EARLY PREDICTION OF CRITICAL EVENTS IN INFANTS WITH SINGLE VENTRICLE PHYSIOLOGY IN CRITICAL CARE USING ROUTINELY COLLECTED DATA

by

Victor Manuel Ruiz Herrera

Bachelor of Science, Pontificia Universidad Javeriana, 2011

Master of Science, University of Pittsburgh, 2014

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Victor Manuel Ruiz Herrera

It was defended on

January 11, 2019

and approved by

Dr. Rich Tsui, Associate Professor, Biomedical Informatics

Dr. Shyam Visweswaran, Associate Professor, Biomedical Informatics

Dr. Douglas Landsittel, Assistant Professor, Biomedical Informatics

Dr. Alejandro Lopez-Magallon, Professor, Critical Care Medicine

EARLY PREDICTION OF CRITICAL EVENTS IN INFANTS WITH SINGLE VENTRICLE PHYSIOLOGY IN CRITICAL CARE USING ROUTINELY COLLECTED DATA

Victor Manuel Ruiz Herrera, PhD

University of Pittsburgh, 2019

Intensive care units (ICUs) provide care for critically-ill patients who require constant monitoring and the availability of specialized equipment and personnel. In this environment, a high volume of information and a high degree of uncertainty present a burden to clinicians. In specialized cohorts, such as pediatric patients with congenital heart defects (CHDs), this burden is exacerbated by increased complexity, the inadequacy of existing decision support aids, and the limited and decreasing availability of highly-specialized clinicians.

Among CHD patients, infants with single ventricle (SV) physiology are one of the most complex and severely-ill sub-populations. While SV mortality rates have dropped, patient deterioration may happen unexpectedly in the period before patients undergo stage-2 palliative surgery. Even in expert hands, critical and potentially catastrophic events (CEs), such as cardiopulmonary resuscitation (CPR), emergent endotracheal intubation (EEI), or extracorporeal membrane oxygenation (ECMO) are common in SV patients, and may negatively impact morbidity, mortality, and hospital length of stay.

There is a clinical need of predictive tools that help intensivists assess and forecast the advent of CEs in SV infants. Although ubiquitous, widely adopted ICU severity-of-illness scores or early warning systems (EWS), e.g., PRISM and PIM, have not met this need. They are often

developed for general ICU use and do not generalize well to specialized populations. Furthermore, most EWS are developed for prediction of patient mortality. Among SV patients, however, death is semi-elective. On the other hand, prediction of CEs may help clinicians improve patient care by anticipating the advent of patient deterioration.

In this dissertation, we aimed to develop and validate predictive models that achieve early and accurate prediction of CEs in infants with SV physiology. Such models may provide early and actionable information to clinicians and may be used to perform clinical interventions aimed at preventing CEs, and to reducing morbidity, mortality, and healthcare costs. We assert that our work is significant in that it addresses an unmet clinical need by achieving state-of-the-art, early prediction of patient deterioration in a challenging and vulnerable population.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I am immensely grateful to everyone who made this journey possible.

Thanks to my advisor and committee members: To Dr. Rich Tsui for his scientific mentorship and financial support, and to Drs. Alejandro López, Shyam Visweswaran and Douglas Landsittel for their domain knowledge, feedback, and research advice.

Thanks to Howard Su, Dr. Lucas Saenz, Dr. Henry Ogoe, Gabriella Butler, Celia Pulver, Ashlee Shields, Helen Shi, and Dr. Joe Wu who facilitated data collection and mapping.

Thanks to Dr. Juan Carlos Puyana for his financial support and for making it possible for me to begin graduate school. I hope to one day follow in his footsteps in fostering new scientists and advancing health care and education in my home country, Colombia.

Thanks to all the faculty and staff at the University of Pittsburgh's Department of Biomedical Informatics. Special thanks to Toni Porterfield. Toni, thanks for looking out for me and for pushing me to get things done!

Thanks to my friends who became my surrogate family in the US and who helped me go through hard times. Special thanks to Arturo, Sergio, Brian, Soma, Becca, Mike, Steven, Andy, Travis, and José.

Finally, and most importantly, thanks to my family. This achievement would not have been possible without the love and encouragement of my siblings Sergio and Patricia, and of my wonderful parents Víctor and Azucena. Mom and dad, thank you for giving me life, a beautiful

home, and for being the best role models anyone could wish for. You made innumerable sacrifices so I could come this far and I will never be able to thank you enough. Dad, I hope you were still with us to celebrate this occasion; you are ever present in my heart.

Abbreviations:

APACHE: Acute Physiology and Chronic Health Evaluation

AUC: Area under the receiver operating characteristics curve

BPTT: Back-propagation through time

C-WIN: Cardiac intensive care unit Warning Index

CART: Classification and regression trees

CE: Critical Event

CHD: Congenital Heart Disease

CHOP: Children's Hospital of Philadelphia

CHP: UPMC Children's Hospital of Pittsburgh

CPR: Cardiopulmonary resuscitation

CPT: Conditional Probability Table

DT: Decision Tree

ECG: Electrocardiogram

ECMO: Extra-Corporeal Membrane Oxygenation

EEI: Emergent Endotracheal Intubation

EWS: Early Warning System

FTP: Frequent Temporal Pattern

ICU: Intensive Care Unit

LSTM: Long Short-Term Memory

MSS: Multivariate State Sequence

NB: Naïve Bayes

PRISM: Pediatric Risk or Mortality

PIM: Pediatric Index of Mortality

RNN: Recurrent Neural Network

ROC: Receiver Operating Characteristics curve

SV: Single Ventricle

TA: Temporal Abstraction

TP: Temporal Pattern

1.0 INTRODUCTION

Children and adults with congenital heart disease (CHD) are diverse and complex populations whose management presents a challenge for clinicians and a heavy burden for the healthcare system. In the U.S., 3.7% of all pediatric hospitalizations and 15% of the total cost of pediatric hospitalizations are related to CHD [1].

Infants with single-ventricle (SV) physiology are among the most complex CHD populations, and have high mortality and morbidity risk prior to stage-2 surgical repair [1]. While in-hospital mortality rates have decreased, patient deterioration may happen unexpectedly during the course of critical-care. Even under the care of experienced critical-care teams, potentially-catastrophic critical events (CEs), such as cardiopulmonary resuscitation (CPR), emergent endotracheal intubation (EEI), or extracorporeal membrane oxygenation (ECMO) are common in SV patients. Such events may negatively impact morbidity, mortality, and hospital length of stay [2,3].

To facilitate the detection of unexpected patient deterioration, early warning systems (EWS) assess patients' risk of CEs in real time and provide alerts to clinicians. Their success depends on their ability to (1) predict CEs accurately, (2) alert clinicians with enough time to respond, (3) use objective and readily-available data, and (4) function without increasing the workload of clinicians.

EWS have been used to predict mortality risk in ICUs for decades [4–7], yet there is a scarcity of EWS that predict CEs in pediatric ICUs for specialized populations. Several pediatric EWS predict combined outcomes of mortality, cardiac arrest, or transfer to an ICU in general pediatric populations[8–10]; however, only few pediatric EWS predict the risk of CEs in ICU settings[11–15].

There is very limited research focused on the prediction of CEs in SV infants. To the best of our knowledge, only three studies have attempted to predict CEs in this population. Gupta et al. proposed a model that predicts poor outcomes before or right after stage-1 surgery (Norwood procedure)[16]. This model, however, generates predictions from demographic data, baseline characteristics, and factors related to the Norwood operation. This is a significant limitation because (1) risk of CEs cannot be assessed in real-time, and (2) surgical practices change over time, reducing the life-span of models based on those data. Vu et al. assessed the differences in electrocardiogram (ECG)-lead signals before and after cardiac arrest events and found a statistically-significant difference in ST-vector magnitude and instability between pre and post arrest periods. However, this study did not evaluate this finding in the context of real-time prediction[17]. Finally, Rusin et al. developed a model that predicted CEs with high accuracy in the hour preceding CEs[18]. To the best of our knowledge, this is the only model suitable for real-time prediction of CEs in SV infants currently available. Nonetheless, this state-of-the-art model has two limitations: (1) its accuracy drops rapidly when CEs are predicted more than two hours in advance, and (2) it requires real-time analysis of high-frequency electrocardiogram (ECG) and vital-signs data, which may present technological and financial challenges for many institutions.

To address the need of a real-time EWS for the SV population, we aim at developing and evaluating the *C*ardiac-*I*CU *W*arning *IN*dex (C-WIN) system. This will consist of state-of-the-art predictive models that leverage expert knowledge as well as objective, routinely-collected data for

the early prediction of CEs (CPR, EEI and ECMO) in infants with SV physiology in pediatric ICUs (PICU, NICU, CICU).

1.1 RESEARCH QUESTIONS

1. Is it feasible to build a predictive model from expert clinical knowledge that predicts the onset of catastrophic events (CEs) in the ICU for infants with SV physiology?
2. Can we extract temporal-abstraction features from a longitudinal dataset of objective, routinely-collected clinical data and use said features to build classifiers to predict CEs in real time?
3. How accurately and how early can models built (1) using expert clinical knowledge, (2) using temporal-abstraction features extracted from EHR data, or (3) using raw time-series values, predict CEs?

1.2 SPECIFIC AIMS

We answered the research questions in this dissertation through the following specific aims.

- Aim 1: To build an expert model based on expert clinical knowledge
    - Elicit knowledge from cardiac intensive-care expert clinicians in the form of (1) a list relevant of variables that may predict the onset CEs, (2) SV-specific discretization ranges for numeric variables, and (3) a quantification of the interaction between variables and the risk of CEs.

    - Build a Naïve Bayes model from the variables, discretization ranges, and risk estimations elicited from experts.
- Aim 2: To train state-of-the art models for the early prediction of CEs for SV patients.
    - Retrieve a dataset consisting of variables identified by experts for a cohort of SV ICU admissions.
    - Re-parametrize the expert Bayesian model built in Aim 1 based on the retrieved dataset.
    - Identify trend-summary and temporal-abstraction features and rank them in the order of predictive ability from the retrieved dataset.
    - Build classic machine-learning classifier from static, trend-summary, and temporal-abstraction features, and a dynamic classifier from raw time-series data values.
- Aim 3: To evaluate and compare the performance of models developed in Aims 1 and 2.
    - Evaluate performance of expert, static, and dynamic models on the dataset retrieved from CHP.
    - Evaluate the best-performing classifier on an external dataset, retrieved from the Children's Hospital of Philadelphia (CHOP).

1.3 HYPOTHESES

We tested three hypotheses. First, that models that encode SV-domain-specific knowledge from cardiac intensivists will perform better in predicting CEs than currently-available models. Second, that using clinical data to extract temporal features and train static classifiers will result in

significantly higher performance than that of expert models. Third, that dynamic models that leverage temporal patterns in time-series data will achieve state-of-the-art performance in early prediction of CEs in SV infants.

1.4 CONTRIBUTIONS

We assert that the work presented in this dissertation is significant in four ways. First, we filled the research gap and addressed the unmet clinical need of achieving accurate, early prediction of patient deterioration in a complex, severely-ill pediatric population. Second, we identified novel features that capture the temporal progression of physiological variables and may improve prediction of patient deterioration. Third, we studied and evaluated three modeling strategies including (1) the use of domain expert knowledge, (2) a combination of domain expert knowledge and data-driven modeling, and (3) data-driven modeling including dynamic models that capture the temporal dynamics of physiological data. This approach may be replicated to detect patient deterioration in different ICU populations and hospitals. Fourth and final, we performed the first multi-site validation of models for the real-time prediction of patient deterioration in the single-ventricle population.

2.0 BACKGROUND

2.1 SUPERVISED MACHINE LEARNING

This section presents a short summary of the machine leaning methods necessary for the work in this dissertation.

Machine learning is a subdiscipline of artificial intelligence in which an agent or computer system learns and improves its performance in any given task after making observations. There are three main types of machine learning depending on the kind of feedback that the learning agent obtains while performing the desired task. In unsupervised learning, the agent makes inferences or learns patterns from observed data without receiving any feedback. A common task in this type of learning is to identify groups of instances in observed data that are similar to each other, but differ from those in other groups. This technique is generally-known as clustering, and has been used in biomedical research in tasks such as the identification of novel cancer subtypes from gene-expression data[20]. In reinforcement learning, the agent receives a "reward" or "punishment" as feedback when performing its task. The agent then adapts its behavior to maximize the amount of reward received. This style of learning has been used recently in biomedical research for real-time 3D-landmark detection in CT scans[21]. Finally, the third main type of learning, which will be the one utilized in this dissertation, is known as supervised machine learning[19].

In supervised machine learning, the agent observes input-output example pairs and learns to predict the output of future inputs. Specifically, given a training set of *N* input-output examples $(x_1, y_1), \ldots, (x_N, y_N)$, where the outputs y_i are generated by an unknown function f, such that $y_i = f(x_i)$, the learning task consists in finding a hypothesis function $h(x_i) = \widehat{y_i}$ that approximates the unknown function f. While the agent is learning, it receives feedback in the form of the true output values, which it can compare to its own forecasts. Then, ideally, the learned function *h* will approximate f closely, and will accurately predict the output of input instances even in novel examples not seen during training.

When the output *y* has values that are continuous or discrete numbers, the learning problem is known as regression. Alternatively, when the outcome can take one or many of a finite set of values (e.g., presence or absence of a disease), the learning problem is known a classification.

In this dissertation, we will used supervised classification algorithms to predict whether patient states (training examples consisting of values of a set of physiological variables) correspond to patients who will experience critical events, namely emergent endotracheal intubation, extracorporeal-membrane oxygenation cannulation, or cardiopulmonary resuscitation in a population of single-ventricle infants. Specifically, we will use Naïve Bayes, decision trees, random forests, support-vector machines, and long short-term memory neural network classifiers. The remainder of this section provides a brief introduction of each of these learning algorithms.

2.1.1 Naïve Bayes classifiers

Naïve Bayes (NB) classifiers are a special case of Bayesian Networks with a strong (*naïve*) independence assumption, namely that child nodes (features) in the network are conditionally independent given their parent (class) node[22,23]. They are a type of generative model, i.e., they

learn a joint probability distribution $P(x, y)$ from pair of inputs x and outputs y. Then, they predict the most likely class label from a training example by using Bayes theorem to compute the posterior probability of the class note given an input x.

Consider a training dataset of instance vectors $\boldsymbol{X_i} =< x_{i1}, \dots, x_{iN} >$ where x_{ij} is the value of the *j-th* out of *N* random variables (features) for the *i-th* training instance, and each vector $\boldsymbol{X_i}$ is associated with a label (class) y_i. Given the naïve independence assumption, the probability chain rule can be used to express the conditional probability of $\boldsymbol{X_i}$ given y_i with the equation below.

$$P(X_i\ |y_i) = \prod_{j=1}^{N} P(x_{ij}|y_i)$$

For a simplified case where the class is a binary variable (i.e., it can be either True of False), the Bayes theorem can be used to compute the posterior probability of the class y_i after observing an input $\boldsymbol{X_i}$ as expressed in the equation below.

$$P(y_i = True|\boldsymbol{X_i} = x_{i1}, \dots, x_{iN})$$

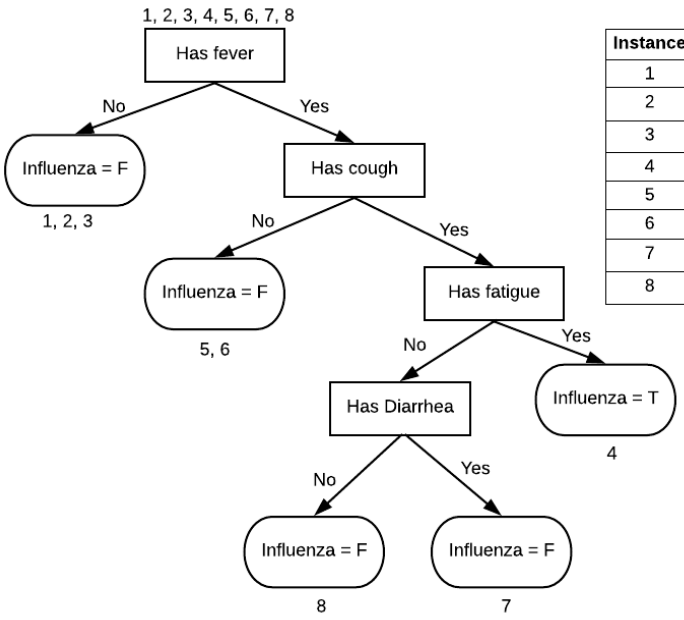
$$= \frac{\prod_{j=1}^{n} P\big(x_{ij}\big|y_i = True\big)P(y_i = True)}{\prod_{j=1}^{n} P\big(x_{ij}\big|y_i = True\big)P(y_i = True) + \prod_{j=1}^{n} P\big(x_{ij}\big|y_i = False\big)P(y_i = False)}$$

Naïve Bayes models have been used in biomedical research since the 1960's and are well-suited for clinical applications[24,25]. Furthermore, although the naïve independence assumption may be unrealistic, empirical evaluation of naïve Bayes classifiers suggests that classification performance is not dependent on the degree of correlation between model features[26].

2.1.2 Decision trees and random forest classifiers

Decision trees are a common approach to multistage decision making, i.e., problems in which a complex decision is broken up into simpler decisions applied in succession. They can be seen as a method that combines different models and where a single model is tasked with making a prediction (classification or regression) for any point in the input space. [27–29]

Consider a dataset of jointly-distributed input-output pairs $\boldsymbol{X}$ and $\boldsymbol{Y}$, where $\boldsymbol{X}$ is an input vector of N components (features), and $\boldsymbol{Y}$ is the class label associated with $\boldsymbol{X}$. The process of choosing a model for $\boldsymbol{X}$ can be seen as a sequential decision process equivalent to the traversal of a binary tree, such as the one shown in **Figure 1**. In this example, the process of model selection starts at the root node, which splits the input space into two regions based on the presence of fever. The region where fever is absent can be further sub-divided into two regions based on the presence of cough, and the process can continue recursively after either all possible sub-divisions have been exhausted or a stopping criterion is met.



Instance	Fever	Cough	Fatigue	Diarrhea	Influenza
1	F	F	F	F	F
2	F	T	F	F	F
3	F	T	T	T	F
4	T	T	T	T	T
5	T	F	T	F	F
6	T	F	F	F	F
7	T	T	F	F	F
8	T	T	F	T	T

Figure 1. Decision tree for influenza case detection. This decision tree depicts a hypothetical decision process that may be used by clinicians to ascertain whether patients' symptoms are indicative an influenza case (adapted from [30]). The input space in this example consists of four-dimensional Boolean vectors such that each dimension corresponds to symptoms (fever, cough, fatigue, diarrhea). The output space consists of a Boolean variable indicating whether influenza is True of False.

The process of learning the structure of a decision tree model from (***X, Y***) can be described as a two-phase process, namely growing and pruning. During the growing phase, the input space is partitioned recursively until either (1) the process reaches a decision tree in which every leaf node is associated with instances of the same class, or (2) a stopping criterion is met, e.g., further partitioning of the training data would result in leaves with a number of instances lower than a specified threshold. This process is described in **Figure 2**. The pruning phase consists of collapsing

branch splits into a single node (leaf), and testing whether the resulting tree improves a given metric, e.g., complexity cost or classification error[27,31].

```
Algorithm: Grown phase during decision tree classifier induction
function grow_tree(X, Y, features)
Input: Training instances X of dimension |features|, class labels Y, and a list of available variables feature
Output: Root of inducted decision tree
Root = Node containing all instances in X
if all instances at Root have the same class do
        create leaf node of corresponding class
        return Root
else do
        Find feature A that maximizes a predefined goodness metric
        foreach value v of A do
                Add new branch below node testing for A = v
                v_instances = subset of instances from X s.t A = v
                v_Y = class labels associated with v_instances
                if v_instances == {∅} do
                        add leaf whose label is the most common value in Y
                else do
                        #below branch add subtree
                        grow_tree(v_instances, v_Y, features – {A})
                end
        end
end
return Root
```

Figure 2. Pseudo code of the growth phase of the induction of a decision tree classifier

As seen in **Figure 2**, during the tree growth phase, a *goodness* metric is used to select the feature used to partition the input space at any given node (and in the case of continuous inputs, also to select the threshold for partitioning based on said feature). Two common choices of feature-selection metrics are the information gain score and Gini impurity, as shown in the equations below.

- Information gain

$$Gain(S,A) = Entropy(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Where S is the set of training instances at the node being split, S_v is the set of instances s.t. $A = v$, and Entropy is defined in the equation below.

$$Entropy(S) = -\sum_{v \in values(Y)} P(v) \log_2 P(v)$$

Where $P(v)$ is the probability of class v in S

- Gini impurity

$$Gini(S) = \sum_{v \in values(Y)} P(v)(1 - P(v))$$

Where S is the set of training instances at the node being split and $P(v)$ is the probability of class v in S

Random forests[32] are an ensemble method classifier built from multiple decision trees. These classifiers predict the class of training instances by aggregating the output of N_{tree} individual decision trees. These trees are inducted from N_{tree} bootstrap samples of the training dataset, respectively, and at every split node the best splitting variable is selected from a random set of all available features.

2.1.3 Support vector machines

Support vector machines (SVMs)[29] are a decision-margin maximization classifier. Consider a training dataset of N input vectors $X_1, \dots, X_N$ with corresponding class labels $y_1, \dots, y_N \in \{1, -1\}$, and suppose that there exists a linear classifier whose output is given by the expression below.

$$f(x) = W^T \phi(x) + b$$

Where ϕ is a function that transforms the feature space and b is a bias term.

Assuming that the dataset is linearly separable in its feature space, there may exist multiple solutions for W and b that satisfy the constraint that $y_i * f(x_i) > 0$. Support vector machines strive to minimize generalization error by maximizing the distance between training distances and a hyperplane that separates said instances according to their class label, also known as the decision boundary depicted in **Figure 3**. Specifically, SVMs maximize the distance from the decision boundary to the closest training instances (support vectors).

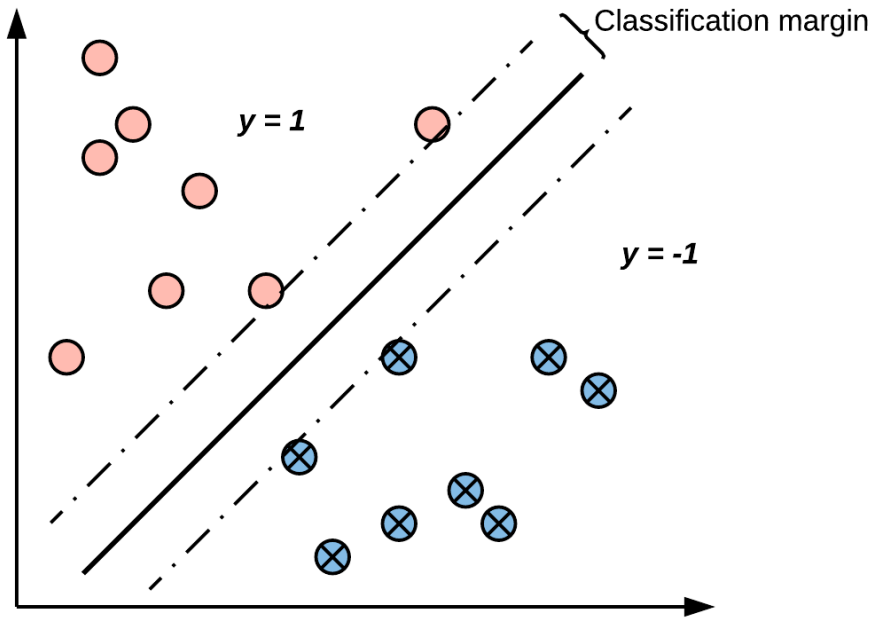


Figure 3. Decision boundary representation for a binary classification task. The solid diagonal line is a hyperplane that separates data instances according to class label $y \in \{-1, 1\}$, known as the decision boundary. Support vector machines minimize generalization error by maximizing the minimum perpendicular distance between the decision boundary and data instances closest to it. This distance is known as the classification margin.

Generally, data instances may not be linearly separable in their original feature space. SVMs then rely on kernel functions that map instances into a higher-dimensional space such that linear separation is achieved. Common choices of kernel functions include the linear, radial basis function, and sigmoid kernels. In this general scenario, SVMs assign a class label to an input data instance $\boldsymbol{x}$ with the expression below[33].

$$f(\boldsymbol{x}) = sign(\sum_{i=1}^{N} y_i \alpha_i K(\boldsymbol{x_i x}) + b)$$

Where x_i and y_i are the training instances and their associated class labels, K is the kernel function, and α and b are learned model parameters.

2.1.4 Long short-term memory neural networks

Long short-term memory (LSTM)[34] networks are a variant of recurrent neural networks (RNN). RNNs are well-suited for learning from temporal data because through their feedback connections they can store evolving representations of input sequences. However, there is a limit to the length of dependencies that RNNs can learn from sequential data. In the process of 'back propagation through time' (BPTT), i.e., the iterative optimization process by which RNN weights are learned[35], gradients (errors being propagated) tend to increase or decrease exponentially. When this happens, the network loses its ability to learn from new inputs, a phenomenon known as the 'vanishing and exploding gradients' problem[36].

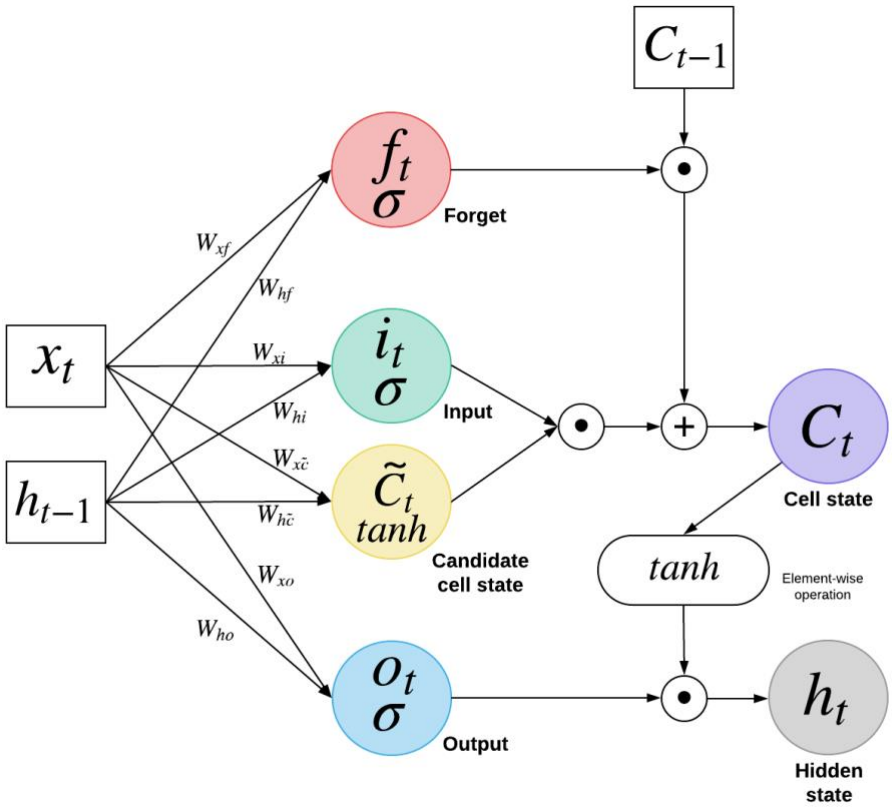


Figure 4. Long short-term memory neural network architecture. The forget (f_t), input (i_t), and output (o_t) gates, as well as the candidate cell state ($\widetilde{C}_t$) are fully-connected layers whose input comprise the input data (x_t) and a feedback connection to the previous hidden state (h_{t-1}). Activations of the cell state and the hidden state are computed via point-wise product, addition, and *tanh* operations. The LSTM gates use sigmoid and tanh activations functions, as depicted in the figure. The dimensions of all quantities are available in **Definition 1.**

The LSTM architecture addresses this limitation by guarantying a constant flow of gradients through the use of special 'units'[34]. These include the cell state, hidden state, input gate, output gate, and forget gate, as depicted in Figure 4. At a given time t (where t indexes data sequentially), the 'cell state' C_t stores an evolving representation of the data and acts as 'long-

term' memory. Analogous to long-term memory in humans, C_t it's not used in its entirety to respond to new inputs. Instead, the LSTM unit relies on the hidden state h_t, which acts like short-term or 'working' memory by storing a transformation of the cell state dependent upon input data.

When an LSTM network receives a new input at time t, it evolves in three different ways based on the input x_t and the last activations of the hidden state h_{t-1}. First, the forget gate f_t determines what should be forgotten from the long-term memory C_t. Second, the input gate i_t determines which elements of a candidate cell state $\widetilde{C_t}$ should be committed to long memory. Third and finally, the output gate o_t determines which elements of the cell state should be 'loaded' into the hidden state. The hidden state may be then connected to a classifier layer to generate predictions based on all previous inputs. The mathematical definitions of the units in the LSTM cell are described in **Definition 1.**

Definition 1. Long short-term memory units

Let X be a dataset of sequential data instances, T be the size of sequential batches of data instances, t be the index of the *t-th* batch of data instances $x_t \subset X$, and P be the number of features of x_t. Let also $|C|$ be the cardinality (number of neurons) of C_t, C_t be the cell state, h_t the hidden state, f_t the forget gate, i_t the input gate, $\widetilde{C_t}$ the candidate cell state, and o_t the output gate of the LSTM cell at time t. Unit activations are determined by the following expressions:

$$
\begin{aligned}
f_t &= \sigma\left(W_{hf} * h_{t-1} + W_{xf} * x_t + b_f\right) \\
i_t &= \sigma(W_{hi} * h_{t-1} + W_{xi} * x_t + b_i) \\
\widetilde{C_t} &= \tanh(W_{hc} * h_{t-1} + W_{xc} * x_t + b_c) \\
C_t &= f_t \cdot C_{t-1} + i_t \cdot \widetilde{C_t} \\
o_t &= \sigma(W_{ho} * h_{t-1} + W_{xo} * x_t + b_o) \\
h_t &= o_t \cdot \tanh(C_t)
\end{aligned}
$$

Where $\{f_t, i_t, o_t, \widetilde{C_t}, C_t, h_t\} \in \mathbb{R}^{Tx|C|}$, $x_t \in \mathbb{R}^{TxP}$, $\{W_{xf}, W_{xi}, W_{xo}\} \in \mathbb{R}^{Px|C|}$, and $\{W_{hf}, W_{hi}, W_{ho}\} \in \mathbb{R}^{|C|x|C|}$

As seen in the equations above, LSTM units use sigmoid and *tanh* activation functions. Although the selections of activations or parameter initialization strategies are known to influence

the predictive performance of neural networks, the original LSTM architecture specification has not been surpassed by alternative configurations [37].

2.2 FREQUENT TEMPORAL PATTERN MINING (FTP)

The FTP mining process can be broadly summarized into four steps: (1) temporal abstraction, (2) state sequence representation, (3) temporal-pattern representation, and (4) FTP mining. Below, we present each step in detail, as available in [38–40].

2.2.1 Temporal abstraction (TA)

TA is the process of mapping a timestamped series of variable values into a sequence of higher-level concepts that represent some temporal aspect of the original series.

Definition 2. Temporal abstraction

TA is a sequence $< v_i[s_i, e_i]: s_i \leq e_i \wedge\ i \in \mathbb{N} >$, where $v_i \in \sum$ is an abstraction valid between times s_i and e_i.

TAs can be categorized in terms of their duration. Interval (trend) TAs hold during a specified interval ($s_i > e_i$), while point (value) TAs hold only at a specified time point ($s_i = e_i$). **Figure 5** shows a time series of diastolic blood pressure values (**Figure 5**(a)), as well as examples of possible interval and point TAs. These include a "gradient" TA, which indicates whether values in the series are increasing, decreasing, or remain constant during a specified interval (**Figure 5**(b)). While this gradient TA maps value changes to a discrete, 3-value set {-1, 0, 1}, indicating the value-change direction, an alternative gradient TA could compute the slope of two continuous

measurements. **Figure 5**(c) and **Figure 5**(d) show point TAs that discretize instant variable values into {Low, Normal, High} or {Normal, Abnormal} values, respectively.

It is evident that TAs can be used to explicitly introduce expert knowledge into the FTP-extraction process. For instance, a point TA could use expert-defined cut-points to discretize diastolic pressure values into {Low, Normal, High} bins, as was done in **Figure 5**(c).

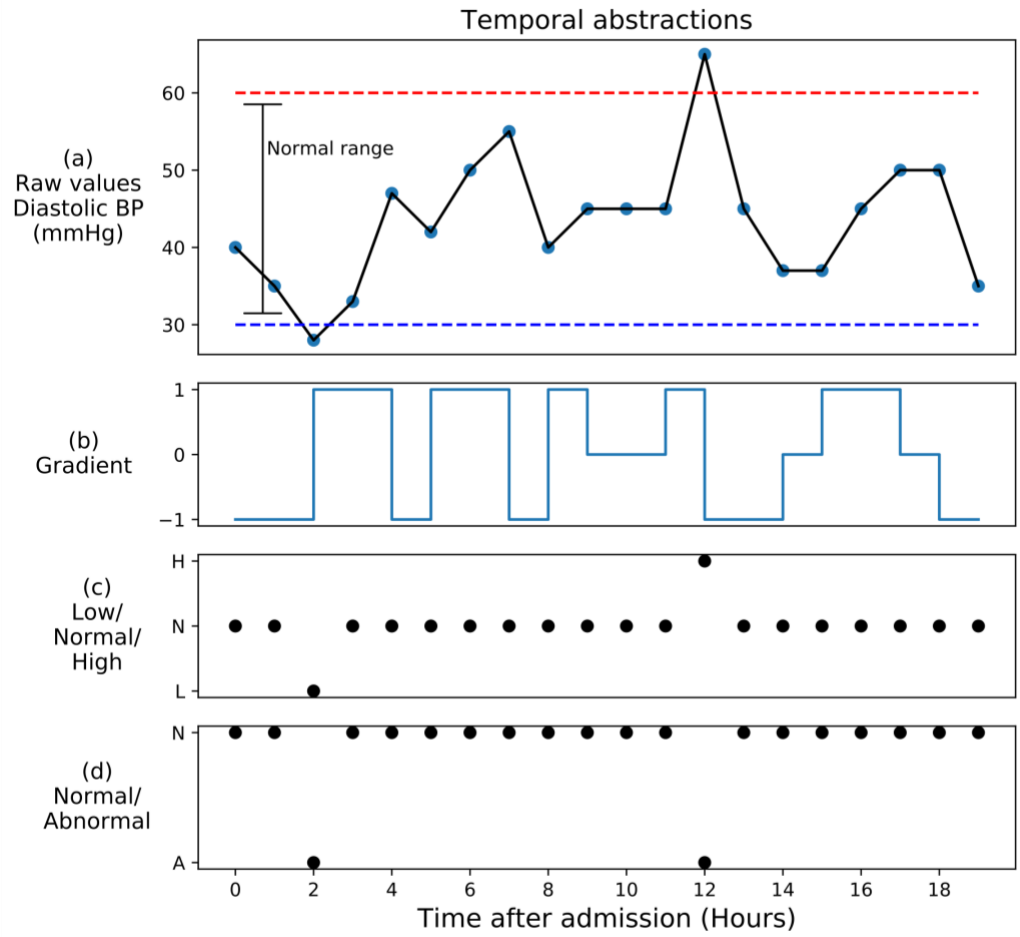


Figure 5. Temporal abstractions from time-series data

2.2.2 State-sequence representation

A multivariate state sequence (MSS) is an ordered sequence that aggregates all time-series-variables' TAs into a single array.

> **Definition 3. Multivariate state sequence (MSS)**
>
> A MSS is a finite sequence $Z =< E_i : i \in \mathbb{N} \wedge E_i.s \leq E_{i+1}.s >$, Where E_i is a state interval (F, V, s, e), F is a time-series variable (e.g., blood pressure), and $V \in \Sigma$ is an abstraction function that holds in the interval between times s and e.s

It follows from **Definition 3** that state intervals in an MSS are ordered by their start time, regardless of their end time, i.e., E_{i+1} may start before the end time of E_i. Once the time-series for each variable have been abstracted individually, a patients' record can be represented by a single MSS. **Figure 6** shows TAs for heart rate and oxygen saturation time-series values from the same patient. The resulting MSS from these TAs would be $Z =< (HR,\ High, 0, 3),$ $(O_2sat, Normal, 1, 4),\ (HR, Normal, 3, 5),\ (O_2sat, High, 4, 6),\ (HR, Low, 5, 7),$ $(O_2sat, Normal, 6, 9), (HR, High, 7, 9) >.$

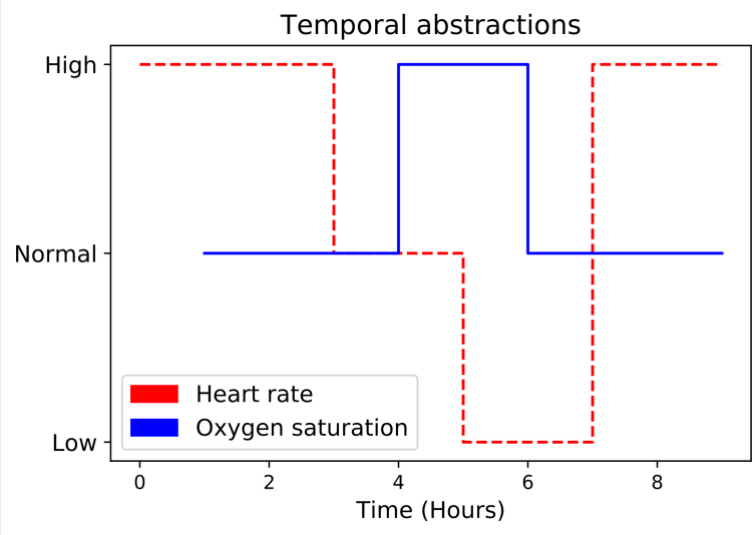


Figure 6. Multivariate temporal abstractions

2.2.3 Temporal-pattern representation

Temporal patterns (TP) are sequence of states, i.e., (F, V) pairs (see **Definition 3**) ordered in terms of their temporal relations. The most widely-adopted set of temporal relations were proposed by James Allen in a seminal publication of a formalism based on temporal logic[40]. Allen proposed a total of seven possible relationships, which are based on the start and end times of state intervals. They are listed below.

- *X before Y*: X ends before Y starts
- *X equals Y*: X and Y have the same start and end times
- *X meets Y*: Y starts at the same time that X ends
- *X overlaps* Y: X starts before Y, and there is a non-zero overlap between X's and Y's intervals.
- *X during Y*: X starts after and ends before Y.
- *X starts Y*: Y starts at the same time as but ends before Y.

- *X finishes Y*: X starts after and ends at the same time as Y.

Typically, only a subset of these temporal relations is used in the context of TP mining. For instance, due to documentation lags in clinical time-series data, asserting that two state intervals end at the same time is not a reliable statement. Furthermore, some of the relations described above may lead to the extraction of different patterns with very similar clinical interpretations, a phenomenon known as pattern fragmentation[38]. Hence, a simplified subset of temporal relations is often used. Specifically, we will use the following two temporal relations.

- *X before (b) Y*: X starts before Y, regardless of their end times.
- *X and Y co-occur (c)*: X starts before Y. There is a non-zero overlap between X's and Y's time intervals.

Definition 4. Temporal pattern

A temporal pattern (TP) is a sequence $P = (< S_i : i \in \mathbb{N} \wedge S_i.start \leq S_{i+1}.start >, R)$, where S_i is a temporal-abstraction state. Consecutive states in a TP are temporally related, and their relationships are specified by an upper-triangular matrix R_{NxN}, where N is the number of states S_i, and $R_{ij} \in \{before\ (b), co-occur\ (c)\}$ is the temporal relationship between states S_i and S_j. A *k-TP* is a TP of length *k*.

A MSS Z is said to contain a temporal pattern TP P if all states in P exist in Z, and all temporal relations R in P are satisfied in Z. For instance, consider two MSS $Z_1 =< (HR, Normal, 0,\ 2), (BP, Normal, 3, 5), (HR, High, 4, 7) >$ and $Z_2 =< (HR, Normal, 0, 2.5), (BP, Normal, 3, 4), (HR, High, 5, 8) >$, and two TPs $P_1 =< (HR, Normal)\boldsymbol{c}(BP, Normal) >$ and $P_2 =< (BP, Normal)\boldsymbol{b}(HR, High) >$, where $\boldsymbol{b}$ and $\boldsymbol{c}$ are the before and co-occur temporal relations, respectively, HR is heart rate, and BP is blood pressure. We can assert that both Z_1 and Z_2 contain P_1. However, only Z_2 contain P_2, because while the before relationship between

$(BP, Normal)$ and $(HR, High)$ is satisfied, in Z_2, it is not in Z_1, where the co-occur relation exist between those two states.

Definition 5. Recent temporal pattern

Let P be a TP $(< (S_i): i = 1, 2, \dots, n >, R)$, and g be a quantity of time units, or gap. P is a recent temporal pattern (RTP) in a MSS $Z =< (E_i: i = 1, 2, \dots, k) >$, if for a specified g, all the following conditions are met:

1. Z contains P
2. S_n can be mapped to a recent state interval in Z, i.e., $\exists E_i = (S_n, start, end): E_i.end \leq E_k.end$.
3. Every consecutive pair of sates S_i and S_{i+1} can be mapped to a pair of consecutive state intervals $E_l, E_{l+1} \in Z: E_l.end - E_{l+1}.start \leq g$

Definition 6 Horizontal and vertical support

Let D be a dataset of N patient records, abstracted into a sequence of MSS, i.e., $D =< MSS_i: i = 1,2, \dots, N >$, D_y be the subset of MSS in D labeled with class y, and g the maximum gap to define TPs as RTPs. The horizontal support of a RTP P in MSS_i is the count of times that P is contained in MSS_i, noted as $hsup_g(P, MSS_i)$. Similarly, the vertical support of P in D_y, or $vsup_g(P, D_y)$ is the number of MSS_i in D_y that contain P at least once.

From **Definition 5** and **Definition 6**, it follows that a recent frequent temporal pattern (RFTP), is a RTP such that $vsup_g(P, D_y) \geq \delta$.

2.2.4 FTP mining and vectoral feature representation

FTP mining is a two-stage process. First, all possible FTPs from a dataset D of N MSS are identified by means of a pattern-generation routine. Second, the horizontal and vertical support of each FTP is computed, and a matrix M_{NxK} is generated, where K is the number of FTPs whose vertical support is above a pre-specified threshold δ, and M_{ij} quantifies the association between

the *i-th* MSS and the *j-th* FTP. This quantification may be defined in several ways, including a binary indication (1 if a MSS contains a FTP and 0 otherwise), or a real-valued metric such as horizontal support or FTP mean-duration for each MSS. Detailed routines for pattern-candidate generation and FTP-mining are available in [38].

3.0 AIM 1: EXPERT MODEL DEVELOPMENT

In Aim 1 we focused on two goals. First, the elicitation of knowledge from pediatric cardiologists that specialize in the treatment of infants with SV physiology. Second, we developed a predictive Bayesian model that encodes the knowledge elicited from experts.

3.1 METHODS

3.1.1 Expert knowledge elicitation

We consulted experts including two pediatric cardiologists and two critical-care nursing specialists. Elicited expert knowledge was comprised of (1) a list of all clinical variables that experts believed are relevant for the prediction of CEs, (2) a list of discretization ranges for each numeric variable, and (3) a quantification of the interaction between expert-selected variables and the risk of experiencing CEs.

Expert variables

We asked experts to individually answer the following question: "Based on your clinical experience, what variables do you rely on when assessing a patient's risk of experiencing a CE, i.e., EEI, ECMO, or CPR?" Then, we conducted an interview with all experts present where they reviewed each variable and determined if they could be used in real time for assessing patient

deterioration. Discussion was encouraged during this session for experts to identify new variables that were not included in their original answers.

Discretization ranges and risk estimation

When dealing with SV patients, what is considered an abnormal value for a physiological variable may differ from the general infant population. For example, whereas an oxygen saturation of 85% in a SV infant who has undergone stage 1 palliative surgery may be considered favorable, it would be considered a hypoxemic event for an infant without SV physiology. After compiling the final list of expert variables, we asked the two senior experts (pediatric cardiologists) to provide a list of meaningful value ranges for all numeric variables.

Quantification of interaction between expert variables and risk of CEs

We elicited from each senior expert via a self-administered, computer-assisted questionnaire their expert estimation of the distribution of values for each expert variable in a hypothetical cohort of 100 cases (patients at risk of CEs) and 100 controls. For the remainder of this dissertation, we will refer to these experts as ‘expert 1’ and ‘expert 2’. Each expert answered the questionnaire individually without discussing their answers with each other. A sample of the questionnaire is shown in **Figure 7**.

Clinical variable	In a group of 100 patients **at risk** of critical events, how many are expected to have the following values?			In a group of 100 patients **at NO risk** of critical events, how many are expected to have the following values?		
Heart rate (bpm)	HR ≤ 120	120 < HR ≤ 160	HR > 160	HR ≤ 120	120 < HR ≤ 160	HR > 160
	40	20	40	15	70	15
Systolic blood pressure (mmHg)	SBP < 60	60 < SBP ≤ 90	SBP > 90	SBP < 60	60 < SBP ≤ 90	SBP > 90
	40	20	40	10	80	10
Oxygen saturation (%)	O_2sat ≤ 70	70 < O_2sat ≤ 85	O_2sat > 85	O_2sat ≤ 70	70 < O_2sat ≤ 85	O_2sat > 85
	40	20	40	15	70	15

Instructions: Answers should be based on infants with SV, younger than 6 months of age, before undergoing state 2 palliative surgery. Critical events (CEs) include emergent intubation, ECMO cannulation, and administration of cardiopulmonary resuscitation (CPR) outside the operating room. For each row of answers, orange cells (case group) and green cells (control group) must each add to 100.

Figure 7. Self-administered questionnaire used to elicit experts' estimation of interaction between expert variables and the risk of experiencing critical events. This sample includes three out of 54 variables identified by experts as relevant for the prediction of critical events for infants with SV physiology.

3.1.2 Expert model construction

We built a Naïve Bayesian network model (NB) from the answers provided by experts to the questionnaire described in **Figure 7**. This was achieved by first using questionnaire answers as discrete conditional probability tables (CPTs) and then using said CPTs to parametrize a discrete NB network as shown in **Figure 8**. This process was repeated for each senior expert, resulting in two expert models, NB-expert-1 and NB-expert-2.

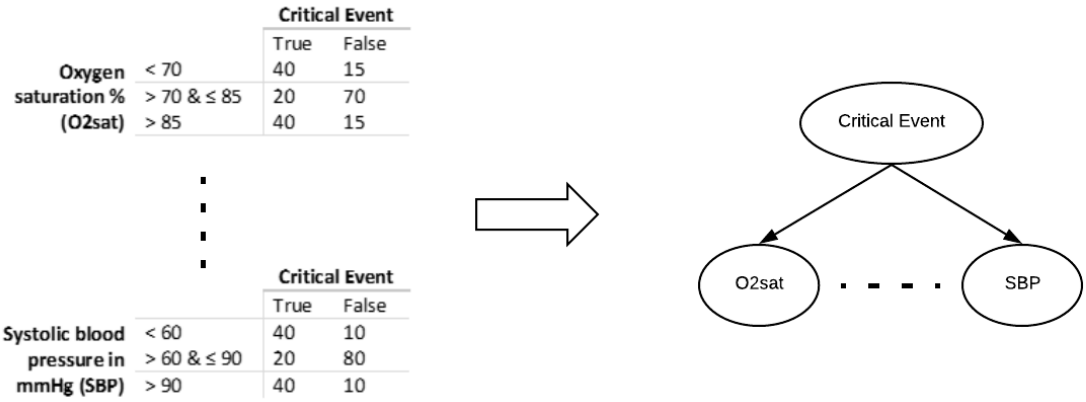


Figure 8. Construction of naïve Bayes model from experts' questionnaire answers.

3.2 RESULTS

Expert variables

Experts identified a total of 52 variables as relevant for the prediction of CEs in SV infants. These variables include laboratory test results (e.g., creatinine, bicarbonate ion), blood gasses (e.g., carbon dioxide, oxygen saturation), vital signs (e.g., respiratory rate, diastolic blood pressure), surgical-related factors (e.g., sternal closure time, Blalock-Taussig shunt abnormalities), and imaging-test-related variables (e.g., chest X-ray effusion, electrocardiogram ST-segment elevation or depression greater than 1 mm). These variables were deemed relevant for predicting a combined outcome of EEI, ECMO, and CPR events. Although outside of the scope of this dissertation, experts also identified a subset of variables that they considered important for the prediction of each type of event individually. Finally, experts identified sixteen variables as the minimum set of relevant predictors of CEs. A complete list of variables, as well as details of variables relevant for individual CE types is available in **Table 14**.

Quantification of interaction between expert variables and risk of CEs

Each expert provided contingency tables for each variable for a combined outcome of EEI, ECMO, and CPR. The full list of variables, as well as the contingency tables for expert 1 are presented in **Table 14.**

4.0 AIM 2: DEVELOPMENT OF STATE-OF-THE-ART MODELS FOR PREDICTION OF CRITICAL EVENTS IN SV INFANTS

Aim 2 had four main objectives. First, the retrieval of a longitudinal dataset of SV ICU admissions containing the variables identified by experts as relevant in Aim 1. Second, the re-parametrization of expert models using the retrieved SV dataset. Third, the derivation of two sets of features, namely trend-summary and temporal-abstraction features. Fourth and final, the development of static and dynamic classifiers from derived features and raw time-series data, respectively.

4.1 METHODS

4.1.1 Retrieval of SV cohort and expert-identified variable values from CHP's EHR system

After approval by the institutional review board (PRO17020157), which included a waiver of collection of informed consent, we retrieved clinical data from infants admitted to the CICU, PICU, and NICU units at the Children's Hospital of Pittsburgh of UPMC (CHP). The inclusion criteria were (1) age less than six months, (2) hospital admission between January 1, 2014 and August 30, 2017, and (3) a primary diagnosis of single ventricle (SV) physiology, i.e., any diagnostic ICD-9 code amongst 745.3, 746.1, 746.3, 746.5, 746.7, 746.01, 747.22, or any ICD-10

code amongst Q22.0, Q23.4, Q20.4, Q22.6, Q23.2. We excluded hospital admissions of patients that had already undergone second surgical palliation (Bidirectional Glenn) at the time of admission. Hence, clinical data used for model development an evaluation only included ICU encounters of SV patients prior to second-stage repair.

4.1.1.1 Outcome definitions

Critical event (CE) cases

We defined cases as any instance of EEI, ECMO cannulation, or CPR that occurred while patients were admitted to the CICU, PICU, or NICU, and that occurred at least eight hours after patients' first ICU admission. We considered multiple CEs experienced by the same patient as separate cases if they occurred at least eight hours after the end time of any previous CE. We excluded cases that happened within less than eight hours after ICU admission or the end of another CE for evaluation purposes. Thus, we ensure that predictive performance comparisons between different prediction horizons (e.g., at 2 hours vs at 8 hours before CEs) are made on the same set of instances.

We defined the start and end times of CEs by means of timestamps that we retrieved retrospectively from the Cerner® EHR system at CHP. For intubation events, we defined the start time (intubation) as the time when a change of airway from natural to artificial was documented in nursing charts, and the end time (extubation) as the time when a change of air-way from artificial to natural was documented. Airway changes were documented manually by nurses as part of routine care. For ECMO events, we defined start and end times as the times of cannulation and de-cannulation, respectively. For CPR events, we defined start times as the time when any keyword amongst *arrest, arrest sheet, arrest code, chest compressions, condition A, or CPR* were documented in nursing arrest sheets.

Non-event controls

We defined controls as periods of ICU stays longer than 24 hours from patients who did not experience CEs during their hospital admission. We divided long ICU stays into multiple 24-hour period windows and considered each window as a distinct control. The data retrieval and the case and control definition processes are summarized in **Figure 9**.

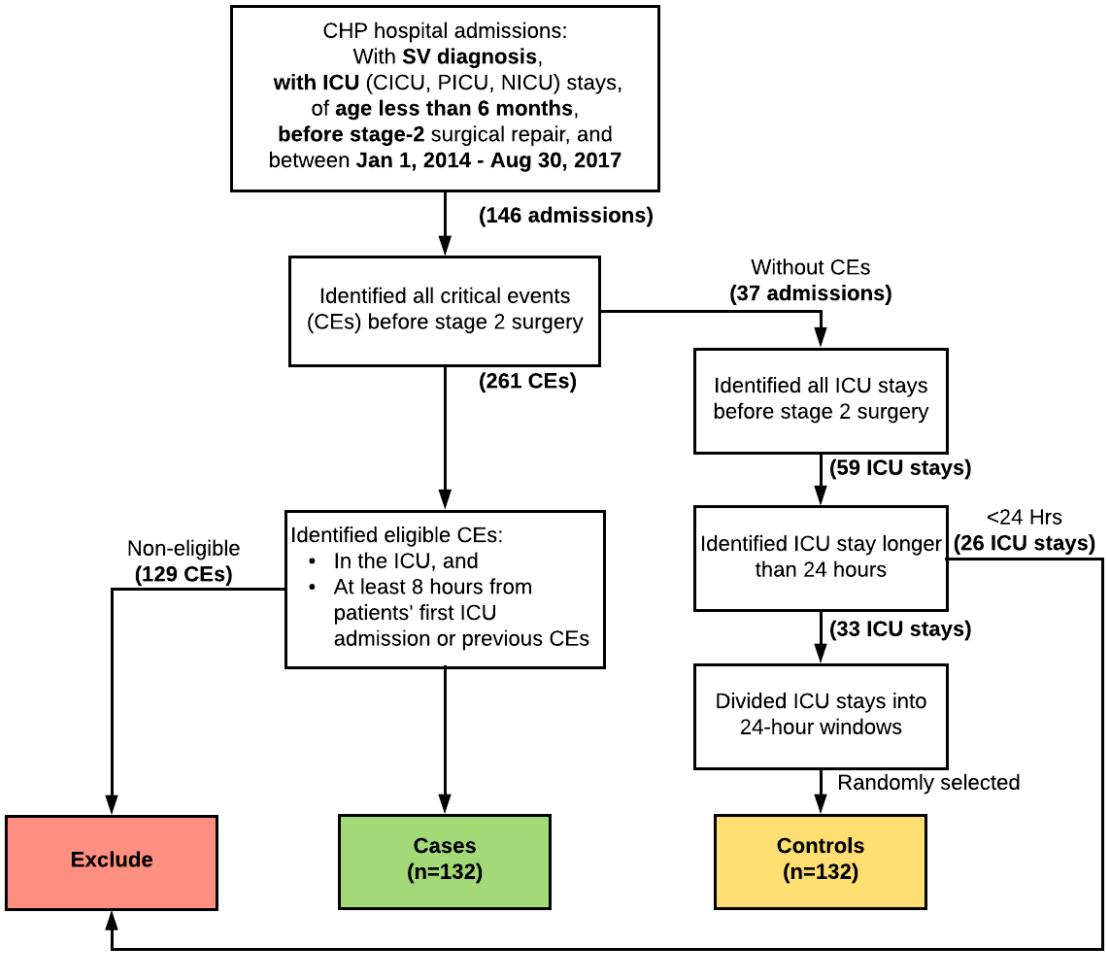


Figure 9. Data retrieval and case/control definition process. Data retrieval and case-control definition process. Multiple critical events (CE) during the same ICU stay were considered as different cases if they occurred at least eight hours after either ICU admission or the end of a previous CE. Non-eventful ICU stays of at least forty hours of duration were selected as controls. Long ICU stays were divided into 24-hour windows, and all windows considered as separate controls. We randomly selected 132 control windows to match the number of CE cases

4.1.1.2 Mapping of clinical events in EHR data to expert variables

With help from the Information Technology team at CHP, we compiled a list of codes from the EHR's clinical event table that could represent any expert variable. We then asked experts to validate the code list.

We found that several EHR event codes may represent a single expert variable. For example, diastolic blood pressure (DBP) was represented by at least two local EHR event codes: 'arterial diastolic pressure', which represents an invasive DBP measurement, and 'diastolic blood pressure' which represents a noninvasive measurement. We identified all local EHR codes that represented every expert-identified variable and aggregated their measurements.

The concurrent availability of multiple EHR codes for a single expert variable presented a practical problem, i.e., how to select a single value for a variable at any given time. We addressed this issue by defining priority levels for each variable in consultation with experts. We then used priority levels to choose a single value whenever values for multiple EHR codes were available for the same expert variable simultaneously. Following the previous example, when both invasive and non-invasive DBP were available simultaneously, we used invasive DBP values and discarded non-invasive measurements; if only non-invasive DBP was available, we used it as the value for the DBP variable.

4.1.1.3 Processing of expert variable values

Uniform time interval resampling

Clinical variables are usually measured at irregular time intervals (frequencies). While some variables (e.g., blood pressure, heart rate) are measured approximately every hour, other variables (e.g., lactate, base excess, creatinine, BUN, BNP) may be measured at irregular intervals of several hours. To estimate the risk of a critical event regular time intervals, we resampled variable values

in uniform steps of 30 minutes, ending at the time of presentation of CEs for cases and at the time of ICU discharge for controls. When multiple values for the same variable were available within the same 30-minute window, we used their mean value.

Missing value imputation

Unavailability of variable values during each 30-minute window leads to missing data issues, caused by the nature of how clinical information becomes available. We imputed missing values with first the last known value for the same variable up to six hours in the past, and second with the variable's mean value if no previous observations were available for any given case or control. We assumed that variable values measured within the last six hours still reflected the state of a patient and could be used in the model to compute the risk of critical events.

4.1.2 Re-parametrization of expert model using clinical data

The baseline expert model used CPTs defined by domain experts as described in section 3.1.2. This had the purpose of explicitly encoding expert knowledge into predictive models. However, expert-defined CPTs are susceptible to cognitive biases. We re-parametrized CPTs with maximum-likelihood estimates derived from retrieved clinical data. Additionally, we used information gain scores[41] with an empirical threshold of 0.01 to perform feature selection. Thus, we created four additional models, NB-ML-full-1 and NB-ML-full-2 which used all available features, and NB-ML-lean-1 and NB-ML-lean-2 which used features selected with information gain scores. The '1' and '2' suffixes denote the discretization bins used in each model, i.e., variable ranges defined by experts 1 and 2, respectively.

4.1.3 Derivation and ranking of temporal features

The expert models developed in Aim 1 used mostly instantaneous, cross-sectional measurements (the last value known for each variable at the time of prediction). However, experts indicated that variable value trends are also important for clinical judgment, and suggested the inclusion of two trend variables, namely creatinine and SvO_2 changes from baseline values.

We aimed at improving upon the feature space utilized to evaluate our baseline (expert) and re-parametrized NB models by extracting temporal-abstraction features from time-series data. First, we generated trend-summary features. Then, we identified multivariate frequent temporal patterns. We describe these two approaches below.

4.1.3.1 Univariate trend-summary features

In the first level of temporal abstraction, we derived a subset of the univariate trend-summary features proposed by Valko and Hauskrecht[42] from the SV dataset. These features summarize the time-series data available in a patient's EHR into an a-temporal vector representation suitable for static machine-learning classifiers.

The original set of trend-summary features was validated in two contexts, namely, the prediction of physician orders[42], and the detection and alerting of anomalous patient-management decisions.[43] Although related, these prediction tasks are different to the prediction of CEs insofar they are affected by routine-care processes. For instance, the time since a laboratory test was last ordered may predict that the same laboratory is likely to be ordered again because some tests are ordered periodically. This feature however, does not reflect the state or progression of an individual patient. In contrast, the last value of a laboratory test result, or the highest observed value of said test *does* reflect the state of individual patients. Hence, we focused on the extraction of summary

features that reflect patient states and ignore those that may be indicative of routine-care workflows. **Table 1** lists and defines 14 trend-summary features that we extracted from numeric variables. An example of a time series for this data type is shown in **Figure 10**.

Table 1. Trend-summary feature definitions for numeric variables

Feature	*Description*	*Definition*
y_last	Last value	y_f
y_diff_last2	Last value difference	$y_f - y_{f-1}$
y_diff_percent_last2	Last percentage change	$\frac{y_f - y_{f-1}}{y_{f-1}}$
y_nadir	Nadir (lowest value)	y_{min}
y_diff_nadir	Nadir difference	$y_f - y_{min}$
y_diff_percent_nadir	Nadir percentage difference	$\frac{y_f - y_{min}}{y_{\min}}$
y_apex	Apex (highest value)	y_{max}
y_diff_apex	Apex difference	$y_f - y_{max}$
y_diff_percent_apex	Apex percentage difference	$\frac{y_f - y_{max}}{y_{\max}}$
y_first	Baseline (first value)	y_0
y_diff_first	Drop from baseline	$y_f - y_0$
y_diff_percent_first	Drop from baseline percentage	$\frac{y_f - y_0}{y_0}$
y_avg_lastwindow	Average of the N values observed during the last W-hours window, where W is the window's width	$\frac{\sum_{i=1}^{N} y_i}{N}$
y_slope_last2	Slope of the last 2 values	$\frac{y_f - y_{f-1}}{x_f - x_{f-1}}$

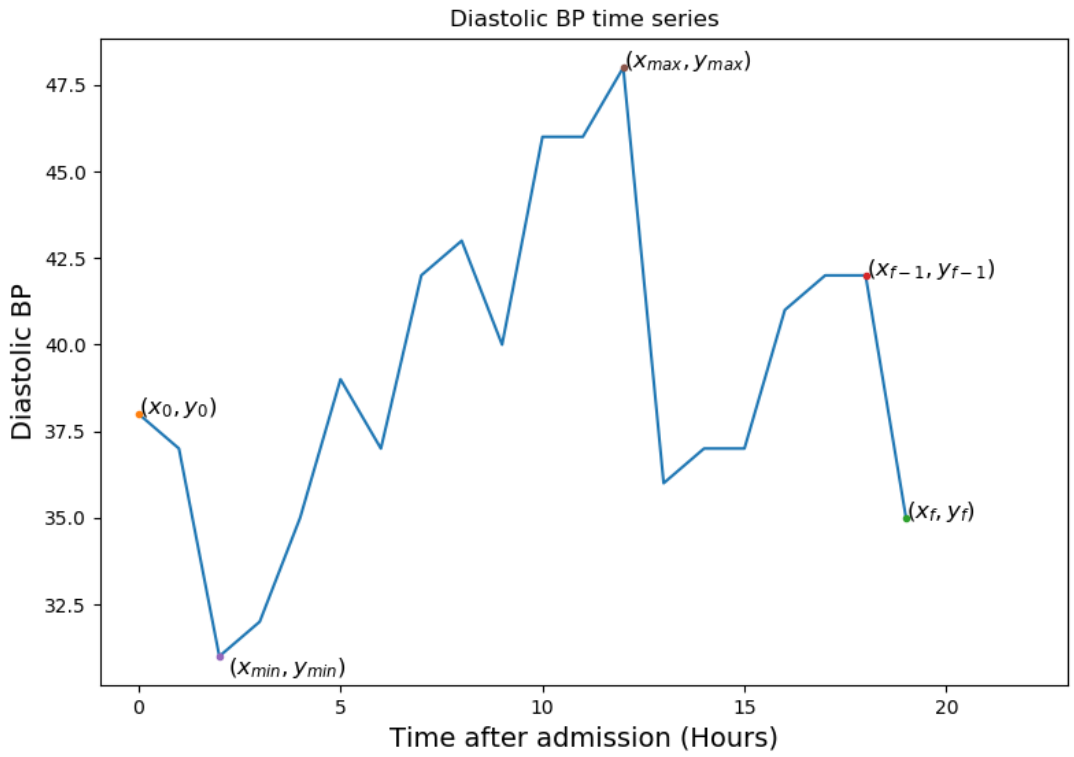


Figure 10. Diastolic blood pressure time series. BP: blood pressure; x_0, y_0: time and value of first BP measurement; x_{min}, y_{min} time and value of the minimum BP measurement (nadir); x_{max}, x_{max} time and value of the maximum BP measurement (apex); x_f, y_f: time and value of last BP measurement; x_{f-1}, y_{f-1}: time and value of the next-to-last BP measurement.

4.1.3.2 Extraction of multivariate frequent temporal patterns

We extracted multivariate frequent temporal patterns (FTP) as a second level of temporal abstraction. The building blocks for this technique were proposed by Yuval Shahar in his seminal publication[44], which provided a domain-independent framework for the abstraction of temporal concepts from time-series data. In recent years, FTP-mining algorithms have been proposed and evaluated in the clinical domain in the context of prediction of heparin-induced thrombocytopenia [38,45], disease diagnoses in diabetic patients[46], hepatitis type after diagnosis[47,48], and administration of outcome-event procedures[49]. A detailed description of the temporal abstraction (TA) and FTP

extraction methodology is available in section 2.2, and a high-level description of the mining process is shown in **Figure 11**.

Algorithm: High-level description of frequent temporal pattern mining

$function\ RFTP\_mining(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Kmax}, \boldsymbol{\sigma}, \boldsymbol{\delta}, \boldsymbol{g})$

Input :

- $\boldsymbol{X}$: training time-series data
- $\boldsymbol{Y}$: Class labels associated with each instance in X
- $\boldsymbol{Kmax}$: maximum pattern length
- $\boldsymbol{\sigma}$: minimum support (percentage of instances for which an FTP is consistent) to determine whether a temporal pattern is frequent
- $\boldsymbol{\delta}$: Bayesian score threshold to determine whether an FTP is predictive
- $\boldsymbol{g}$: maximum gap (in time units) between states in an FTP or between the last state and prediction horizon to determine whether an FTP is recent (RFTP)

Output:

- $\boldsymbol{RFTPs}$: recent frequent temporal patterns in $\boldsymbol{X}$
- $\boldsymbol{FTP\_matrix}$: vectoral binary features for training static classifiers based on FTP features

$RFTPs = \{\}$
$X_{train} = Random\ sample\ (80\%)\ of\ X$
$X_{val} = \{x_i \in X : x_i \notin X_{train}, i = 1, \ldots |X|\}$

foreach $label \in values(Y)$ **do**
    $label\_data = \{x_i \in X_{train} : y_i \in Y_{train} = label, i = 1, \ldots |X|\}$
    $k = 0$
    while $k < K\_max$ **do**
        $k++$
        $candidates = consistent\_KFTPs(label\_data, k)$
        $kFTPs = \{TP \in candidates : BS(TP, X_{val}) \geq \delta\ and\ support(TP) \geq \sigma\ and\ \text{MaxGap(TP)} \leq \text{g}\}$
        $RFTPs = RFTPs \cup kFTPs$
    end
end

$FTP_{matrix} = Binary\ matrix\ of\ dimension\ |X| x |RFTPs|$
foreach i in $1, 2, \ldots, |X|$ **do**
    foreach j in $1, 2, \ldots, |RFTPs|$ **do**
        $FTP_{matrix}(i, j) = \begin{cases} 1\ if\ x_i \in X\ contains\ RFTP_i \in RFTPs \\ 0\ otherwise \end{cases}$
    end
end

return $RFTPs, FTP\_matrix$

Figure 11. High level description of the recent temporal pattern mining process. Temporal patterns are comprised of a sequence of temporal-abstraction states *P* and a temporal-relations matrix *R* that specifies the temporal relations between any two states in *P*. Recent frequent

temporal patterns (FTPs) are those whose last state is at most g time unites from the time of prediction. An FTP is consistent is there is at least one multivariate state sequence MSS in the training set such that MSS contains all the states in P and those states satisfy the temporal relations in R. The support of an FTP is the number of instances in the training set for which the pattern is consistent. The Bayesian score of an FTP measures how predictive a pattern is for a class data instances of label y compared to a more general group of instances (e.g., complete training dataset). This Bayesian score was first proposed by Batal et al.[50] and subsequently applied to the scoring of FTPs in clinical datasets[38].

We utilized three types of temporal abstractions, namely (1) discretization of variable values into bins using the expert-defined ranges identified in section 3.1.1, which we will refer to as 'ExpertBins', (2) indicators of whether variable values are increasing, decreasing, or stable, as depicted in **Figure 5**(b), which we will refer to as 'DiscreteGradient', and (3) discretized number of standard deviations away from the mean, which we will refer to as 'NDeviations'. Abstractions of the type NDeviations were discretized by rounding to the nearest integer value and were capped at two deviations. Thus, values of this abstraction may take one of five values, i.e., $\{\leq -2, , -1, 0, 1, >= 2\}$. We mined FTPs using each abstraction type separately, and also combining ExpertBins and DiscreteGradient, and ExpertBins and NDeviations abstractions, respectively.

4.1.4 Predictive model training

After we built our baseline expert models, which we improved by using retrieved clinical data to re-compute said models' CPTs, we built predictive models for four additional sets of features, namely (1) static, cross-sectional variable values known at the time of prediction, which we will

refer to as ‘LastNumericValues’; (2) static, cross-sectional variable values plus the trend-summary features described in section 4.1.3.1, which we will refer to as ‘TrendSummaries’; (3) binary features from mined frequent temporal patterns as described in **Figure 11**, which we will refer to as ‘FTPs’; and (4) raw time-series data for each variable up until the time of prediction, which we will refer to as ‘TimeSeries’. From LastNumericValues, TrendSummaries, and FTPs features, we trained naïve Bayes (NB), decision trees (DT), random forests (RF), and support vector machine (SVM) classifiers. Additionally, for those classifiers, we performed feature selection by ranking features using reliefF[51], information gain[41], and feature-importance score derived from a fitted random-forest classifier with Gini-index splitting[52], and then selecting the best *k* features for model training. From TimeSeries data, we trained long short-term memory networks (LSTMs) without prior feature selection. A high-level description of the model building process is shown in **Figure 12**.

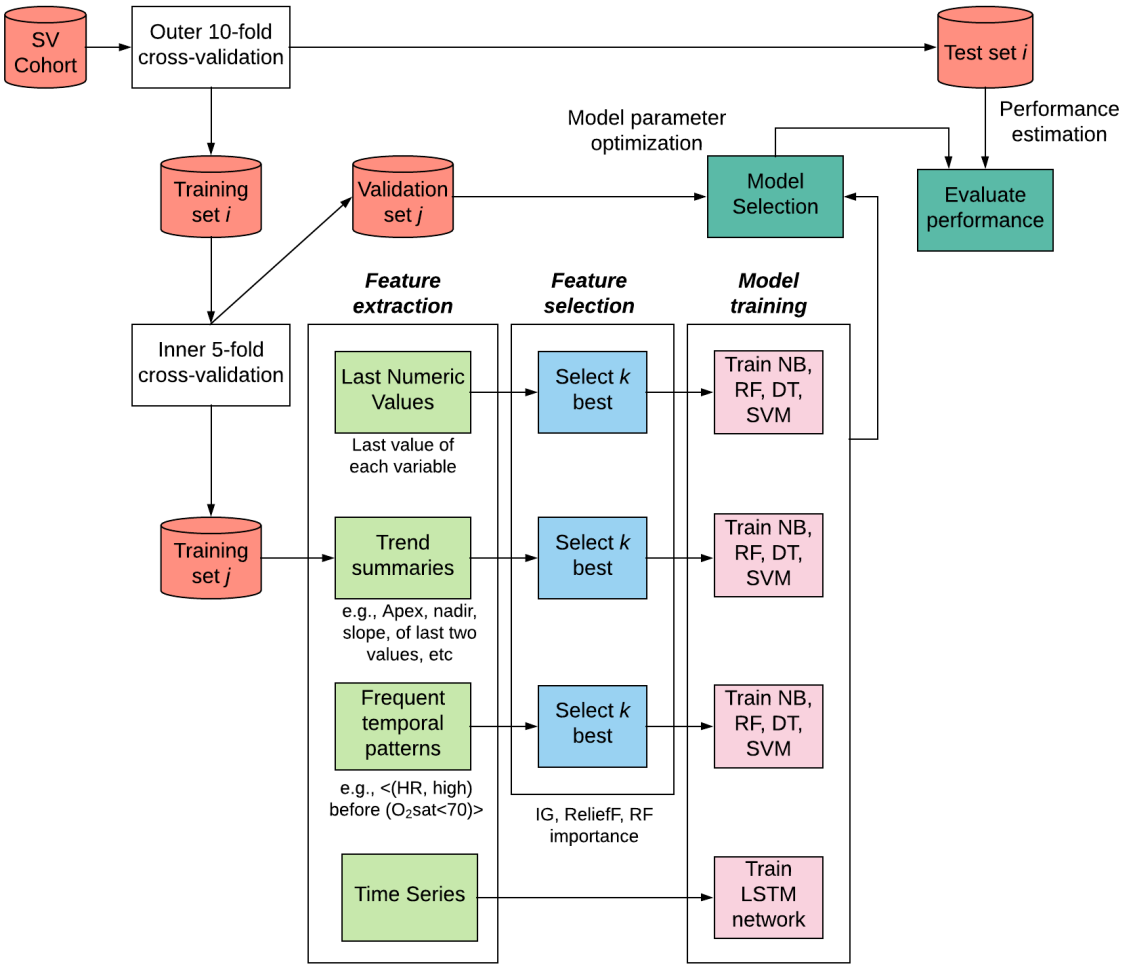


Figure 12. Predictive-model building process. We trained and evaluated predictive models in nested cross-validation. We measured performance in the outer 10-fold cross-validation, and performed classifier hyper-parameter optimization in the inner 5-fold cross-validation. We trained NB, RF, DT, and SVM classifiers from three different types of features, namely 'LastNumericValues', which are cross-sectional a-temporal patient states; Trend-summaries, which are static vectoral representation of training instances that include features that summarize temporal trends; and frequent temporal patterns, which are multi-variate sequence of variable states and corresponding temporal relations. SV: Single-ventricle physiology; O2Sat: oxygen

saturation; HR: heart rate; NB: naïve Bayes; DT: decision tree; RF: random forest; SVM: support vector machine.

For static classifiers (NB, DT, RF, SVM) we pre-processed data as follows:

- We standardized variable values to the [0, 1] range by using the following expression

$$f(x) = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- We Imputed missing variable values with previously-known values within the previous six hours. If previous values were not available, we imputed the mean of the variable in the training dataset.

For dynamic (LSTM) classifiers, we pre-processed data as follows:

- We transformed data into sequential instances, i.e., a 3-dimensional dataset of shape *D*x*T*x*Q*, where the *D* is the number of data instances (cases or controls), *T* is the number of time steps allowed in each instance (30-minute windows), and *Q* is the number of variables available for training.
- We standardized variable values to the [0, 1] range as described above.
- We imputed missing values with -1. The rationale for this is that given that all values were scaled to the [0, 1] range, an LSTM classifier, through backpropagation, should learn that -1 has a special meaning, i.e., missing data.

4.2 RESULTS

4.2.1 Retrieval of SV cohort

During the study period, we identified 146 hospital admissions of patients with SV diagnosis, of age of less than six months, and with ICU admissions before stage-2 palliation. These corresponded to 120 patients who experienced 261CEs. Ninety-five patients and 132 CEs met our inclusion criteria for analysis, i.e., they occurred in the ICU at least eight hours after patients' first ICU stay during a hospital admission or after previous CEs. The set of included CEs was comprised of 119 EEIs, 9 ECMO, and 4 CPR events.

Patients did not experience any CEs in 37 of the included hospital admissions. We identified 33 ICU stays during these hospital admissions that lasted 24 hours or longer. We divided long ICU stays into 24-hour control periods, and randomly-selected 132 controls to match the number of included CEs, as shown in **Figure 9**. The case and control groups were comprised of 77 and 26 unique patients, respectively. Eight patients were present in both the case and control groups. This occurred because these patients had at least two hospital admissions, among which there was at least one admission with CEs and one admission without CEs.

While most patients had one CE while admitted to the ICU, some patients experienced up to eight CEs during a single hospital admission. **Figure 13** shows the distribution of the number of CEs per hospital admission as well as the number of eligible cases depending on the amount of available data, i.e., the amount of time from patient's first ICU or a previous CE, and the onset of a given CE.

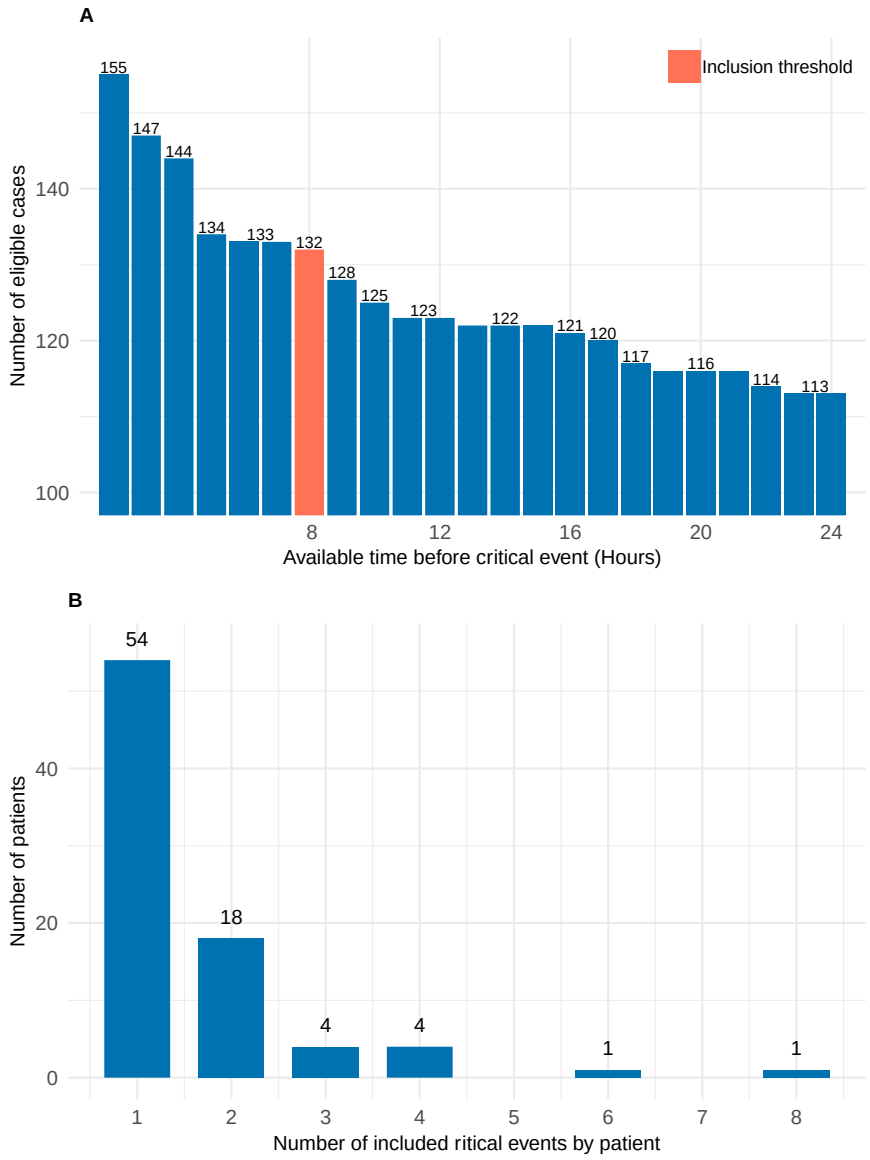


Figure 13. Distribution of critical events by data availability and inpatient encounter. A) Number of cases (critical events) available for analysis depending on data availability, i.e., the time between the first ICU admission or the end of a previous critical event and the onset of a critical event. **B**) Distribution of the number of included critical events by hospital admission.

Hypoplastic left heart syndrome was the most common primary diagnosis (42.1% of study population) followed by mitral stenosis (19%), as shown in **Table 2**. Among patients' palliation

procedures, Norwood was the most common (28.4% of study population) followed by Blalock-Taussig shunt (15.8%), as shown in **Table 3**. The majority (37.4%) of included CEs occurred seven or more days after patients underwent palliative procedures. **Figure 14** shows the distribution of times when CEs occurred relative to palliative procedures.

Table 2. Distribution of primary diagnosis among included single ventricle patients

Diagnosis	Number of patients (%)
Hypoplastic left heart syndrome	40 (42.1%)
Congenital mitral stenosis	18 (19%)
Pulmonary valve atresia	12 (12.7%)
Congenital atresia and stenosis of aorta	7 (7.4%)
Hypoplastic right heart syndrome	6 (6.3%)
Tricuspid atresia and stenosis, congenital	5 (5.3%)
Common ventricle	3 (3.2%)
Congenital stenosis of aortic valve	2 (2.1%)
Double inlet left ventricle	2 (2.1%)
Total	**95 (100%)**

Table 3. Classification of single-ventricle patients in the study population by type of palliative surgical procedure

Type of procedure	Number of patients (%)
Norwood	27 (28.4%)
Modified Blalock-Taussig shunt	15 (15.8%)
Other	15 (15.8%)
Pulmonary artery banding	9 (9.5%)
Hybrid	7 (7.4%)
Bi-directional Glenn	2 (2.1%)
Non-surgical	20 (21%)
Total	**95 (100%)**

Patients in the non-surgical category met the inclusion criteria for analysis but did not undergo palliative procedures during the study period. Two patients underwent bi-directional Glenn as their

first palliative procedure. However, only ICU admissions before patients' stage-2 repair were used for model development and evaluation. Procedures in the 'other' category included aortic arch repair, tetralogy of Fallot repair after Blalock-Taussig shunt, right ventricle to pulmonary artery conduit repair, aortopexia, valvuloplasty, stent placement in catheterization laboratory, atrioventricular septal defect repair after pulmonary artery banding and aortic arch repair, and Nikaidoh after Blalock-Taussig shunt.

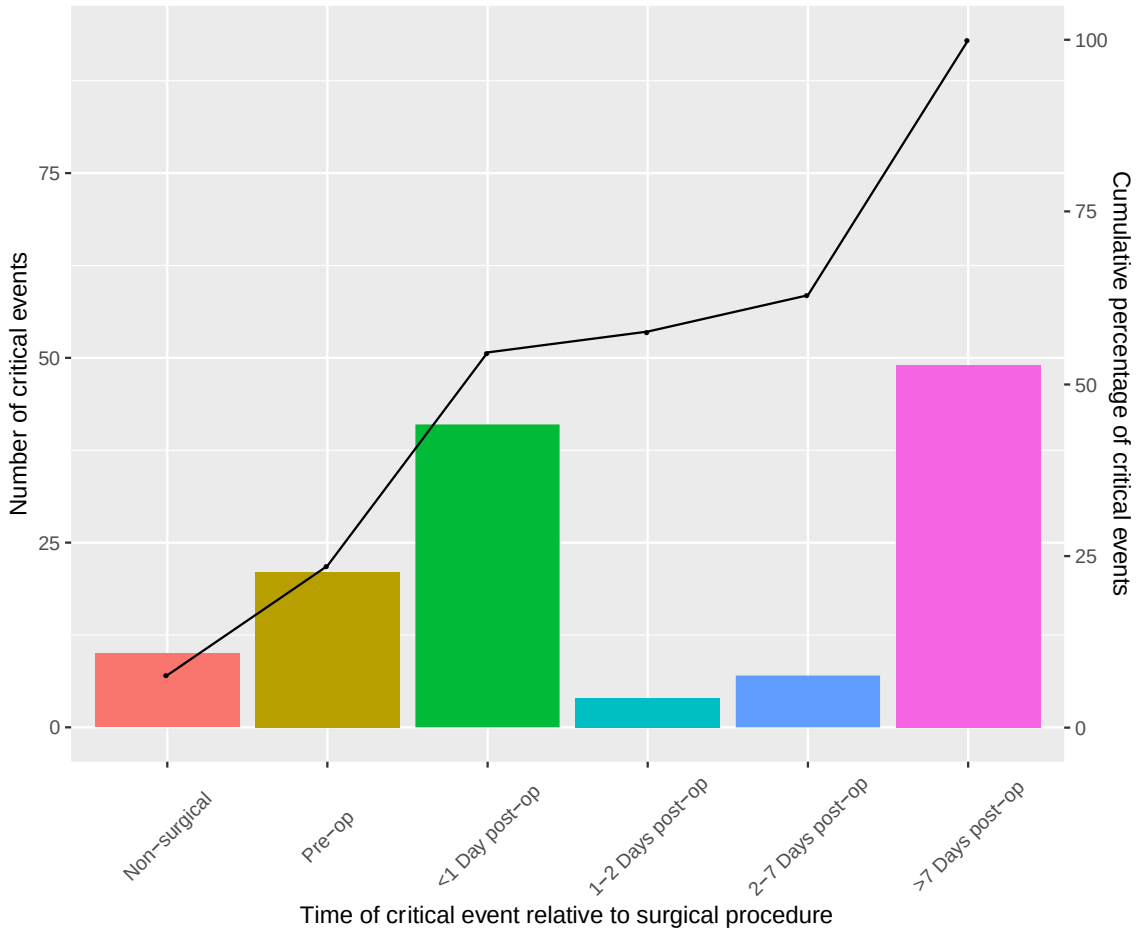


Figure 14. Time of presentation of critical events relative to the time of palliative procedures. Critical events included 132 emergent endotracheal intubations, extracorporeal-membrane oxygenation cannulations, and cardiopulmonary resuscitations experienced by single-ventricle infants before stage-2 palliative surgery. The non-surgical category corresponds to critical events experienced by patients who did not undergo palliative procedures during the study period. The black line shows a cumulative percentage of critical events.

4.2.2 Mapping of clinical events in CHP's EHR data to expert variables

We identified 68 concepts in CHP's Cerner® EHR system that mapped to 34 of the 52 variables identified by clinical experts. The 18 variables that could not be retrieved retrospectively included mainly imaging and surgery-related variables, e.g., echocardiogram coarctation of the aorta, electrocardiogram ST segment elevation or depression greater than 1 mm, or chest X-ray cardiomegaly. **Table 14** and **Table 15** include the full list of expert variables, as well as the list of variables that could be mapped to EHR concepts. **Table 15** also includes the priority rules used to choose variable values when multiple EHR were available simultaneously.

5.0 AIM 3: EVALUATION OF PREDICTIVE MODELS

In this specific aim, we measured and compared the predictive performance of the models trained in Aims 1 and 2, namely (1) models based exclusively on expert knowledge (Expert), (2) models combining expert knowledge and data-derived CPTs (ExpertRetrained), (3) models based on a-temporal, cross-sectional patient states (LastValsNumeric), (4) models incorporating trend-summary features (TrendSummaries), (5) models based on frequent temporal pattern mining, and (6) dynamic (LSTM) models based on the time series data (TimeSeries).

5.1 METHODS

5.1.1 Internal validation

We measured the performance of the models built in chapters 3.0 and 4.0 in nested cross-validation, as shown in **Figure 12**. The outer validation was performed with stratified 10-fold cross-validation, and estimated model performance. The inner validation was performed in stratified 5-fold cross-validation and was used for model hyper-parameter optimization and model selection.

We measured model discrimination with the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, f-measure, and accuracy metrics, and we measured the statistical significance between AUC differences by means of false discovery rate (FDR)-corrected[53] two-sided DeLong tests[54]. We assessed model calibration with the Brier skill score[55], Hosmer-Lemeshow test[56], and calibration curves. **Table 4** shows the list of experiments conducted in this evaluation.

Table 4. Experiments for evaluation of predictive models

Experiment Name	Description of feature set	Models	Experiment description
Expert	Cross-sectional patient states with the last value available at the time of prediction for each variable. Variables were discretized with expert-defined bins	NB-Expert	Models built from expert knowledge. Full and Lean models used all available features or a minimal set of features identified by experts, respectively.
ExpertRetrained		NB-ML	Models built with expert-discretization and CPTs estimated from clinical data. Full and Lean models used all available features, or features selected based on information gain, respectively
LastValsNumeric	Cross-sectional patient states with the last value available at the time of prediction for each variable. All variable values are continuous.	NB, DT, RF, SVM	Models built with a-temporal, cross-sectional representations of patient states
TrendSummaries	Cross-sectional patient states with last variable values augmented with features that summarize temporal trends	NB, DT, RF, SVM	We developed models with two feature sets, (1) *with_filter*: Last values + trend-summary features from variables for which at least 25% of the training dataset had at least two available values; (2) *no_filter:* Last values + trend-summary features from all variables.
FTPs	Vectoral binary indication of whether training instances contained mined FTPs	NB, DT, RF, SVM	We built models from FTPs mined from five sets of temporal abstractions: ExpertBins, DiscreteGradient, NDevaitions, ExpertBins + DiscreteGradient, and NDeviations + DiscreteGradient.
TimeSeries	Sequential time series where all variables are uniformly sampled and all instances have the same sequence length	LSTM	We trained recurrent (LSTM) neural networks from continuous time-series data.

We implemented Expert and ExpertRetrained models using the WEKA data mining software[57]. We implemented NB, DT, RF, and SVM models in the LastValsNumeric, TrendSummaries, and FTPs experiments using the SciKit-learn machine learning framework[58]. We implemented LSTM models in the TimeSeries experiment using the Keras deep-learning framework[59]. We computed AUCs and conducted DeLong tests with the pROC R package[60]. We

computed BSSs and Hosmer-Lemeshow values with the Verification and ResourceSelection R packages, respectively[61,62].

5.1.2 External validation of predictive models

We measured the AUC of the predictive models developed from data from UPMC Children's Hospital of Pittsburgh (CHP) and that were evaluated in section 5.1.1 on an external dataset of SV admissions to the Children's Hospital of Philadelphia (CHOP). This dataset was comprised of hospital admissions to CHOP between January 1, 2015 and September 30, 2018. During this period, 466 patients were admitted to an ICU before undergoing bidirectional Glenn surgery. We identified 385 CEs, of which 164 happened at least eight hours after patients' first ICU admission or presentation of a previous CE, and were included in the test set for external validation. These CEs included 161 EEIs and 3 ECMO events. In the same fashion as in the models trained with CHP data, we selected 164 controls to match the number of cases in the CHOP dataset.

5.2 RESULTS

5.2.1 Performance of predictive models trained on CHP data

5.2.1.1 Expert models

Expert-based models achieved modest performance from one to four hours before CEs, as shown in **Table 5**. With the exception of the full models at 6 hours before CEs, the NB-expert-1 model achieved equal or higher AUCs than those of the NB-expert-2 model for all prediction horizons

except for four hours before CEs. Full and lean models for both experts lost all predictive ability at six hours before CEs.

All expert models exhibited poor calibration, as indicated by the Brier skill scores and Hosmer-Lemeshow test p-values shown in **Table 5**. The calibration curves of the NB-expert1-full model for all prediction horizons are shown in **Figure 15**.

Table 5. Prediction performance of Expert Naïve Bayes models

Metric	Horizon	NB-Expert1-Full	NB-Expert2-Full	NB-Expert1-Lean	NB-Expert2-Lean
AUC	-1	**0.67 (0.6-0.74)**	0.58 (0.51-0.65)	0.6 (0.53-0.67)	0.5 (0.43-0.57)
	-2	**0.71 (0.64-0.77)**	0.61 (0.54-0.67)	0.64 (0.57-0.71)	0.54 (0.46-0.61)
	-4	**0.58 (0.51-0.65)**	0.57 (0.51-0.64)	0.48 (0.41-0.56)	0.5 (0.43-0.58)
	-6	0.52 (0.45-0.59)	**0.54 (0.47-0.61)**	0.45 (0.38-0.52)	0.49 (0.42-0.56)
	-8	**0.49 (0.42-0.56)**	0.46 (0.39-0.54)	0.41 (0.34-0.48)	0.42 (0.36-0.49)
BSS	-1	-0.73	-0.36	-0.70	-0.41
	-2	-0.77	-0.15	-0.70	-0.32
	-4	-0.80	-0.53	-0.78	-0.68
	-6	-0.78	-0.60	-0.77	-0.69
	-8	-0.80	-0.58	-0.75	-0.65
HL	-1	0	0	0	0
	-2	0	0	0	0
	-4	0	0	0	0
	-6	0	0	0	0
	-8	0	0	0	0
Best f1	-1	0.67	0.67	0.67	0.67
	-2	0.68	0.67	0.67	0.67
	-4	0.67	0.67	0.67	0.67
	-6	0.67	0.67	0.67	0.67
	-8	0.67	0.67	0.67	0.67
f1 at 0.5	-1	0.07	0.28	0.13	0.23
	-2	0.03	0.49	0.10	0.33
	-4	0.01	0.19	0.04	0.03
	-6	0.03	0.14	0.06	0.03
	-8	0.02	0.11	0.07	0.09
Best accuracy	-1	0.68	0.62	0.62	0.58
	-2	0.69	0.65	0.66	0.59
	-4	0.59	0.60	0.55	0.55
	-6	0.56	0.58	0.52	0.56
	-8	0.53	0.56	0.52	0.53

Values in parentheses show 95% confidence intervals computed with 2000 bootstrap replicates. Bold-face values show the best AUCs for each prediction horizon, i.e., the number of hours before critical events when predictions were generated; NB: Naïve Bayes; AUC: Area under the receiver operating characteristic curve; BSS: Brier skill score; HL: p-value of Hosmer-Lemeshow test.

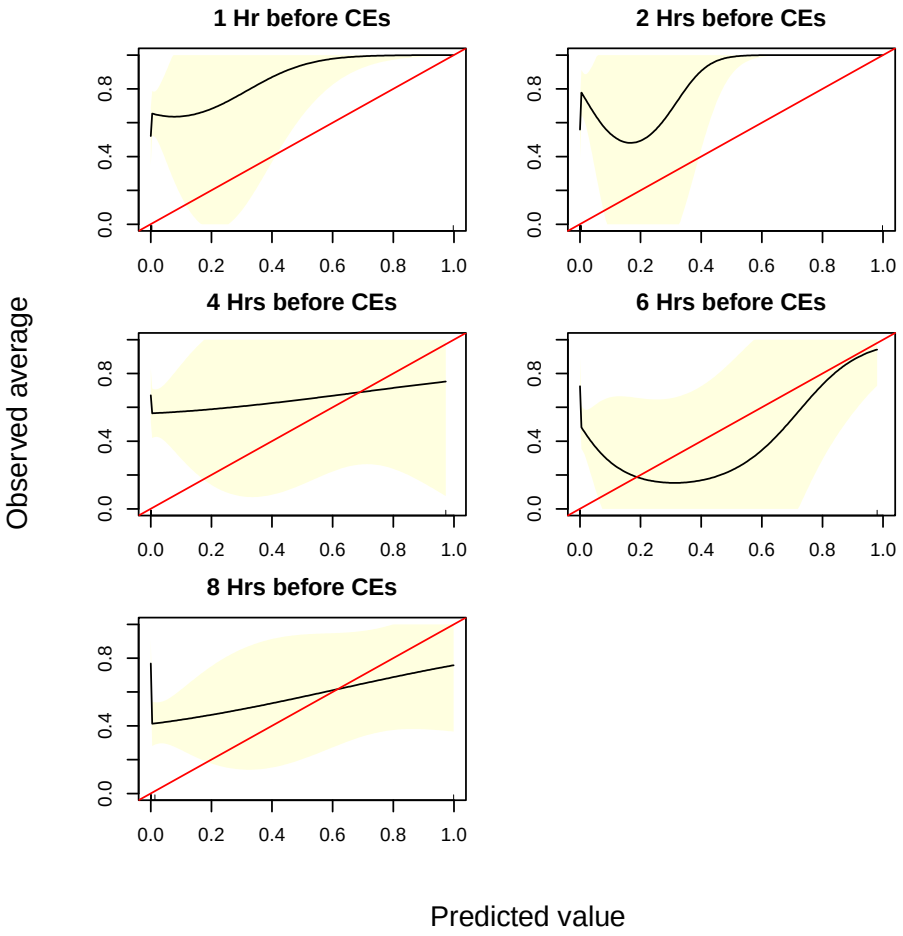


Figure 15. Calibration curves of the NB-expert-1 model across different prediction horizons

5.2.1.2 ExpertRetrained models

The re-parametrized expert models achieved moderate discrimination. As shown in **Table 6**, the NB-ML-1-full model, which used the full list of 34 variables identified by clinical experts and expert bins provided by Expert 1, achieved AUCs between 0.74-0.88. Similarly, the NB-ML-1-lean model, which used variables selected with information gain scores and a threshold of 0.01, achieved AUCs between 0.74-0.87, matching the AUC of the full model with a reduced feature space of 24 variables. Compared to the discrimination performance of the expert models described in **Table 5**, both NB-ML models had consistently higher AUCs than those of the NB-expert-1. In all cases, the difference was statistically significant (adjusted p-value < 0.01).

Table 6. Prediction performance of re-calibrated Naïve Bayes models

Metric	Horizon	NB-ML-1-Full	NB-ML-1-Lean	NB-ML-2-Full	NB-ML-2-Lean
AUC	-1	**0.88 (0.83-0.92)**	0.87 (0.83-0.91)	0.87 (0.82-0.91)	0.87 (0.82-0.91)
	-2	**0.87 (0.82-0.91)**	**0.87 (0.82-0.91)**	0.86 (0.81-0.9)	0.86 (0.81-0.9)
	-4	**0.79 (0.73-0.84)**	**0.79 (0.74-0.85)**	0.78 (0.73-0.84)	**0.79 (0.73-0.84)**
	-6	**0.74 (0.68-0.8)**	**0.74 (0.68-0.8)**	**0.74 (0.68-0.8)**	0.73 (0.67-0.79)
	-8	**0.74 (0.67-0.8)**	**0.74 (0.68-0.8)**	**0.74 (0.68-0.8)**	**0.74 (0.68-0.8)**
BSS	-1	0.40	0.37	0.36	0.34
	-2	0.38	0.37	0.35	0.34
	-4	0.21	0.22	0.19	0.20
	-6	0.13	0.17	0.12	0.14
	-8	0.10	0.13	0.08	0.10
HL	-1	0.00	0.00	0.00	0.00
	-2	0.00	0.00	0.00	0.00
	-4	0.00	0.00	0.00	0.00
	-6	0.00	0.00	0.00	0.00
	-8	0.00	0.00	0.00	0.00
Best f1	-1	0.82	0.82	0.82	0.82
	-2	0.81	0.81	0.80	0.81
	-4	0.75	0.74	0.74	0.74
	-6	0.71	0.72	0.71	0.72
	-8	0.72	0.73	0.73	0.72
f1 at 0.5	-1	0.78	0.77	0.77	0.76
	-2	0.79	0.79	0.78	0.78
	-4	0.72	0.71	0.70	0.70
	-6	0.68	0.68	0.68	0.68
	-8	0.63	0.64	0.63	0.65
Best accuracy	-1	0.83	0.82	0.81	0.81
	-2	0.81	0.82	0.80	0.80
	-4	0.75	0.75	0.75	0.75
	-6	0.72	0.73	0.73	0.73
	-8	0.73	0.73	0.72	0.72

Values in parentheses show 95% confidence intervals computed with 2000 bootstrap replicates. Bold-face values show the best AUCs for each prediction horizon, i.e., the number of hours before critical events when predictions were generated. The NB-ML-Full model used the full set of 34 variables identified by expert clinicians as relevant for the prediction of critical events in single-

ventricle infants. The NB-ML-lean model used a subset of those variables selected with information gain scores with a threshold of 0.01. Both models were trained and evaluated on a cohort of 95 patients and 132 critical events in 10-fold cross-validation. NB: Naïve Bayes; AUC: Area under the receiver operating characteristic curve; BSS: Brier skill score; HL: p-value of Hosmer-Lemeshow test.

Hosmer-Lemeshow p-values for re-parametrized models suggest that calibration was still poor. However, the NB-ML models had positive BSS values in all prediction horizons, while all NB-expert models had negative BSS values. The calibration curves shown in **Figure 16** also show that calibration of NB-ML models is better (closer to the diagonal line) than that of the NB-expert models for all prediction horizons.

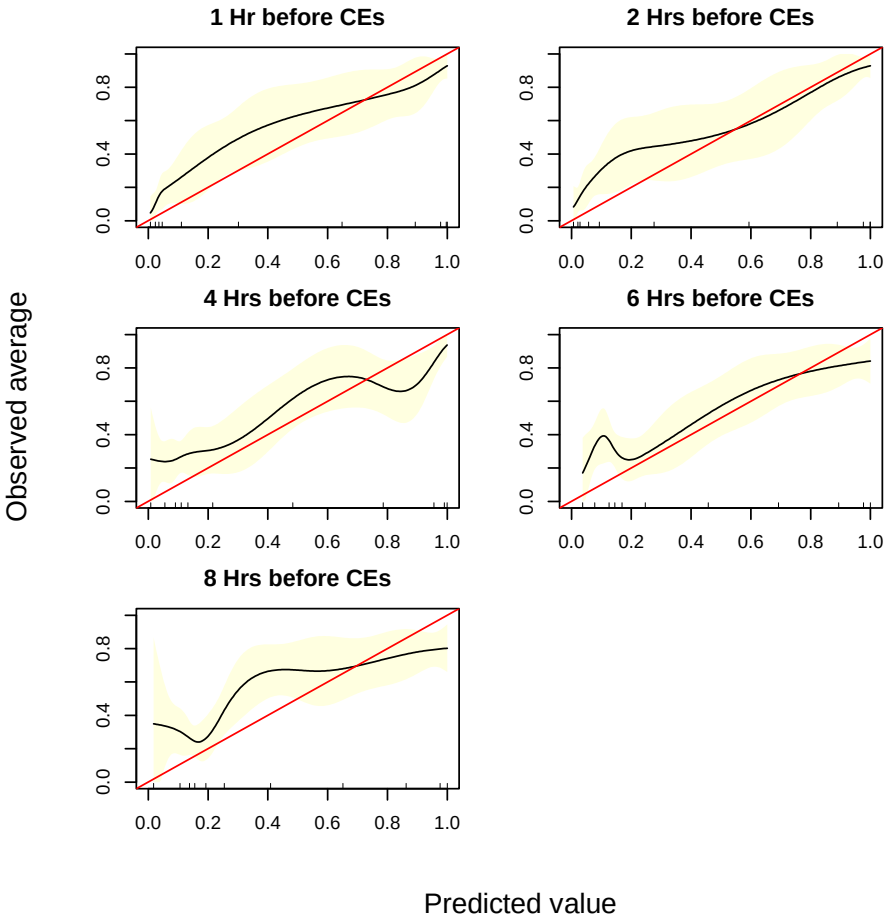


Figure 16. Calibration curves of the NB-ML-1-Full model across different prediction horizons

Receiver operating characteristic (ROC) analysis showed that the NB-ML-1-Full model had a sensitivity of 83.3% at the 81.1% specificity level one hour before CEs. Eight hours before CEs, it had a sensitivity of 56.8% at the 80.3% specificity level. Selecting a prediction threshold that resulted in a specificity level of 95%, the BN-ML-1 model had sensitivities of 48.5% and 25.8% one and eight hours before CEs, respectively. **Figure 17** shows specificity, sensitivity, and f1 values one hour before CEs for the best expert and re-parametrized models.

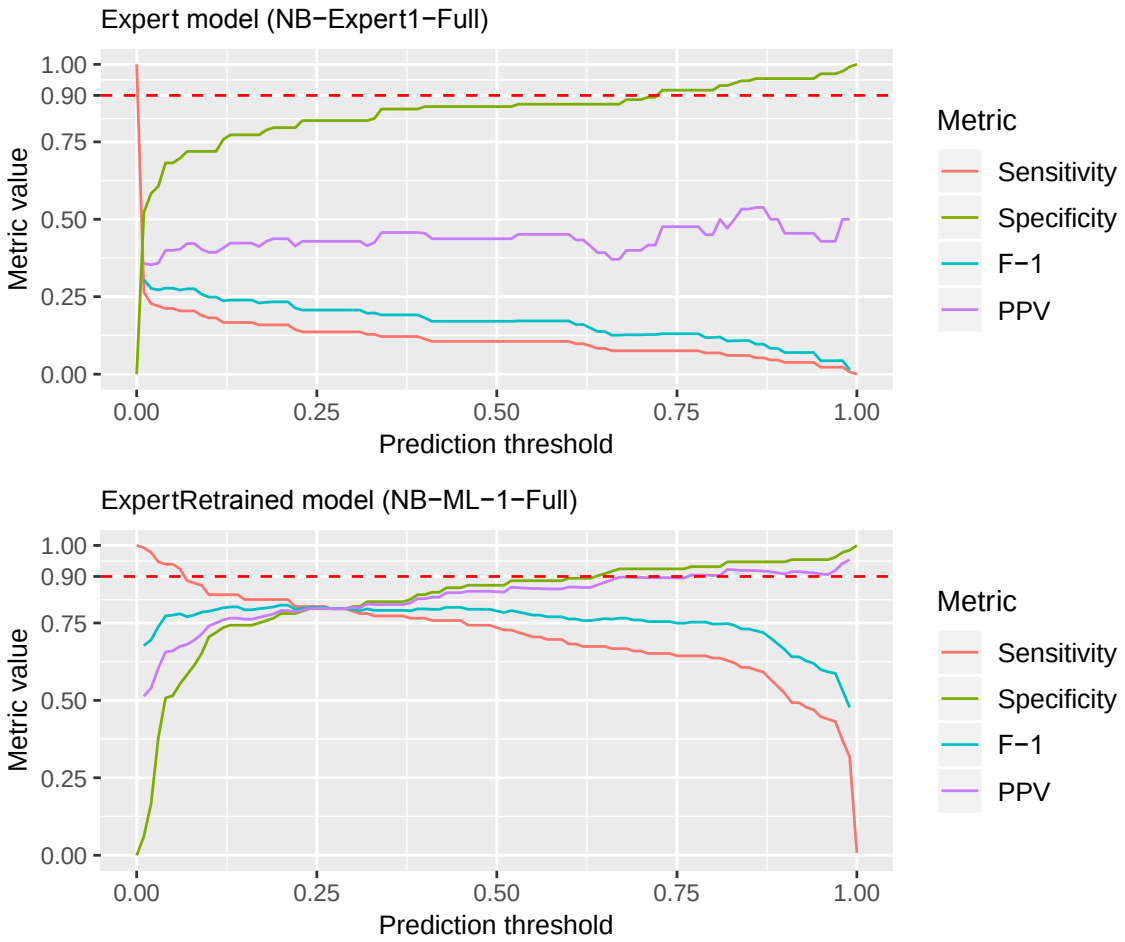


Figure 17. Comparison of performance metrics of the Expert and ExpertRetrained models at one hour before critical events. PPV: positive predictive value.

5.2.1.3 Performance of LastValsNumeric models

Models based on cross-sectional continuous values achieved high accuracy, with AUCs ranging from 0.77 (0.71-0.82) to 0.91 (0.88-0.95) at eight and one hours before CEs, respectively. Random forest models outperformed NB, DT, and SVM models at every prediction horizon, and also achieved higher BSS values. As seen in the calibration curves in **Figure 19**, RF models in this experiment tended to be under-confident for all prediction horizons, in contrast to the ExpertRetrained curves in **Figure 16**, which exhibited over-confidence in all prediction horizons.

Random forest models in this experiment had sensitivities of 89% and 56% at the 82% and 95% specificity levels at the one-hour prediction horizon, respectively. At eight hours before CEs, they had sensitivities of 61% and 36% at the 79% and 95% specificity levels, respectively. A detailed description of the behavior of the best models at one and eight hours before CEs for all prediction thresholds is shown in **Figure 18**.

Table 7. Prediction performance of LastValsNumeric models

Metric	Horizon	NB	SVM	DT	RF
AUC	-1	0.84 (0.79-0.89)	0.81 (0.76-0.87)	0.76 (0.7-0.82)	**0.91 (0.88-0.95)**
	-2	0.82 (0.77-0.87)	0.73 (0.67-0.79)	0.83 (0.78-0.87)	**0.89 (0.85-0.93)**
	-4	0.7 (0.64-0.76)	0.6 (0.53-0.67)	0.72 (0.65-0.78)	**0.78 (0.72-0.83)**
	-6	0.68 (0.61-0.74)	0.61 (0.54-0.67)	0.72 (0.66-0.78)	**0.78 (0.72-0.83)**
	-8	0.67 (0.6-0.73)	0.54 (0.48-0.61)	0.63 (0.57-0.7)	**0.77 (0.71-0.82)**
BSS	-1	0.17	0.28	0.20	0.46
	-2	0.14	0.16	0.30	0.42
	-4	-0.14	0.03	0.10	0.19
	-6	-0.41	0.05	0.10	0.19
	-8	-0.34	-0.01	-0.05	0.16
HL	-1	0.00	0.22	0.00	0.00
	-2	0.00	0.25	0.00	0.00
	-4	0.00	0.01	0.00	0.00
	-6	0.00	0.03	0.00	0.00
	-8	0.00	0.44	0.00	0.00
best f1	-1	0.81	0.76	0.77	0.86
	-2	0.78	0.71	0.78	0.82
	-4	0.69	0.67	0.70	0.74
	-6	0.69	0.68	0.70	0.72
	-8	0.67	0.67	0.67	0.72
f1 at 0.5	-1	0.73	0.75	0.76	0.86
	-2	0.71	0.65	0.76	0.82
	-4	0.58	0.61	0.70	0.72
	-6	0.48	0.62	0.68	0.72
	-8	0.50	0.55	0.58	0.69
best accuracy	-1	0.80	0.76	0.75	0.86
	-2	0.78	0.70	0.75	0.83
	-4	0.69	0.61	0.70	0.74
	-6	0.65	0.59	0.69	0.74
	-8	0.64	0.57	0.61	0.72

Values in parentheses show 95% confidence intervals computed with 2000 bootstrap replicates. Bold-face values show the best AUCs for each prediction horizon, i.e., the number of hours before critical events when predictions were generated. Models were trained and evaluated on a cohort of

95 patients and 132 critical events in 10-fold cross-validation. NB: Naïve Bayes; DT: Decision Tree; SVM: Support Vector Machine; RF: Random Forest.

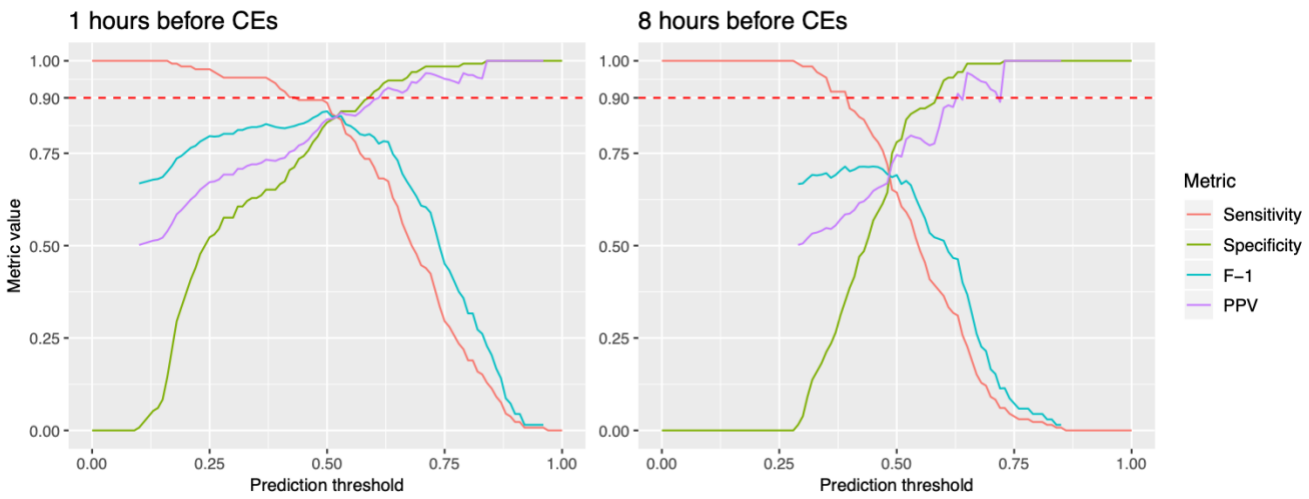


Figure 18. Performance metrics of LastValsNumeric models at one and eight hours before critical events. LastValsNumeric models presented in the graph are random forest classifiers trained with cross-sectional patient states without any temporal (longitudinal) trend features. PPV: positive predictive value.

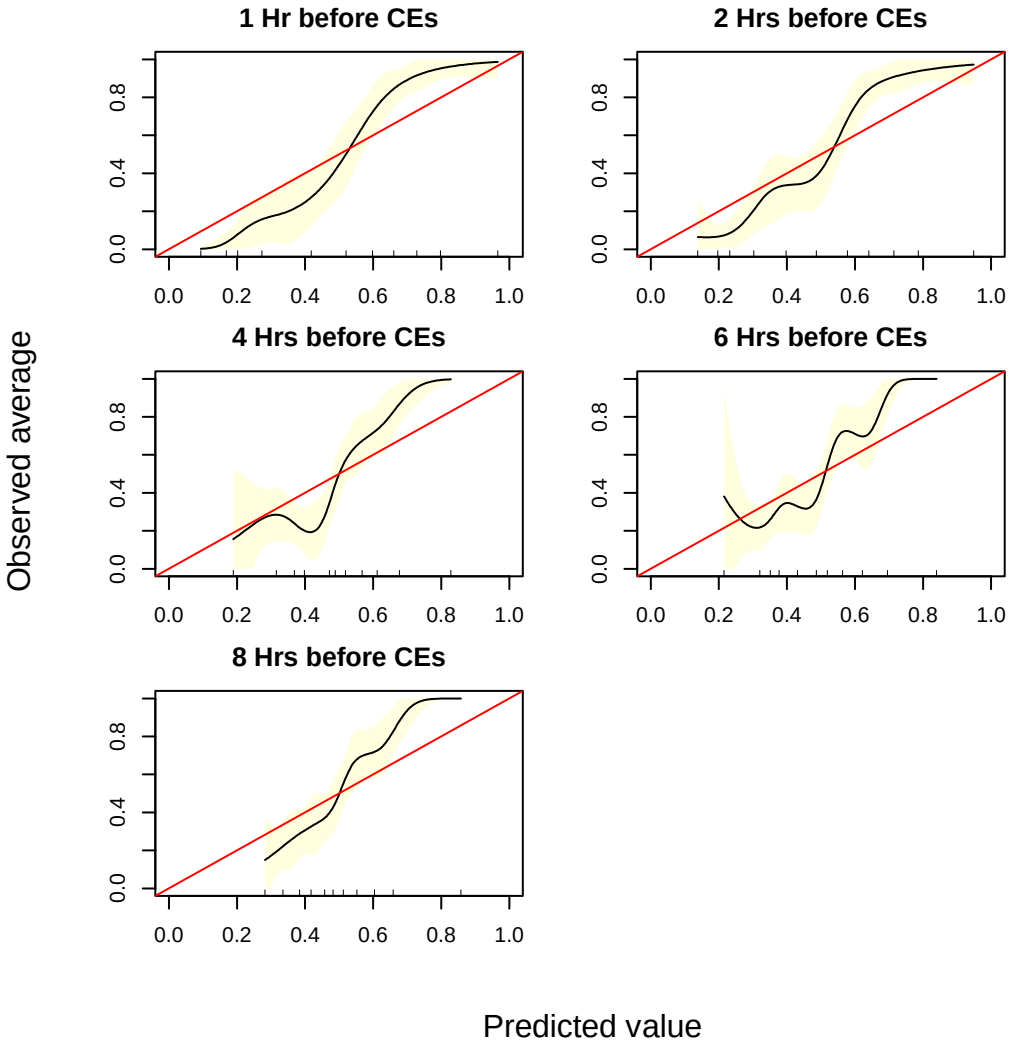


Figure 19. Calibration curves of random forest models trained with continuous-valued a-temporal patient states

5.2.1.4 Models including trend-summary features

Table 8 shows the performance of models trained with the TrendSummaries feature set, which included the last values of each variable as well as features that summarized temporal trends. In this table we present the results of the '*with_filter*' experiment, in which we only derived TrendSummaries from features for which two or more values were available for at least 25% of the training dataset. AUCs in this experiment ranged from 0.73 (0.67-0.79) to 0.87 (0.83-0.91) at

eight and one hours before CEs. Discrimination was lower in this experiment compared to the LastValsNumeric models with the exception of the model trained at four hours before CEs (AUC 0.80 vs 0.78). Again, RF classifiers achieved the highest discrimination, and had good calibration as shown by positive BSS values and HL values greater than 0.05 in four out of five prediction horizons. As seen in **Figure 20**, the best-performing RF models in this experiment tended to be slightly under-confident.

Table 8. Prediction performance of TrendSummary models

Metric	Horizon	NB	SVM	DT	RF
AUC	-1	0.82 (0.77-0.87)	0.8 (0.74-0.85)	0.8 (0.75-0.85)	**0.87 (0.83-0.91)**
	-2	0.85 (0.8-0.9)	0.79 (0.74-0.84)	0.73 (0.66-0.79)	**0.88 (0.84-0.92)**
	-4	0.69 (0.62-0.75)	0.63 (0.56-0.69)	0.72 (0.65-0.78)	**0.8 (0.74-0.85)**
	-6	0.69 (0.62-0.75)	0.64 (0.57-0.71)	0.64 (0.56-0.7)	**0.75 (0.69-0.81)**
	-8	0.64 (0.57-0.71)	0.63 (0.56-0.69)	0.6 (0.53-0.67)	**0.73 (0.67-0.79)**
BSS	-1	0.10	0.29	0.28	0.39
	-2	0.16	0.25	0.11	0.40
	-4	-0.21	0.06	0.04	0.25
	-6	-0.28	0.05	-0.05	0.16
	-8	-0.38	0.06	-0.13	0.14
HL	-1	0.00	0.01	0.00	0.13
	-2	0.00	0.00	0.00	0.02
	-4	0.00	0.02	0.00	0.08
	-6	0.00	0.00	0.00	0.06
	-8	0.00	0.18	0.00	0.14
best f1	-1	0.79	0.78	0.78	0.79
	-2	0.80	0.76	0.73	0.82
	-4	0.68	0.67	0.69	0.75
	-6	0.68	0.67	0.67	0.72
	-8	0.68	0.67	0.67	0.72
f1 at 50	-1	0.73	0.76	0.77	0.77
	-2	0.73	0.75	0.73	0.78
	-4	0.61	0.62	0.67	0.73
	-6	0.57	0.62	0.65	0.69
	-8	0.50	0.60	0.57	0.67
best accuracy	-1	0.79	0.78	0.77	0.80
	-2	0.79	0.77	0.71	0.81
	-4	0.67	0.64	0.68	0.73
	-6	0.66	0.63	0.67	0.72
	-8	0.64	0.61	0.61	0.69

Values in parentheses show 95% confidence intervals computed with 2000 bootstrap replicates. Bold-face values show the best AUCs for each prediction horizon, i.e., the number of hours before critical events when predictions were generated. Models were trained and evaluated on a cohort of

95 patients and 132 critical events in 10-fold cross-validation. NB: Naïve Bayes; DT: Decision Tree; SVM: Support Vector Machine; RF: Random Forest.

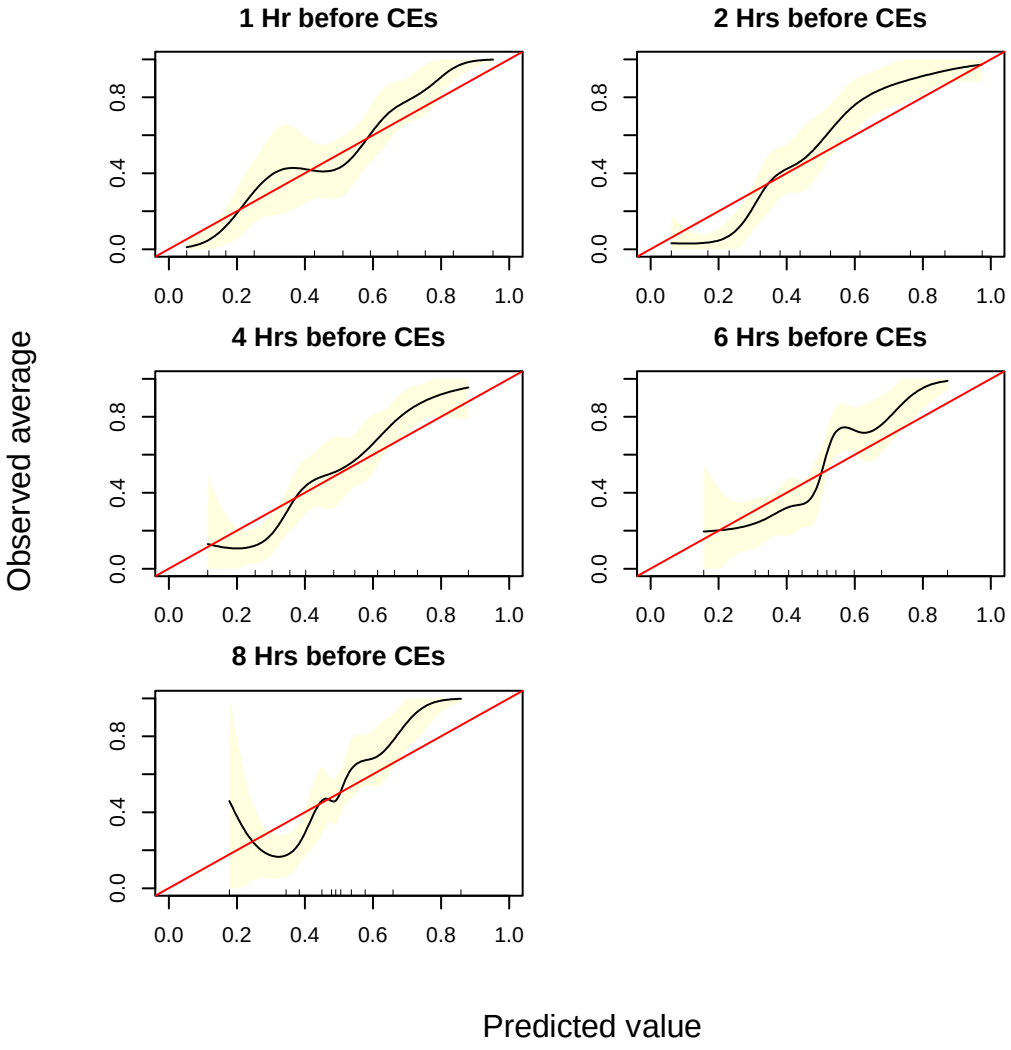


Figure 20. Calibration curves of random forest classifiers trained with static cross-sectional patient states augmented with trend-summary features

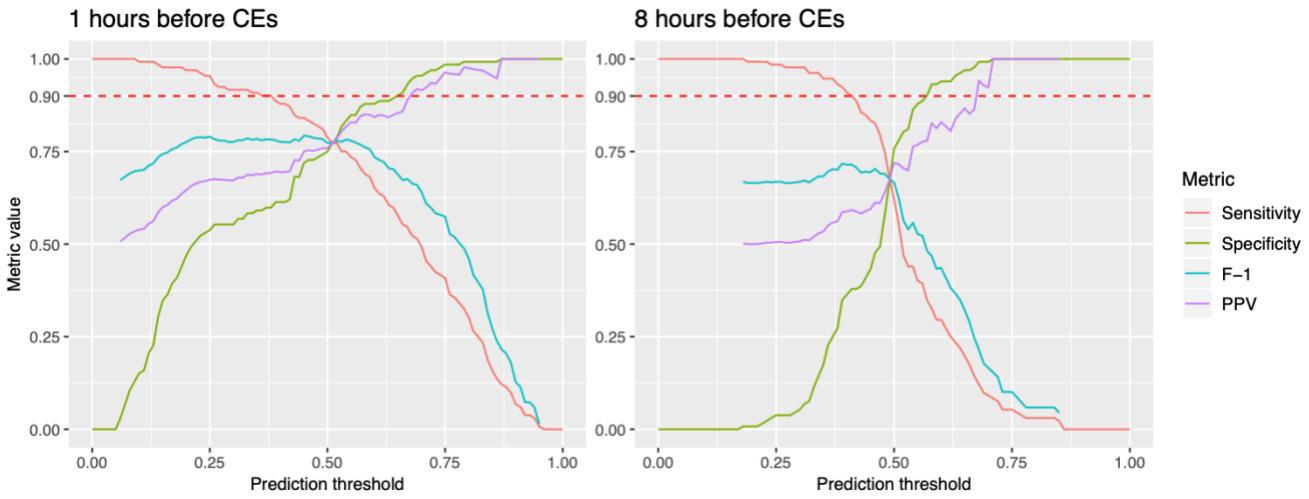


Figure 21. Performance metrics of TrendSummaries models at one and eight hours before critical events. TrendSummaries models presented in the graph are random forest classifiers trained with cross-sectional patient states and temporal features derived from longitudinal data (e.g., difference between apex and last value). PPV: positive predictive value.

5.2.1.5 Models based on frequent temporal patterns

Table 9 shows the performance of models trained with the FTPs feature set. In this table we present the results of the ExpertBins experiment, in which we only derived FTPs using expert-defined discretization bins to derive temporal abstractions. AUCs in this experiment ranged from **0.7 (0.64-**0.76) to 0.84 (0.79-0.89) at eight and one hours before CEs. Models derived from FTPs achieved the highest AUC across all experiments at four hours before CEs, but had lower AUCs than the LastValuesNumeric models for all other prediction horizons. RF classifiers achieved the highest discrimination with the exception of the two hours prediction horizon. Models based on FTPs exhibited positive BSS for all prediction horizons and HL values greater than 0.05 from four to eight hours before CEs. As seen in **Figure 22**, the best-performing RF models in this experiment tended did not show a consistent under or over-confidence behavior.

Table 9. Prediction performance of FTP models

Metric	Horizon	NB	SVM	DT	RF
AUC	-1	**0.84 (0.79-0.89)**	0.82 (0.77-0.87)	**0.84 (0.79-0.88)**	**0.84 (0.79-0.89)**
	-2	**0.82 (0.76-0.87)**	0.8 (0.75-0.86)	0.75 (0.69-0.81)	0.79 (0.74-0.85)
	-4	0.77 (0.71-0.82)	0.79 (0.73-0.85)	0.8 (0.74-0.85)	**0.82 (0.76-0.87)**
	-6	0.73 (0.67-0.79)	0.65 (0.58-0.72)	0.68 (0.61-0.73)	**0.74 (0.67-0.79)**
	-8	0.68 (0.61-0.75)	0.66 (0.6-0.73)	0.67 (0.6-0.73)	**0.7 (0.64-0.76)**
BSS	-1	0.17	0.32	0.39	0.37
	-2	0.12	0.30	0.17	0.30
	-4	-0.10	0.29	0.28	0.31
	-6	0.05	0.08	0.05	0.17
	-8	-0.06	0.07	-0.01	0.12
HL	-1	0.00	0.33	0.00	0.00
	-2	0.00	0.80	0.00	0.02
	-4	0.00	0.09	0.00	0.67
	-6	0.00	0.24	0.00	0.23
	-8	0.00	0.13	0.00	0.79
Best f1	-1	0.80	0.78	0.80	0.79
	-2	0.76	0.76	0.69	0.75
	-4	0.72	0.74	0.73	0.76
	-6	0.72	0.68	0.69	0.71
	-8	0.68	0.68	0.67	0.68
f1 at 50	-1	0.76	0.71	0.80	0.76
	-2	0.71	0.71	0.68	0.71
	-4	0.64	0.73	0.72	0.71
	-6	0.66	0.50	0.63	0.67
	-8	0.62	0.60	0.62	0.63
best accuracy	-1	0.80	0.77	0.81	0.79
	-2	0.78	0.76	0.75	0.76
	-4	0.71	0.77	0.73	0.74
	-6	0.69	0.62	0.63	0.73
	-8	0.67	0.64	0.64	0.66

Values in parentheses show 95% confidence intervals computed with 2000 bootstrap replicates. Bold-face values show the best AUCs for each prediction horizon, i.e., the number of hours before critical events when predictions were generated. Models were trained and evaluated on a cohort of 95 patients and 132 critical events in 10-fold cross-validation. Features used for model training

included frequent temporal patterns mined from the training dataset and based on expert-defined discretization abstractions. NB: Naïve Bayes; DT: Decision Tree; SVM: Support Vector Machine; RF: Random Forest.

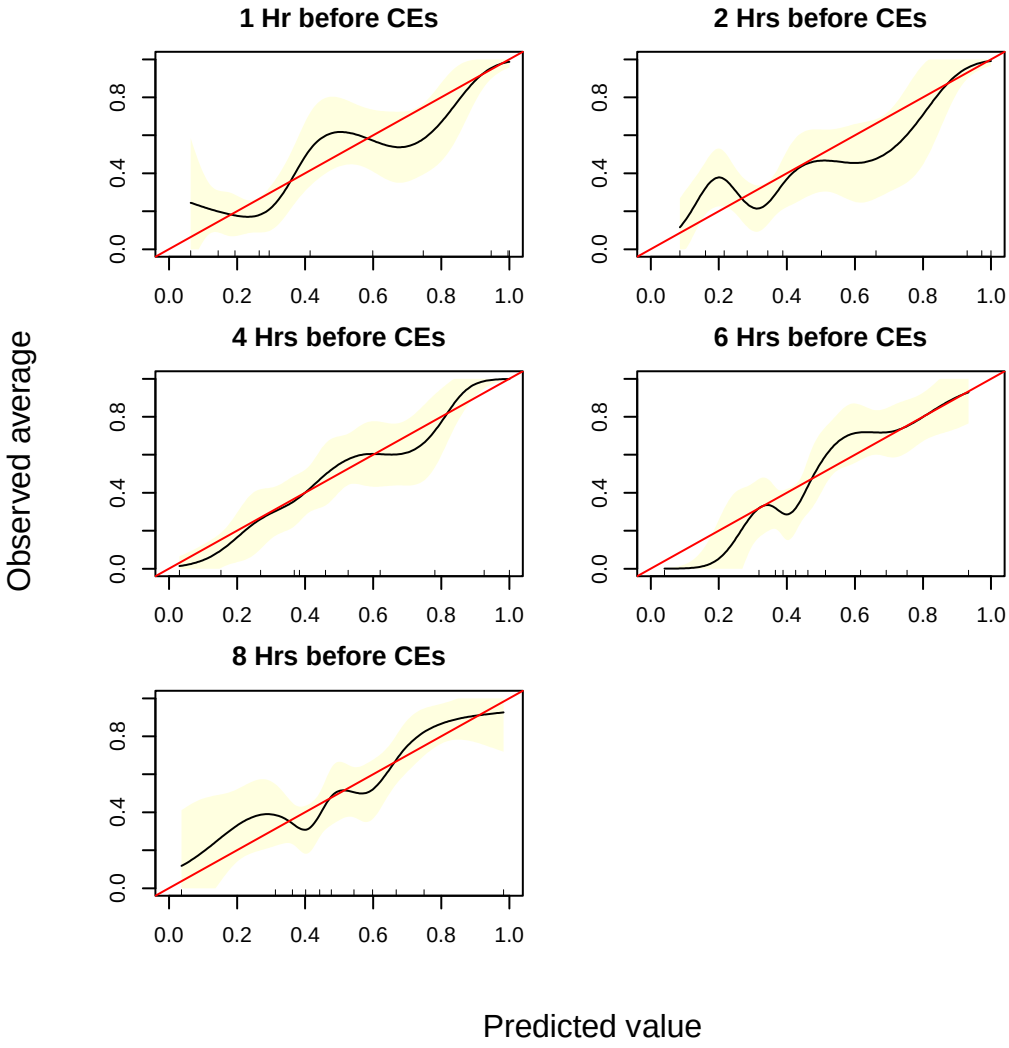


Figure 22. Calibration curves of random forest models derived from frequent temporal patterns with expert-binning temporal abstractions

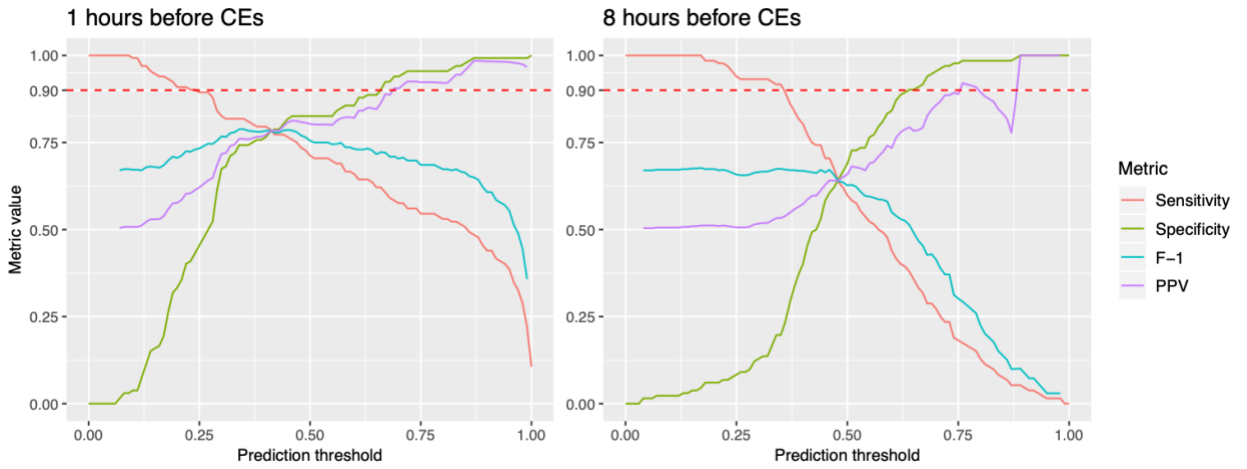


Figure 23. Performance metrics of FTP models at one and eight hours before critical events. Frequent temporal pattern (FTP) models presented in the graph are random forest classifiers trained with temporal features derived from longitudinal data. PPV: positive predictive value.

5.2.1.6 TimeSeries models

Table 10 shows the performance of models trained with time-series data. In this, experiments, LSTM classifiers achieved high AUCs ranging from 0.77 (0.71-0.82) to 0.9 (0.86-0.94) at eight and one hours before CEs. Models derived from time-series data achieved the highest AUC across all experiments at four and eight hours before CEs, but had lower AUCs than the LastValuesNumeric models at all other prediction horizons. LSTM Models had positive BSS for all prediction horizons but lack of fit as indicated by HL values. As seen in **Figure 24**, LSTM models in this experiment tended to be over-confident.

Table 10. Performance of long short-term memory models trained from time series data

Horizon	AUC	BSS	HL	Highest f1	F1 at 0.5	Accuracy
-1	0.9 (0.86-0.94)	0.49	0.00	0.84	0.83	0.84
-2	0.88 (0.83-0.91)	0.42	0.00	0.80	0.79	0.81
-4	0.82 (0.76-0.87)	0.28	0.00	0.78	0.75	0.77
-6	0.75 (0.69-0.8)	0.13	0.00	0.73	0.67	0.71
-8	0.77 (0.71-0.82)	0.18	0.00	0.73	0.71	0.74

Values in parentheses show 95% confidence intervals computed with 2000 bootstrap replicates. Bold-face values show the best AUCs for each prediction horizon, i.e., the number of hours before critical events when predictions were generated. Models were trained and evaluated on a cohort of 95 patients and 132 critical events in 10-fold cross-validation. NB: Naïve Bayes; DT: Decision Tree; SVM: Support Vector Machine; RF: Random Forest.

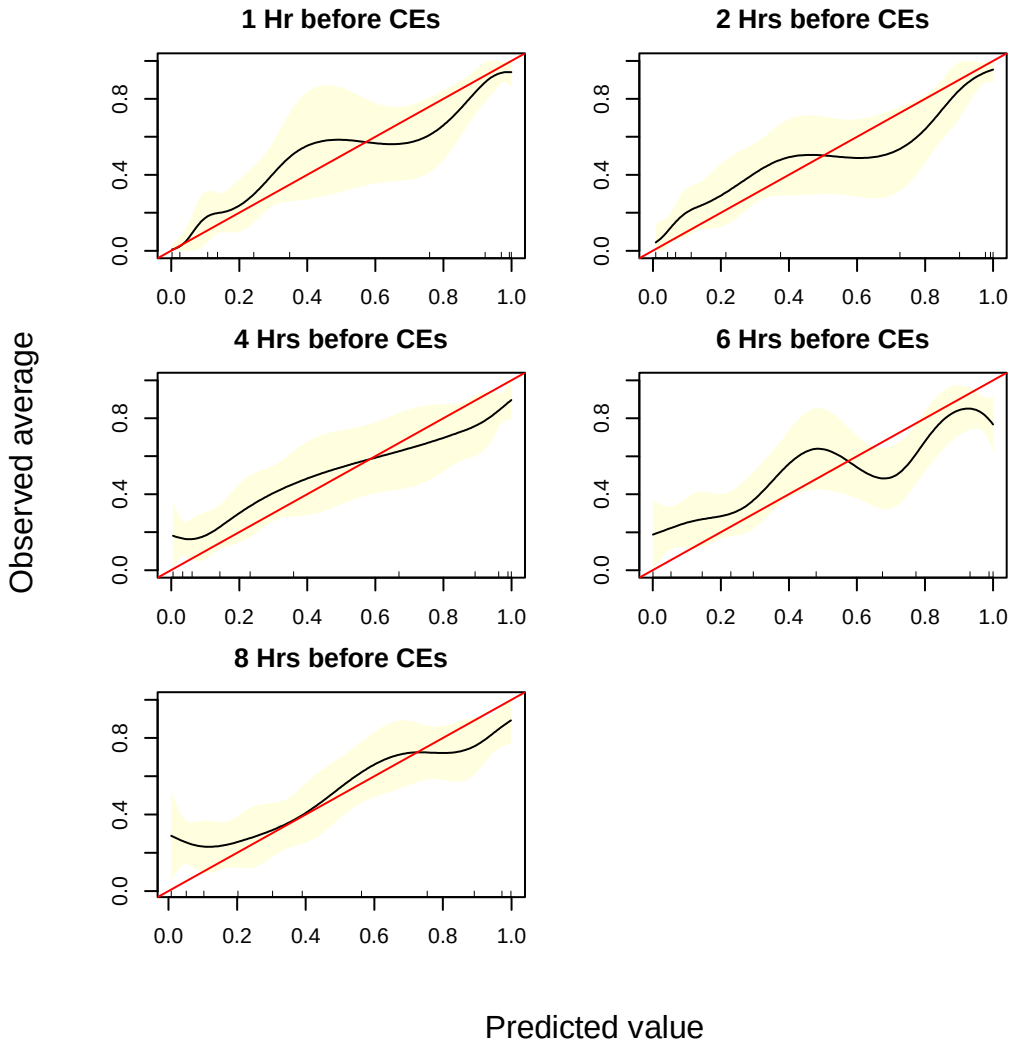


Figure 24. Calibration curves of LSTM models trained from time-series data

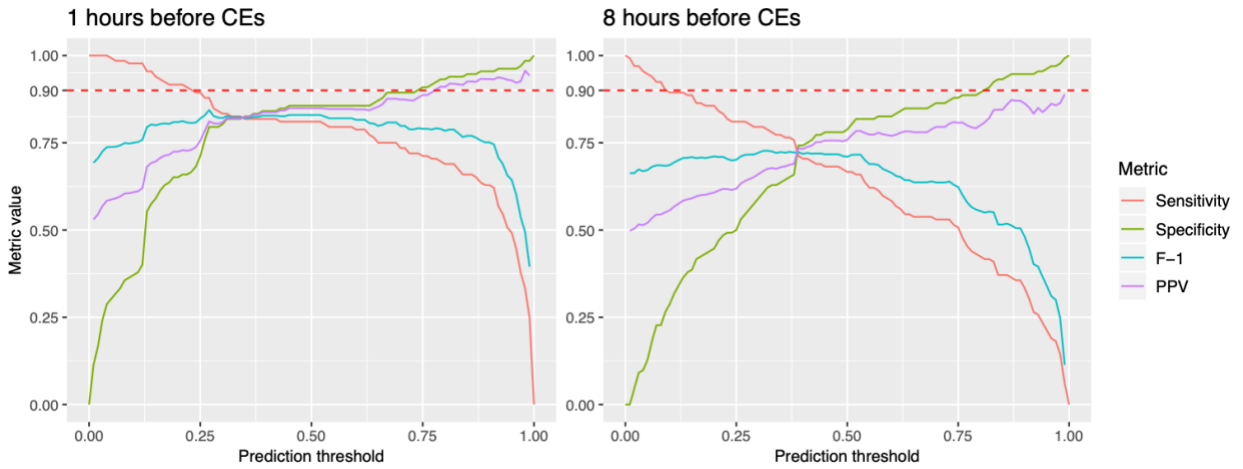


Figure 25. Performance metrics of TimeSeries models at different prediction thresholds. TimeSeries models presented in the graph are long short-term memory classifiers trained from time-series data. PPV: positive predictive value.

5.2.1.7 Performance summary

Among all feature sets, models trained from continuous-valued, a-temporal patient states (LastValsNumeric) achieved the best AUC in most prediction horizons, with the exception of four hours before CEs. At that horizon, the FTPs and TimeSeries models achieved the highest discrimination performance. In most experiments, random forest classifiers achieved the best performance. Details about individual classifier performance are available in **Table 7 - Table 10**.

Table 11. Predictive performance for each feature set and prediction horizon

Horizon	Expert	ExpertRetrained	LastValsNumeric	TrendSummaries	FTPs	TimeSeries
-1	0.67 (0.6-0.74)	0.88 (0.83-0.92)	**0.91 (0.88-0.95)**	0.87 (0.83-0.91)	0.84 (0.79-0.89)	0.9 (0.86-0.94)
-2	0.71 (0.64-0.77)	0.87 (0.82-0.91)	**0.89 (0.85-0.93)**	0.88 (0.84-0.92)	0.83 (0.78-0.87)	0.88 (0.83-0.92)
-4	0.58 (0.51-0.65)	0.79 (0.73-0.84)	0.78 (0.72-0.83)	0.81 (0.75-0.86)	**0.82 (0.76-0.87)**	**0.82 (0.77-0.87)**
-6	0.54 (0.47-0.61)	0.74 (0.68-0.8)	**0.78 (0.72-0.83)**	0.75 (0.69-0.81)	0.74 (0.67-0.79)	0.75 (0.69-0.81)
-8	0.49 (0.42-0.56)	0.74 (0.67-0.8)	**0.77 (0.71-0.82)**	0.74 (0.68-0.8)	0.7 (0.64-0.76)	**0.77 (0.7-0.82)**

Values in parentheses show 95% confidence intervals computed with 2000 bootstrap replicates. Bold-face values show the best AUCs for each prediction horizon, i.e., the number of hours before critical events when predictions were generated. Models were trained and evaluated on a cohort of 95 patients and 132 critical events in 10-fold cross-validation. For each feature set except for time-series data, we trained naïve Bayes, decision trees, random forests, and support-vector machine classifiers. This table shows the best performing model for each prediction horizon. The values in the TimeSeries column correspond to long short-term memory classifiers.

Table 12. Statistical comparison between best AUCs per feature set

Horizon	Expert	ExpertRetrained	LastValsNumeric	TrendSummaries	FTPs	TimeSeries
-1	5.1e-9*	0.012*	**N/A**	0.007*	0.002*	0.53
-2	1.1e-5*	0.16	**N/A**	0.58	0.002*	0.56
-4	5.4e-7*	0.59	0.49	0.65	**N/A**	**N/A**
-6	1.7e-6*	0.13	**N/A**	0.42	0.29	0.42
-8	2.5e-8*	0.3	**N/A**	0.31	0.08	**N/A**

Values are two-sided DeLong p-values of comparisons between the AUC of the best performing model at each horizon and all other (lower) AUCs. Cells with N/A values represent the highest-performing model in each prediction horizon. P-values were FDR corrected by prediction horizon. *P-values < 0.05; AUC: Area under the receiver operating characteristics curve; FDR: False-discovery rate.

5.2.2 Validation of CHP models on CHOP data

The expert models achieved higher AUCs in the CHOP dataset than in the CHP dataset at six and eight hours before CEs. Otherwise, all classifiers had lower AUCs at all prediction horizons when tested on the CHOP dataset. FTP models had the highest AUCs at one and six hours before CEs, and LastValsNumeric values had the highest AUCs at two, four, and eight hours before CEs, as shown in **Table 13**. Among all external validation experiments, the LastValsNumeric at two hours before CEs achieved the highest performance, with an AUC of 0.79 (0.74-0.84).

Table 13. AUCs of models trained with PGH data when tested on CHOP data

Horizon	Expert	ExpertRetrained	LastValsNumeric	TrendSummaries	FTPs	TimeSeries
-1	0.62 (0.56-0.69)	0.58 (0.52-0.64)	0.63 (0.57-0.68)	0.6 (0.54-0.66)	**0.66 (0.6-0.72)**	0.55 (0.49-0.61)
-2	0.61 (0.55-0.67)	0.56 (0.49-0.62)	**0.79 (0.74-0.84)**	0.6 (0.54-0.66)	0.61 (0.54-0.67)	0.55 (0.49-0.61)
-4	0.59 (0.53-0.65)	0.56 (0.5-0.63)	**0.66 (0.6-0.72)**	0.57 (0.51-0.63)	0.63 (0.57-0.7)	0.53 (0.47-0.59)
-6	0.57 (0.51-0.63)	0.62 (0.56-0.68)	0.61 (0.55-0.66)	0.58 (0.53-0.64)	**0.69 (0.63-0.75)**	0.56 (0.5-0.62)
-8	0.58 (0.51-0.64)	0.63 (0.57-0.69)	**0.64 (0.59-0.7)**	0.58 (0.51-0.64)	0.61 (0.55-0.67)	0.56 (0.49-0.62)

Values in parentheses show 95% confidence intervals computed with 2000 bootstrap replicates. Bold-face values show the best AUCs for each prediction horizon, i.e., the number of hours before critical events when predictions were generated. Models were trained on a cohort of 95 patients and 132 critical events, and validated on an external dataset comprised of 164 cases and 164 controls. For each feature set except for time-series data, we evaluated naïve Bayes, decision trees, random forests, and support-vector machine classifiers. For TimeSeries data, we trained and evaluated long short-term memory models. This table shows the best performing model for each horizon.

6.0 DISCUSSION

Hospital care of SV infants is complex because of their unique physiology, elevated severity of illness, and unpredictable clinical deterioration. In this dissertation, we developed the C-WIN models, which achieved early and accurate prediction of CEs in this population. We addressed this need by developing predictive models that use objective and routinely-collected data that can be retrieved automatically from an EHR system. Rather than following a completely data-driven approach, we incorporated expert knowledge into our models, which we believe is important to facilitate the adoption of predictive models into clinical workflows.

We tested the hypothesis that models that encode SV-domain-specific knowledge from cardiac intensivists would perform better in predicting CEs than currently-available models. We found that purely-expert-derived models achieved modest performance from one to four hours prior to CEs. However, they lost predictive power after six hours prior to CEs. A possible explanation for this behavior is that the CPTs provided by clinicians reflect the characteristics of patients close to the time of decompensation. Moreover, unlike models developed later on in this dissertation, the expert models predicted CEs at all prediction horizons with fixed CPTs. However, we found that applying machine-learning techniques increased expert-model performance significantly. Computing CPTs from retrieved training data and using the same variables and discretization ranges provided by experts in a naïve Bayes model (ExpertRetrained models) resulted in statistically-significantly-higher performance compared to using expert-defined CPTs.

Furthermore, using information gain reduced model complexity. With a conservative selection threshold, the NB-ML-lean models achieved similar performance than that of NB-ML-full models using fewer variables.

ExpertRetrained models achieved slightly lower AUCs than that to the state-of-the-art model for our prediction task[18] in the hour preceding critical events (0.88 vs. 0.91). However, clinical experts considered that patient deterioration may already be expected at that time. In contrast, our ExpertRetrained models achieved higher performance than that of the state-of-the-art model from two to eight hours before CEs.

Our second hypothesis was that using clinical data to extract temporal features and train static classifiers would result in significantly higher performance than that of expert models. We found that all models trained from trend-summary features and frequent temporal patterns achieved statistically-significantly-higher AUCs than those of the expert models for all prediction horizons. However, we found that classifiers that used continuous-valued variables without any longitudinal-changes or temporal information (LastValsNumeric) achieved the highest prediction among all experiments for all prediction horizons except for that at four hours before CEs.

Our third hypothesis was that dynamic models that leverage temporal patterns in time-series data would achieve state-of-the-art performance in early prediction of CEs in SV infants. Long short-term memory models trained from longitudinal data sequences achieved the highest AUCs at four and eight hours before CEs, and their AUCs were not statistically-significantly lower than the highest performing models for any other prediction horizon.

In our final experiment, we conducted an external validation of our models on dataset retrieved from a different institution and geographical location. We found that model performance was lower when applied at a different site, and that LastValsNumeric models were again the best

performing, achieving the highest AUC in three out of five prediction horizons. AUCs in the external validation ranged between 0.64 and 0.79, whereas when they were evaluated at the same site where data was retrieved for training, their AUCs were in the 0.77-0.91 range. This can be attributed to several factors. First, we could not retrieve urine output values in normalized units by weight and time since last measurement, and we did not include this variable in our evaluation. Second, some variables, such as SVO_2 values are not frequently used at the external validation site, as expressed by clinicians in said institution. Finally, peri-operative management and SV patient characteristics are variable across sites, making model generalization challenging.

We used naïve Bayes classifiers for developing the expert models and as a baseline for all other experiments. While naïve Bayesian networks are relatively simple compared to other modeling strategies (e.g., deep artificial neural networks, support vector machines), we believe that they are an appropriate baseline for this work. They have been used in biomedical research since the 1960's and are well-suited for clinical applications[24,25]. Their computation is efficient, they can operate with missing and categorical data explicitly, and can incorporate prior knowledge from clinicians or clinical data. Moreover, their simple structure and small number of parameters are beneficial in the absence of a large training dataset, which is especially challenging in specific populations such as SV infants.

In recent years there has been an increased interest in early warning systems and a variety of predictive models being used in clinical settings. These models are often validated in the adult population, and there have been efforts to adapt them to pediatric patients. However, the performance of generic early warning systems has suffered in certain populations where specific scores are better suited[10,63,64]. To the best of our knowledge, the model by Rusin et al. is the only model available for real-time prediction of CEs in SV infants[18]. Our best performing model

achieved the same performance as this model in the hour preceding CEs, and achieved higher performance from two to eight hours before CEs. We achieved this while using variables that are routinely available in most hospital EHR systems. Therefore, our models could be more feasibly implemented in institutions that do not have the technological or financial resources required to collect and analyze data such as ECG waveforms in real time. Finally, we retrieved data from a larger cohort to develop our models, including 95 patients and 132 CEs.

6.1 LIMITATIONS

The work presented in this dissertation had limitations. First, we could not retrospectively retrieve some variables that experts believed to be relevant, including electrocardiography, echocardiography, and X-ray imaging. However, those additional data may not be readily available in many hospitals and EHR systems. Nevertheless, it may be feasible to develop a more comprehensive model that leverages both routinely-collected and high-frequency ECG data in the future. Thus, the same model may be applied in hospitals with varying levels of technological resources. Second, we imputed missing values with data available within a six-hour period. While this may be a reasonable assumption, some variables may change rapidly and more elaborate imputation techniques may improve performance. Third, our test set included a limited number of ECMO and CPR events and we were not able to ascertain the performance of our models for these events separately. Fourth, we did not include variables related to clinical interventions. Although it is true that some interventions may signal imminent deterioration (e.g., order of ADAMTS13 activity test), for the first iteration of our models we decided to focus exclusively on physiological

variables. Finally, event times and variable values collected from EHRs were often entered manually as part of routine care and may have been subject to potential data-entry errors.

6.2 CONCLUSIONS

Early, real-time prediction of CEs may help clinicians reduce morbidity, mortality, length of stay, healthcare costs, and the suffering of patients and guardians. However, to fully realize the benefits of implementing such systems in clinical practice, they should be specific enough to minimize alert fatigue, and should not increase the burden of clinicians or divert resources from other aspects of care. From a clinical standpoint, our models may enable early interventions and avoidance of up to 56% of CEs with a specificity of 95% (based on performance one hour before CEs). This would allow clinicians to focus on patients truly at risk of CEs while minimizing alert fatigue. Furthermore, because our models utilize physiological variables that may be extracted automatically from an EHR system, an early warning system based on our models may operate autonomously and at low operational cost without disrupting clinicians' workflow.

We envision a potential implementation of our models as an alert-triggering system with a two-tiered set of responses. First, a model calibrated for increased sensitivity and earliest response may be monitored via virtual surveillance in the Tele-ICU setting, following a well-established model in adult patients[65]. Then, a second model calibrated for high specificity may be used to prompt a rapid response from bedside clinicians.

APPENDIX

Table 14. Variables identified by pediatric cardiologists as relevant for the prediction of critical events in SV infants

Clinical variable	In a group of 100 patients at risk of critical events, how many are expected to have the following values?				In a group of 100 patients at NO risk of critical events, how many are expected to have the following values?			
Heart rate [a, b, c, d]	<120	120-160	>160		<120	120-160	>160	
	40	20	40		15	70	15	
Respiratory rate [a, b, c, d]	<35	35-60	>60		<35	35-60	>60	
	40	20	40		15	70	15	
Oxygen saturation [a, b, c, d]	<70%	70-85%	>85%		<70%	70-85%	>85%	
	40	20	40		15	70	15	
Lactate [a, b, c, d]	<1.5	>1.5			<1.5	>1.5		
	20	80			80	20		
Mixed venous saturation [a, b, c, d]	<50%	50-60%	>60%		<50%	50-60%	>60%	
	85	5	10		10	80	10	
Partial pressure of oxygen [a, b, c, d]	<30	30-44	>44		<30	30-44	>44	
	40	5	55		10	80	10	
Systolic blood pressure [a, b, c, d]	<60	60-90	>90		<60	60-90	>90	
	40	20	40		10	80	10	
Diastolic blood pressure [a, b, c, d]	<30	30-60	>60		<30	30-60	>60	
	50	20	30		10	80	10	
Extubation time	<=4 days postop	>4 days postop			<=4 days postop	>4 days postop		
	20	80			80	20		
Sternal closure	<=3 days postop	>3 days postop			<=3 days postop	>3 days postop		
	20	80			80	20		
Residual lesions, common [b, c, d]	no	yes			no	yes		
	20	80			80	20		
Blood urea nitrogen [a, c, d]	<=40	>40			<=40	>40		
	60	40			40	60		
Brain natriuretic peptide [b, c, d]	<100	>100			<100	>100		
	20	80			80	20		
Chest X-ray effusion [b, c, d]	no	yes			no	yes		
	40	60			60	40		
Chest X-ray cardiomegaly [b, c, d]	No	mild	severe		No	mild	severe	
	10	20	70		70	20	10	

Table 14 continued

Chest X-ray congestive lungs [b, c, d]	no	Yes			no	yes		
	30	70			70	30		
Echocardiogram ventricular dilation [b, c, d]	No	Yes			No	Yes		
	40	60			60	40		
Echocardiogram diastolic flow reversal in descending aorta [b, c, d]	No	Yes			No	Yes		
	40	60			60	40		
Echocardiogram ventricular dysfunction [b, c, d]	No	Yes			No	Yes		
	40	60			60	40		
Echocardiogram atrioventricular valve regurgitation	No - mild	Mod - severe			No - mild	Mod - severe		
	40	60			60	40		
Echocardiogram systemic ventricle outflow obstruction [b, c, d]	No	Yes			No	Yes		
	40	60			60	40		
Echocardiogram coarctation of the aorta [b, c, d]	No	Yes			No	Yes		
	30	70			70	30		
Electrocardiogram ST elevation/depression > 1mm	No	Yes			No	Yes		
	20	80			80	20		
Electrocardiogram T-wave inversion	No	Yes			No	Yes		
	20	80			80	20		
Electrocardiogram decreased R-R variability	No	Yes			No	Yes		
	60	40			60	40		
Necrotizing enterocolitis	No	Yes			No	Yes		
	20	80			80	20		
Carbon dioxide [b, c, d]	<30	30-55	>55		< 30	30-55	>55	
	80	10	10		10	80	10	
Bicarbonate ion [a, b, c, d]	<20	20-32	>32 mEq/L		<20	20-32	>32 mEq/L	
	70	20	10		10	70	20	
Sodium	<128 mEq/L	128-146 mEq/L	>146 mEq/L		<128 mEq/L	128-146 mEq/L	>146 mEq/L	
	40	20	40		10	80	10	
Potassium	<2.5 mEq/L	2.5-5.5 mEq/L	>5.5 mEq/L		<2.5 mEq/L	2.5-5.5 mEq/L	>5.5 mEq/L	
	40	20	40		20	60	20	
Calcium [c, d]	<1.0 mmol/L	1-1.4 mmol/L	>1.4 mmol/L		<1.0 mmol/L	1.1-1.4 mmol/L	>1.4 mmol/L	
	40	20	40		10	80	10	

Table 14 continued

Glucose	<40 mg/dl	40-180 mg/dl	>180 mg/dl		<40 mg/dl	40-180 mg/dl	>180 mg/dl	
	40	20	40		10	80	10	
Hemoglobin [b, c, d]	<9 mg/dL	9-16 mg/dl	>16 mg dl		<9 mg/dL	9-16 mg/dl	>16 mg dl	
	40	20	40		10	80	10	
Hematocrit [b, c, d]	<30	30-50	>50		<30	30-50	>50	
	40	20	40		10	80	10	
International normalized ratio [c, d]	<1	1-1.4	>1.4		<1	1-1.4	>1.4	
	20	20	60		10	80	10	
Partial thromboplastin time [c, d]	<22	22-60	>60		<22	22-60	>60	
	20	20	60		10	80	10	
Platelets [c, d]	<100	100-400	>400		<100	100-400	>400	
	60	20	20		10	80	10	
Fibrinogen	<200 mg/dl	200-500 mg/dl	>500 mg/dl		<200 mg/dl	200-500 mg/dl	>500 mg/dl	
	60	10	30		10	80	10	
Creatinine [a, c, d]	<0.5	0.5-0.9	>0.9		<0.5	0.5-0.9	>0.9	
	10	10	80		80	10	10	
Alanine aminotransferase [c, d]	<50 U/L	50-200	>200		<50 U/L	50-200	>200	
	20	40	40		80	10	10	
Total bilirubin [c, d]	<=3	>3			<=3	>3		
	20	80			80	20		
Anti-Xa [c, d]	<0.3	0.3-0.7 IU/ml	>0.7		<0.3	0.3-0.7 IU/ml	>0.7	
	80	10	10		10	80	10	
ADAMTS-13 activity	<=57%	>57%			<=57%	>57%		
	80	20			20	80		
Urine output cc/Kg/Hr [a, b, c, d]	<1 cc/hr/kg	1-2 cc/hr/kg	>2 cc/kg/hr		<1 cc/hr/kg	1-2 cc/hr/kg	>2 cc/kg/hr	
	60	20	20		10	10	80	
Near-infrared spectroscopy [a, b, c, d]	<40	40-50	>50		<40	40-50	>50	
	70	20	10		5	15	80	
Preoperative intubation	yes	no			yes	no		
	80	20			20	80		
Central venous pressure [a, b, c, d]	<5	5-12	12-15	>15	<5	5-12	12-15	>15
	10	20	40	30	50	30	15	5
Base excess [a, b, c, d]	<-5	-3 to -5	-3 to 0	>0	<-5	-3 to -5	-3 to 0	>0
	70	10	10	10	3	7	20	70
Fraction of inspired oxygen [a, b, c, d]	<21%	21-30%	30-60%	>60%	<21%	21-30%	30-60%	>60%
	35	15	15	35	40	40	10	10

Table 14 continued

Pulmonary artery abnormalities	No	Yes			No	Yes		
	5	95			95	5		
Blalock-Taussig shunt abnormalities [b, c, d]	No	Yes			No	Yes		
	5	95			95	5		
Mixed venous saturation change from baseline	<25%	25-50%	>50%		<25%	25-50%	>50%	
	10	30	60		80	10	10	
Creatinine increase from baseline	<25%	25-50%	>50%		<25%	25-50%	>50%	
	10	30	60		90	8	2	

[a]Experts considered this variable as part of the minimal set of variables necessary to predict critical events; [b]Experts considered this variable relevant for prediction of emergent endotracheal intubation; [c]Experts considered this variable relevant for the prediction of extracorporeal-membrane oxygenation; [d]Expert considered this variable relevant for the prediction of cardiopulmonary resuscitation

Table 15. Mapping of clinical variables to available electronic health record concepts

Model variable	EHR clinical concept	Priority
ADAMTS 13 Activity	ADAMTS 13 Activity	1
ALT	ALT/SGPT	1
Anti-Xa	Anti Xa Unfract Heparin	1
	Anti -Xa Assay for Enoxaparin	1
Base excess[a]	Base Excess, Arterial	1
	Base Excess, Capillary	1
	Base Excess, Venous	1
	Base Excess, Oxygenator	1
	Base Deficit Oxygenator	1
	Base Deficit Capillary	1
	Base Deficit, Venous	1
	Base Deficit Arterial	1
	Base Deficit Venous	1
Bicarbonate anion[a]	HCO3a	1
	HCO3v	2
Blood urea nitrogen[a]	BUN	1
Brain natriuretic peptide	B-Type Natriuretic Peptide	1
Carbon dioxide	PaCO2	1
	PvCO2	2
Central venous pressure[a]	Central Venous Pressure	1
Creatinine[a]	Cr	1
Creatinine change from baseline[b]	N/A	N/A
Diastolic blood pressure[a]	Arterial Diastolic Pressure	1
	Diastolic BP	3
	Arterial Diastolic Pressure #2	2
Fibrinogen	Fibrinogen Level	1
Fraction of inspired oxygen[a]	FiO2	1
	FiO2 (oxygen %)	2
Glucose	Glucose Meter	1
	Glucose, Whole Blood	2
	Glucose	3
Heart rate[a]	Heart Rate	1
	Heart Rate - SPO2	2
	Pulse	3
Hematocrit	Hct, Whole Blood	1
	Hct Derived- Venous	2
	Hct Derived - Arterial	2
	Hct	3

Table 15 continued

Hemoglobin	Hgb	1
	Hgb, Venous	2
	Hgb, Arterial	2
International normalized ratio	INR	1
Ionized Calcium	Ionized Ca, Whole Blood	1
Lactate[a]	Lactate	1
	Lactate, Whole Blood	2
Mixed venous oxygen saturation[a]	O2 sat-Mixed Venous	1
	SvO2	1
Mixed venous oxygen saturation change from baseline[b]	N/A	N/A
Near infrared spectroscopy[a]	NIRS Cerebral Oxygenation-L	1
	NIRS Cerebral Oxygenation-R	1
	NIRS Tissue Oxygenation	2
	NIRS Cerebral Oxygenation	3
	NIRS Cerebral Oxygenation #2	3
Oxygen saturation[a]	SaO2	1
	SpO2 Bedside Monitor	1
Partial pressure of oxygen in arterial blood*	PaO2	1
Partial thromboplastin time	PTT	1
Platelets	Platelets	1
Potassium	K, Whole Blood	1
	K	2
Respiratory rate[a]	Respiratory Rate	1
Sodium	Na, Whole Blood	1
	Na	2
Systolic blood pressure[a]	Arterial Systolic Pressure	1
	Arterial Systolic Pressure #2	2
	Systolic BP	3
Total Bilirubin	Bili, Total	1
	Bilirubin	1
Urine output cc/Hg/Hr[a]	Urine Output 24 hour (weight based)	1
	Urine Output 8 hour (weight based)	1

Model variables were identified by cardiac intensivists. The *EHR clinical concept* column shows the hospital-specific codes that represent variables in the dataset used for model development and

evaluation. The *priority* values were used to select variable values when multiple EHR-event-code values were available simultaneously. [a] Variables included in the minimal subset that experts identified as essential for the prediction of critical events; [b] Variable derived from another variable in this table.

BIBLIOGRAPHY

1. Tabbutt S, Ghanayem N, Ravishankar C, et al. Risk factors for hospital morbidity and mortality after the Norwood procedure: A report from the Pediatric Heart Network Single Ventricle Reconstruction trial. *J Thorac Cardiovasc Surg*. 2012;144(4):882-895. doi:10.1016/j.jtcvs.2012.05.019

2. Meza JM, Hickey EJ, Blackstone EH, et al. The Optimal Timing of Stage 2 Palliation for Hypoplastic Left Heart Syndrome: An Analysis of the Pediatric Heart Network Single Ventricle Reconstruction Trial Public Data Set. *Circulation*. 2017;136(18):1737-1748. doi:10.1161/CIRCULATIONAHA.117.028481

3. Barron DJ, Kilby MD, Davies B, Wright JG, Jones TJ, Brawn WJ. Hypoplastic left heart syndrome. *Lancet*. 2009;374(9689):551-564. doi:10.1016/S0140-6736(09)60563-8

4. Pollack MM, Patel KM, Ruttimann UE. The Pediatric Risk of Mortality III--Acute Physiology Score (PRISM III-APS): a method of assessing physiologic instability for pediatric intensive care unit patients. *J Pediatr*. 1997;131(4):575-581. doi:10.1016/S0022-3476(97)70065-9

5. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients*. *Crit Care Med*. 2006;34(5):1297-1310. doi:10.1097/01.CCM.0000215112.84523.F0

6. Thiagarajan RR, Nathan M. Pediatric Index of Cardiac Surgical Intensive Care Mortality. *Pediatr Crit Care Med*. 2015;16(9):885-886. doi:10.1097/PCC.0000000000000510

7. Slater A, Shann F, Pearson G, Paediatric Index of Mortality (PIM) Study Group. PIM2: a revised version of the Paediatric Index of Mortality. *Intensive Care Med*. 2003;29(2):278-285. doi:10.1007/s00134-002-1601-2

8. Chapman SM, Wray J, Oulton K, Pagel C, Ray S, Peters MJ. "The Score Matters": wide variations in predictive performance of 18 paediatric track and trigger systems. *Arch Dis Child*. 2017;102(6):487-495. doi:10.1136/archdischild-2016-311088

9. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. *QJM*. 2001;94(10):521-526.

10. da Silva YS, Hamilton MF, Horvat C, et al. Evaluation of Electronic Medical Record Vital Sign Data Versus a Commercially Available Acuity Score in Predicting Need for Critical Intervention at a Tertiary Children's Hospital. *Pediatr Crit Care Med*. 2015;16(7):644-651. doi:10.1097/PCC.0000000000000444

11. Niles DE, Dewan M, Zebuhr C, et al. A pragmatic checklist to identify pediatric ICU patients at risk for cardiac arrest or code bell activation. *Resuscitation*. 2016;99:33-37. doi:10.1016/j.resuscitation.2015.11.017

12. Snoek KG, Capolupo I, Morini F, et al. Score for Neonatal Acute Physiology-II Predicts Outcome in Congenital Diaphragmatic Hernia Patients. *Pediatr Crit Care Med*. 2016;17(6):540-546. doi:10.1097/PCC.0000000000000738

13. Moss TJ, Lake DE, Calland JF, et al. Signatures of Subacute Potentially Catastrophic Illness in the ICU: Model Development and Validation. *Crit Care Med*. 2016;44(9):1639-1648. doi:10.1097/CCM.0000000000001738

14. Fenix JB, Gillespie CW, Levin A, Dean N. Comparison of Pediatric Early Warning Score to Physician Opinion for Deteriorating Patients. *Hosp Pediatr*. 2015;5(9):474-479. doi:10.1542/hpeds.2014-0199

15. Kennedy CE, Aoki N, Mariscalco M, Turley JP. Using Time Series Analysis to Predict Cardiac Arrest in a PICU. *Pediatr Crit Care Med*. 2015;16(9):e332-9. doi:10.1097/PCC.0000000000000560

16. Gupta P, Chakraborty A, Gossett JM, Rettiganti M. A prognostic tool to predict outcomes in children undergoing the Norwood operation. *J Thorac Cardiovasc Surg*. 2017;154(6):2030-2037.e2. doi:10.1016/j.jtcvs.2017.08.034

17. Vu EL, Rusin CG, Penny DJ, et al. A Novel Electrocardiogram Algorithm Utilizing ST-Segment Instability for Detection of Cardiopulmonary Arrest in Single Ventricle Physiology: A Retrospective Study. *Pediatr Crit Care Med*. 2017;18(1):44-53. doi:10.1097/PCC.0000000000000980

18. Rusin CG, Acosta SI, Shekerdemian LS, et al. Prediction of imminent, severe deterioration of children with parallel circulations using real-time processing of physiologic data. *J Thorac Cardiovasc Surg*. 2016;152(1):171-177. doi:10.1016/j.jtcvs.2016.03.083

19. Russell SJ, Norvig P. Learning from examples. In: *Artificial Intelligence: A Modern Approach*. 3rd ed. Pearson Education Limited; 2016.

20. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999;286(5439):531-537. http://www.ncbi.nlm.nih.gov/pubmed/10521349.

21. Ghesu F-C, Georgescu B, Zheng Y, et al. Multi-Scale Deep Reinforcement Learning for Real-Time 3D-Landmark Detection in CT Scans. *IEEE Trans Pattern Anal Mach Intell*. 2019;41(1):176-189. doi:10.1109/TPAMI.2017.2782687

22. Pearl J. Bayesian Networks. In: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. ; 1988.

23. Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks: The combination of knowledge and statistical data. *Mach Learn*. 1995;20(3):197-243. doi:10.1007/BF00994016

24. Warner HR, Toronto AF, Veasey LG, Stephenson R. A Mathematical Approach to Medical Diagnosis. *JAMA*. 1961;177(3):177. doi:10.1001/jama.1961.03040290005002

25. López Pineda A, Ye Y, Visweswaran S, Cooper GF, Wagner MM, Tsui F (Rich). Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. *J Biomed Inform*. 2015;58:60-69. doi:10.1016/j.jbi.2015.08.019

26. Rish I. An empirical study of the naive Bayes classifier. *IJCAI 2001 Work Empir methods Artif Intell*. 2001;3(22):41-46.

27. Kotsiantis SB. Decision trees: a recent overview. *Artif Intell Rev*. 2013;39(4):261-283. doi:10.1007/s10462-011-9272-4

28. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. *IEEE Trans Syst Man Cybern*. 1991;21(3):660-674. doi:10.1109/21.97458

29. Bishop CM. Sparse kernel machines. In: *Pattern Recognition and Machine Learning*. ; 2016.

30. Zimmerman RK, Balasubramani GK, Nowalk MP, et al. Classification and Regression Tree (CART) analysis to predict influenza in primary care patients. *BMC Infect Dis*. 2016;16(1):503. doi:10.1186/s12879-016-1839-x

31. Quinlan JR. Induction of Decision Trees. *Mach Learn*. 1986;1(1):81-106. doi:10.1023/A:1022643204877

32. Tin Kam Ho. Random decision forests. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol 1. IEEE Comput. Soc. Press; :278-282. doi:10.1109/ICDAR.1995.598994

33. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273-297. doi:10.1007/BF00994018

34. Hochreiter S, Schmidhuber J. Long Short-Term Memory. *Neural Comput*. 1997;9(8):1735-1780. doi:10.1162/neco.1997.9.8.1735

35. Williams RJ, Zipser D. Gradient-based learning algorithms for recurrent networks and their computational complexity. *Backpropagation Theory, Archit Appl*. 1995;1:433-486.

36. Hochreiter S. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *Int J Uncertainty, Fuzziness Knowledge-Based Syst*. 1998;06(02):107-116. doi:10.1142/S0218488598000094

37. Greff K, Srivastava RK, Koutnik J, Steunebrink BR, Schmidhuber J. LSTM: A Search Space Odyssey. *IEEE Trans Neural Networks Learn Syst*. 2017;28(10):2222-2232. doi:10.1109/TNNLS.2016.2582924

38. Batal I, Cooper GF, Fradkin D, Harrison J, Moerchen F, Hauskrecht M. An efficient pattern mining approach for event detection in multivariate temporal data. *Knowl Inf Syst*. 2016;46(1):115-150. doi:10.1007/s10115-015-0819-6

39. Moskovitch R, Shahar Y. Classification of multivariate time series via temporal abstraction and time intervals mining. *Knowl Inf Syst*. 2015;45(1):35-74. doi:10.1007/s10115-014-0784-5

40. Allen JF. Towards a general theory of action and time. *Artif Intell*. 1984;23(2):123-154. doi:10.1016/0004-3702(84)90008-0

41. Battiti R. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans Neural Networks*. 1994;5(4):537-550. doi:10.1109/72.298224

42. Valko M, Hauskrecht M. Feature importance analysis for patient management decisions. *Stud Health Technol Inform*. 2010;160(Pt 2):861-865. http://www.ncbi.nlm.nih.gov/pubmed/20841808. Accessed March 1, 2018.

43. Hauskrecht M, Batal I, Valko M, Visweswaran S, Cooper GF, Clermont G. Outlier detection for patient monitoring and alerting. *J Biomed Inform*. 2013;46(1):47-55. doi:10.1016/J.JBI.2012.08.004

44. Shahar Y. A framework for knowledge-based temporal abstraction. *Artif Intell*. 1997;90(1-2):79-133. doi:10.1016/S0004-3702(96)00025-2

45. Batal I, Valizadegan H, Cooper GF, Hauskrecht M. A Pattern Mining Approach for Classifying Multivariate Temporal Data. In: *2011 IEEE International Conference on Bioinformatics and Biomedicine*. IEEE; 2011:358-365. doi:10.1109/BIBM.2011.39

46. Batal I, Fradkin D, Harrison J, Moerchen F, Hauskrecht M. Mining recent temporal patterns for event detection in multivariate time series data. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '12*. New York, New York, USA: ACM Press; 2012:280. doi:10.1145/2339530.2339578

47. Batal I, Sacchi L, Bellazzi R, Hauskrecht M. Multivariate Time Series Classification with Temporal Abstractions. *Proc Twenty-Second Int FLAIRS Conf* . 2009. file:///C:/Users/victor/AppData/Local/Temp/48-2445-1-PB-1.pdf. Accessed March 1, 2018.

48. Moskovitch R, Shahar Y. Classification-driven temporal discretization of multivariate time series. *Data Min Knowl Discov*. 2015;29(4):871-913. doi:10.1007/s10618-014-0380-z

49. Moskovitch R, Walsh C, Hripcsak G, Tatonetti N. Prediction of Biomedical Events via Time Intervals Mining. 2014. https://www.researchgate.net/profile/Robert_Moskovitch/publication/267214122_Prediction_of_Biomedical_Events_via_Time_Intervals_Mining/links/54487d740cf22b3c14e31157/Prediction-of-Biomedical-Events-via-Time-Intervals-Mining.pdf. Accessed March 1, 2018.

50. Batal I, Cooper G, Hauskrecht M. A Bayesian Scoring Technique for Mining Predictive and Non-Spurious Rules. *Mach Learn Knowl Discov databases Eur Conf ECML PKDD . proceedings ECML PKDD*. 2012;7524:260-276. doi:10.1007/978-3-642-33486-3_17

51. Kononenko I, Šimec E, Robnik-Šikonja M. Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. *Appl Intell*. 1997;7(1):39-55. doi:10.1023/A:1008280620621

52. Archer KJ, Kimes R V. Empirical characterization of random forest variable importance measures. *Comput Stat Data Anal*. 2008;52(4):2249-2260. doi:10.1016/J.CSDA.2007.08.015

53. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B*. 1995;57(1):289-300. http://www.jstor.org/stable/2346101.

54. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845. http://www.ncbi.nlm.nih.gov/pubmed/3203132. Accessed April 4, 2018.

55. Jolliffe IT, Stephenson DB. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley & Sons; 2003.

56. Hosmer DW, Lemesbow S. Goodness of fit tests for the multiple logistic regression model. *Commun Stat - Theory Methods*. 1980;9(10):1043-1069. doi:10.1080/03610928008827941

57. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software. *ACM SIGKDD Explor Newsl*. 2009;11(1):10. doi:10.1145/1656274.1656278

58. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12(Oct):2825-2830. http://www.jmlr.org/papers/v12/pedregosa11a.html. Accessed December 23, 2018.

59. Keras: Deep Learning for Humans.

60. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12(1):77. doi:10.1186/1471-2105-12-77

61. Laboratory N-RA. Weather forecast verification utilities.

62. Lele, Subhash, Keim, Jonah, Solymos P. Resource Selection (Probability) Functions for Use-Availability Data.

63. Opio MO, Nansubuga G, Kellett J. Validation of the VitalPAC™ Early Warning Score (ViEWS) in acutely ill medical patients attending a resource-poor hospital in sub-Saharan Africa. *Resuscitation*. 2013;84(6):743-746. doi:10.1016/j.resuscitation.2013.02.007

64. Downey CL, Tahir W, Randell R, Brown JM, Jayne DG. Strengths and limitations of early warning scores: A systematic review and narrative synthesis. *Int J Nurs Stud*. 2017;76:106-119. doi:10.1016/j.ijnurstu.2017.09.003

65. Lilly CM, McLaughlin JM, Zhao H, et al. A multicenter study of ICU telemedicine reengineering of adult critical care. *Chest*. 2014;145(3):500-507. doi:10.1378/chest.13-1973