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

[149] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[150] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[151] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[152] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

[153] Aya Abdelsalam Ismail, Mohamed Gunady, Héctor Corrada Bravo,

and Soheil Feizi. Benchmarking deep learning interpretability in time series predictions. *arXiv preprint arXiv:2010.13924*, 2020.

[154] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[155] M Van den Boogaard, P Pickkers, AJC Slooter, MA Kuiper, PE Spronk, PHJ Van der Voort, JG Van Der Hoeven, R Donders, Theo van Achterberg, and L Schoonhoven. Development and validation of pre-deliric (prediction of delirium in icu patients) delirium prediction model for intensive care patients: observational multicentre study. *Bmj*, 344, 2012.

[156] A Wassenaar, MHWA van den Boogaard, Theo van Achterberg, AJC Slooter, MA Kuiper, ME Hoogendoorn, KS Simons, E Maseda, N Pinto, C Jones, et al. Multinational development and validation of an early prediction model for delirium in icu patients. *Intensive care medicine*, 41(6):1048–1056, 2015.

[157] Abhyuday N Jagannatha and Hong Yu. Structured prediction models for rnn based sequence labeling in clinical text. In *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing*, volume 2016, page 856. NIH Public Access, 2016.

[158] S. P. Singh, A. Kumar, H. Darbari, L. Singh, A. Rastogi, and S. Jain. Machine translation using deep learning: An overview. In *2017 Inter-*

national Conference on Computer, Communications and Electronics (Comptelix), pages 162–167, 2017.

[159] 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*, 22(5):1589–1604, 2017.

[160] Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T Dudley, Alberto Lavelli, Fabio Rinaldi, and Venet Osmani. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR medical informatics*, 7(2):e12239, 2019.

[161] Arun Rai. Explainable ai: from black box to glass box. *Journal of the Academy of Marketing Science*, 48(1):137–141, 2020.

[162] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 559–560, 2018.

[163] Wilson Silva, Kelwin Fernandes, Maria J Cardoso, and Jaime S Cardoso. Towards complementary explanations using deep neural networks. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 133–140. Springer, 2018.

[164] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

[165] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[166] Huan Song, Deepta Rajan, Jayaraman J Thiagarajan, and Andreas Spanias. Attend and diagnose: Clinical time series analysis using attention models. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[167] Dongdong Zhang, Changchang Yin, Katherine M Hunold, Xiaoqian Jiang, Jeffrey M Caterino, and Ping Zhang. An interpretable deep-learning model for early prediction of sepsis in the emergency department. *Patterns*, 2(2):100196, 2021.

[168] Kaiser Sun and Ana Marasović. Effective attention sheds light on interpretability. *arXiv preprint arXiv:2105.08855*, 2021.

[169] Felipe Martinez, Catalina Tobar, and Nathan Hill. Preventing delirium: should non-pharmacological, multicomponent interventions be used? a systematic review and meta-analysis of the literature. *Age and ageing*, 44(2):196–204, 2015.

[170] Nicolas Bergeron, M-J Dubois, M Dumont, S Dial, and Y Skrobik. Intensive care delirium screening checklist: evaluation of a new screening tool. *Intensive care medicine*, 27(5):859–864, 2001.

[171] Nathan E Brummel, Eduard E Vasilevskis, Jin Ho Han, Leanne Boehm, Brenda T Pun, and E Wesley Ely. Implementing delirium

screening in the intensive care unit: secrets to success. *Critical care medicine*, 41(9):2196, 2013.

[172] Seyedmostafa Sheikhalishahi, Vevake Balaraman, and Venet Osmani. Benchmarking machine learning models on multi-centre eicu critical care dataset, 2019.

[173] Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

[174] Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. *arXiv preprint arXiv:1908.04211*, 2019.

[175] Aya Abdelsalam Ismail, Mohamed Gunady, Héctor Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time series predictions. *arXiv preprint arXiv:2010.13924*, 2020.

[176] Maria Schubert, Roger Schürch, Soenke Boettger, David Garcia Nuñez, Urs Schwarz, Dominique Bettex, Josef Jenewein, Jasmina Bogdanovic, Marina Lynne Staehli, Rebecca Spirig, et al. A hospital-wide evaluation of delirium prevalence and outcomes in acute care patients-a cohort study. *BMC health services research*, 18(1):1–12, 2018.

[177] Giulia Vilone and Luca Longo. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*, 2020.

[178] Kourosh T Baghaei and Shahram Rahimi. Sepsis prediction: an attention-based interpretable approach. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6. IEEE, 2019.

[179] Peipei Chen, Wei Dong, Jinliang Wang, Xudong Lu, Uzay Kaymak, and Zhengxing Huang. Interpretable clinical prediction via attention-based neural network. *BMC Medical Informatics and Decision Making*, 20(3):1–9, 2020.

[180] Yanni Kang, Xiaoyu Jia, Kaifei Wang, Yiying Hu, Jianying Guo, Lin Cong, Xiang Li, and Guotong Xie. A clinically practical and interpretable deep model for icu mortality prediction with external validation. In *AMIA Annual Symposium Proceedings*, volume 2020, page 629. American Medical Informatics Association, 2020.

[181] Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5(1):1–19, 2021.

[182] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021.

[183] Stefan Hinterstoisser, Vincent Lepetit, Paul Wohlhart, and Kurt Konolige. On pre-trained image features and synthetic images for

deep learning. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[184] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26, 2020.

[185] Rongxiang Weng, Heng Yu, Shujian Huang, Shanbo Cheng, and Weihua Luo. Acquiring knowledge from pre-trained model to neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9266–9273, 2020.

[186] Benjamin Shickel, Anis Davoudi, Tezcan Ozrazgat-Baslanti, Matthew Ruppert, Azra Bihorac, and Parisa Rashidi. Deep multimodal transfer learning for augmented patient acuity assessment in the intelligent icu. *Frontiers in Digital Health*, 3:11, 2021.

[187] Tong Wang and Qihang Lin. Hybrid predictive model: When an interpretable model collaborates with a black-box model. *arXiv preprint arXiv:1905.04241*, 2019.

[188] Stephanie L Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, 26(3):364–373, 2020.