Distribution Agreement

Signature:

____ ______________

Azade Tabaie Date

Predicting Rare Clinical Events in Complex and Dynamic Environments

By

Azade Tabaie
Doctor of Philosophy

Computer Science and Informatics

Rishikesan Kamaleswaran, PhD
Advisor

Evan W. Orenstein, MD
Co-Advisor

Gari D. Clifford, DPhil
Committee Member

Matthew A. Reyna, PhD
Committee Member

Randi N. Smith, MD, MPH
Committee Member

Amy J. Zeidan, MD
Committee Member

Accepted:

Kimberly J. Arriola, PhD, MPH
Dean of the James T. Laney School of Graduate Studies

Date

Predicting Rare Clinical Events in Complex and Dynamic Environments

By

Azade Tabaie
B.Sc., Amirkabir University of Technology, Iran, 2012
M.Sc., Wayne State University, MI, 2015
M.Sc., Emory School of Medicine, GA, 2019

Advisor: Rishikesan Kamaleswaran, PhD

Co-Advisor: Evan W. Orenstein, MD

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2021

Abstract

Predicting Rare Clinical Events in Complex and Dynamic Environments
By Azade Tabaie

Traditional machine learning classification algorithms assume a balanced proportion of classes in the data. However, class-imbalanced data is a challenge for training predictive models in many fields such as the medical domain. Although patient adverse outcomes occur rarely, they are worthy of prediction to improve the quality of care that patients have received; therefore, monitoring systems are needed in the hospital setting to capture the adverse rare events and improve patient health outcomes.

To that end, machine learning and natural language processing (NLP) techniques were used along with clinical expert knowledge to address the issue of rare event classification in a complex environment such as a hospital setting. In particular, two different patient cohort with distinct characteristics and objectives were investigated.

First, strategies were proposed to predict a rare type of infection among hospitalized children with central venous lines (CVLs). This cohort of pediatric patients are at high risk of morbidity and mortality from hospital acquired infections. Many serious infections in hospitalized children are likely preventable through interventions that prevent the infection or identify them early to initiate antimicrobial therapy. Besides being considered as a rare clinical event, the definitions that have been proposed for bloodstream infection commonly have inadequate sensitivity for clinically important infections and may be difficult to generalize across electronic health records (EHR) platforms. To infer the onset of the infection from EHR and eliminate the need for extensive chart reviews, a surrogate definition for bloodstream infection was proposed and validated. Then, two study designs were tested to improve the prediction accuracy of the onset of the infection during hospitalization. Finally, a data fusion approach was undertaken to integrate structured and unstructured information from EHR to boost the prediction performance. Incremental but meaningful improvements in the predictions were observed after each step.

Second, an algorithm was proposed to monitor the visits to an emergency department (ED) to detect intimate partner violence (IPV). IPV is a pervasive social challenge with severe health and demographic consequences. People experiencing IPV may seek care in emergency settings. Despite the urgency of this critical public health issue, IPV continues to be profoundly underdiagnosed and is considered a persistent hidden epidemic. IPV is frequently undercoded, undetected without appropriate screening tools, and underreported, rendering it a rare encounter in EHRs. The early and appropriate detection of and response to such cases is critical in disrupting the cycle of abuse including IPV related morbidity and mortality. Our proposed algorithm benefits from NLP techniques and domain expert knowledge. It can identify victims of IPV with a high precision by analyzing the recorded provider notes and patient narratives.

We argue that all the techniques incorporated in this thesis are transferable to identify other rare clinical events with the ultimate goal of improving the level of care.

Predicting Rare Clinical Events in Complex and Dynamic Environments

By

Azade Tabaie
B.Sc., Amirkabir University of Technology, Iran, 2012
M.Sc., Wayne State University, MI, 2015
M.Sc., Emory School of Medicine, GA, 2019

Advisor: Rishikesan Kamaleswaran, PhD

Co-Advisor: Evan W. Orenstein, MD

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2021

Acknowledgments

Firstly, I would like to express my sincere gratitude to my advisors Dr. Kamaleswaran and Dr. Orenstein, for their continuous support of my Ph.D. study and related research, for their patience, motivation, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis.

Besides my advisors, I would like to thank the rest of my thesis committee: Dr. Clifford, Dr. Reyna, Dr. Smith, and Dr. Zeidan, for their insightful comments and encouragement, which helped widen my research from various perspectives. I also thank Dr. Nemati, Dr. Holder, Dr. Evans and Dr. Basu for their guidance and support in my doctoral training.

I thank all the wonderful people in the Biomedical Informatics department for making this journey pleasant. Special thanks go to Cathy, Julia, Robert, and Jim. I also thank the heart-warming kindness of friends Sahar, Fereshteh, Samaneh, Mohsen, Fatemeh, Nasim, Parisa and the rest of my friends inside and outside of the department.

My gratitude goes out to my parents, without whom I would not have reached my goals. I am thankful for their prayers and supports throughout my life.

This dissertation is dedicated to my husband, Sina. Without your unwavering love and support, this work would not exist. You have helped me realize when I am strong enough to push on, and you have shown me the wisdom in knowing when to ask for help.

I am fortunate, and I am forever grateful.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis contribute to approaches for dealing with extreme class-imbalanced classification problems in complex clinical settings. An imbalanced classification problem is an example of a classification problem where the distribution of examples across the known classes is biased or skewed. The distribution can vary from a slight bias to an extreme imbalance where there is one example in the minority class for hundreds, or thousands of examples in the majority class. The extreme class imbalanced classification problem is also called *rare event* classification problem.

With the rapid advancements of technology, implementing artificial intelligent-based surveillance systems in healthcare facilities is becoming more popular [71, 100, 96, 67]. However, some of the patients' adverse outcomes that are worthy of prediction occur rarely and affect the classification performance of the machine learning model [52, 15, 77].

In this work, we illustrate two examples of rare events in hospitals:

1) Serious bloodstream infection among hospitalized children on central venous lines. Central venous line (CVL) is an intravascular catheter that terminates at or close to the heart, or in one of the great vessels that is used for infusion, withdrawal of blood, or hemodynamic monitoring. Patients with central line are at higher risk of experiencing hospital acquired infections such as serious bloodstream infection, mainly due to having an open wound. This type of infection happens rarely but contributes to higher morbidity, mortality and hospital length of stay. Every year, almost 80k new cases are identified and the patients who developed serious bloodstream infection during hospitalization have 12-25% increased risk of mortality. Predicting the infection ahead of time helps clinicians better identify patients at higher risks for serious infection and achieve balance between early intervention and antimicrobial overuse.

2) Emergency department visits associated with injuries from intimate partner violence. According to CDC, intimate partner violence (IPV) is defined as sexual, physical, psychological, or economic violence that occurs between current or former intimate partners. Individuals who experience IPV experience both short- and long-term adverse health outcomes such as chronic pain, substance use, and mental health disorders. In the United States one in four women and one in nine men have experienced a severe form of IPV at some point in their lifetime; therefore, IPV is not a rare event. But only a small proportion of victims of IPV seek care for this type of injuries which make the IPV a rare event in terms of the data that we have. IPV continues to be underreported and underdiagnosed which makes it critical to identify the IPV-related medical visits in order to disrupt the cycle of abuse and decrease the associated morbidity and mortality.

1.1 Motivation

Traditional machine learning classification algorithms assume a balanced proportion of different classes in the data. However, class-imbalanced data is a challenge when training predictive models in many fields, such as the medical domain. Studies have been conducted in this area, and different algorithms have been proposed [56, 38, 34].

The proposed solutions can be categorized into three subgroups; data level, algorithm level, and cost-sensitive learning approaches.

The most common data level solutions are oversampling the minority class and undersampling the majority class [50, 45, 55, 60]. While these methods are simple to use and effective in some cases, they can lead to overfitting to the training dataset (oversampling the minority class) or insufficiently learning the majority class's patterns (undersampling the majority class). As a result, the Synthetic Minority Oversampling Technique (SMOTE) and its family of algorithms have been proposed [20, 37, 72], in which new synthetic data points from the minority class were generated by the use of the available ones.

Ensemble learning is an approach at the algorithm level, which mainly consists of bagging and boosting methods. In ensemble learning, we benefit from fitting many models instead of one. In bagging, the majority vote of the models is selected as the final class. In boosting, incorrectly predicted samples are up-weighted in subsequent training rounds. Extreme Gradient Boosting (XGBoost) [22] and Random Forest [11] are two of the famous examples of ensemble models.

Most machine learning algorithms assume that all data samples are independent and identically distributed and have the same weight, regardless of coming from a minority or majority class. Equation 1.1 presents how loss value is calculated for a logistic regression model under the aforementioned assumption. In this equation, L demonstrates loss, y is the true class label which takes the value of zero or one, and p is the predicted probability of positive class. This assumption is not valid in modeling class-imbalanced data [99]. Many real-world applications, such as spam and fraud detection and identifying infection among patients, indicate that misclassifying a minority class sample is more expensive than incorrectly classifying a data sample from the majority class.

$$L = -y \cdot \log(p) - (1-y) \cdot \log(1-p) = \begin{cases} -\log(p), & \text{if y=1} \\ -\log(1-p), & \text{otherwise} \end{cases} \tag{1.1}$$

Equation 1.2 presents a modified version of loss function for a binary classification task in presence of class-imbalanced data. This approach is called cost-sensitive learning. Cost-sensitive learning is a field of machine learning that requires defining and using costs in the training process. In cost-sensitive learning, a misclassification cost, such as α in Equation 1.2, is assigned to each class so that the cost is higher for incorrectly classifying the minority class samples [39].

$$L = \begin{cases} -\alpha \log(p), & \text{if y=1} \\ -(1-\alpha)\log(1-p), & \text{otherwise} \end{cases} \tag{1.2}$$

Complex and dynamic data have been commonly incorporated in training machine learning models in the medical domain [64, 24]. In such a case, models should capture the temporal information in the data effectively [40, 89] so that the model can benefit from it, learn patterns efficiently, and tackle the class-imbalanced classification problem. The data level solutions would be challenging to use in temporal data modeling since the data is in the sequence format. On the other hand, ensemble learning techniques could be inefficient in terms of computational costs. Therefore, there should be different approaches that apply to temporal data modeling.

This thesis contributes to addressing these challenges and developing pipeline and model structures that can be incorporated in the case of rare event classification in various healthcare fields.

1.2 Aim of this thesis

This thesis aims to provide generalizable methods for solving the problem of rare event classification in a complex environment such as a hospital setting. To this end, we selected two different domains, infection prediction in the pediatric population and identifying intimate partner violence cases among the adult population, and proposed frameworks transferable to other fields in patient outcome predictions. To achieve our final aim, the following novel research was performed:

- We define a surrogate definition for a rare type of hospital-acquired infection, central line-associated bloodstream infection (CLABSI), which can be inferred from routinely recorded EHRs and eliminate the need for extensive chart reviews. Then, we propose a method based on machine learning techniques to identify rare patient adverse outcomes in a complex environment. We evaluate the method's effectiveness on the serious infection prediction task using six years of data associated with hospitalized children with central venous lines (CVLs).

- A method based on deep neural networks, attention mechanism and a modified loss function to incorporate the temporal information hidden in EHR to predict a rare adverse outcome in a timely manner. We evaluate the effectiveness of this method on the serious bloodstream infection prediction using six years of data associated with hospitalized children with CVLs.

- A method based on deep neural networks and natural language processing techniques to investigate the effect of integrating two data modalities in predicting rare events in a timely manner. We evaluate the effectiveness of this method on the serious bloodstream infection prediction using six years of data, including structured EHR and recorded clinical notes associated with hospitalized children with CVLs.

- A method based on NLP techniques to identify specific rare events from recorded data in free-text format in a dynamic environment. We evaluate the effectiveness of this method in detecting emergency department visits associated with intimate partner violence.

1.3 Thesis outline

The thesis comprises five chapters besides the introduction, all of which (except for the conclusion) have been published or are under review in key journals in the field (see section 1.4).

Chapter 2 presents our proposed surrogate definition for CLABSI, which can be inferred from routinely recorded EHR data. Then, the chapter describes a framework using retrospective data and machine learning capable of predicting a rare clinical event such as a serious bloodstream infection. The chapter concludes by presenting the predictive model's performance on a large-scale dataset of hospitalized pediatric patients.

Chapter 3 proposes a window-wise study design that can be employed as a monitoring system in a healthcare facility for timely prediction of the onset of serious infection among hospitalized children with CVLs. Then, the chapter describes the proposed method using deep learning, attention mechanisms, and a modified loss function to diminish the challenges of an extreme class-imbalanced classification problem.

Chapter 4 extends the work presented in Chapter 3 by adding a new data modality to the input features of the predictive model. This chapter investigates the effect of coupling structured and unstructured EHR data in predicting rare adverse outcomes and proposes a method based on NLP techniques to dynamically capture the embedded information in the provider notes.

Chapter 5 introduces the challenges in detecting the cases of intimate partner violence (IPV) through recorded clinical data and proposes an NLP-based algorithm to utilize providers' notes and patient narratives to identify IPV cases among the visits to the emergency department of a level one trauma center. The chapter concludes by presenting the results of validating the proposed labeling algorithm through manual chart reviews.

Finally, Chapter 6 presents a summary of contributions, limitations, and possible future work.

1.4 List of publications

Work in this thesis has been published in the following journals:

- **A. Tabaie**, E. W. Orenstein, S. Nemati, R. K. Basu, S. Kandaswamy, G. D. Clifford, R. Kamaleswaran, "Predicting Presumed Serious Infection among Hospitalized Children on Central Venous Lines with Machine Learning", Computers in Biology and Medicine, 2021 May 1;132:104289.
  (This publication appears in its entirety in Chapter 2).

- **A. Tabaie**, E. W. Orenstein, S. Nemati, R. K. Basu, G. D. Clifford, R. Kamaleswaran, "Deep Learning Model to Predict Serious Infection among Children with Central Venous Lines", Frontiers in Pediatrics. 2021 Sep 15;9:726870.
  (This publication appears in its entirety in Chapter 3).

- **A. Tabaie**, E. W. Orenstein, S. Kandaswamy, R. Kamaleswaran, "A Machine Learning Pipeline for Integrating Structured and Unstructured Data for Timely Prediction of Bloodstream Infection among Children with Central Venous Lines", Pediatric Research, Under Review.
  (This publication appears in its entirety in Chapter 4).

- **A. Tabaie**, Amy J. Zeidan, Dabney P. Evans, Randi N. Smith, Rishikesan Ka-

maleswaran, “A Novel Technique for Developing a Natural Language Processing Algorithm to Identify Intimate Partner Violence in a Hospital Setting“, BMJ Quality and Safety, Under Review.
(This publication appears in its entirety in Chapter 5).

Chapter 2

Predicting Presumed Serious Infection among Hospitalized Children on Central Venous Lines with Machine Learning

2.1 Abstract

Background: Presumed serious infection (PSI) is defined as a blood culture drawn and new antibiotic course of at least 4 days among pediatric patients with Central Venous Lines (CVLs). Early PSI prediction and use of medical interventions can prevent adverse outcomes and improve the quality of care.

Methods: Clinical features including demographics, laboratory results, vital signs, characteristics of the CVLs and medications used were extracted retrospectively from electronic medical records. Data were aggregated across all hospitals within a single pediatric health system and used to train machine learning models (XGBoost and ElasticNet) to predict the occurrence of PSI 8 h prior to clinical suspicion. Prediction for PSI was benchmarked against PRISM-III.

Results: Our model achieved an area under the receiver operating characteristic curve of 0.84 (95% CI = [0.82, 0.85]), sensitivity of 0.73 [0.69, 0.74], and positive predictive value (PPV) of 0.36 [0.34, 0.36]. The PRISM-III conversely achieved a lower sensitivity of 0.19 [0.16, 0.22] and PPV of 0.30 [0.26, 0.34] at a cut-off of 10. The features with the most impact on the PSI prediction were maximum diastolic blood pressure prior to PSI prediction (mean SHAP = 3.4), height (mean SHAP = 3.2), and maximum temperature prior to PSI prediction (mean SHAP = 2.6). Conclusion: A machine learning model using common features in the electronic medical records can predict the onset of serious infections in children with central venous lines at least 8 h prior to when a clinical team drew a blood culture.

2.2 Introduction

Children with central venous lines (CVLs) are at high risk of morbidity and mortality from hospital acquired infections (HAI), including central-line associated bloodstream infections (CLABSIs) and sepsis. While specific definitions for these entities exist in pediatrics, they often have inadequate sensitivity for clinically important infections and may be difficult to generalize across electronic medical record (EMR) platforms [51, 5]. The presumed serious infection (PSI) case definition was developed initially by adult sepsis epidemiologists to allow for retrospective surveillance of infection and organ dysfunction that could be applied across diverse EMRs [83, 85, 84]. It is defined as at least one blood culture draw followed by at least four consecutive days (or fewer if the patient dies or is transferred out) of antimicrobial agents that were not administered in the week prior to the blood culture draw. The definitions for PSI, as well as organ dysfunction, were adapted for pediatrics by Hsu et al. [42] and have been validated [107]. Successful prediction of PSI, or sepsis in general, among hospitalized children or the adult population could lead to decreased costs while improving the quality of care [69, 82, 71].

Many serious infections and adverse outcomes in hospitalized children are likely preventable through interventions that prevent PSIs or identify them early to initiate antimicrobial therapy. On the other hand, excessive use of antimicrobials can lead to adverse events and worsening antimicrobial resistance. In this setting, predictive models to identify patients at the highest risk for serious infections can help clinicians better achieve the balance between early intervention and antimicrobial overuse.

Machine learning models have been used to address clinical problems [52, 26, 9]. Most of these models have employed biomarkers or clinical risk predictions to predict the onset of events [92, 44]. The deployment of machine learning models and early recognition of the adverse events decreased the mortality and morbidity among patients [85, 79]. However, there are still open challenges about developing machine learning models for low prevalence outcomes [88]. Within the pediatric domain, a recent study reported the incidence of CLABSI in pediatric cardiac ICUs to be 0.32% among children aged between 1 and 18 years [98], representing tremendous challenge for classical machine learning techniques which assume a balanced distribution of cases and controls. These limitations suggest that there is a need to develop robust machine learning algorithms that can adequately predict low prevalence conditions, particularly for pediatric cohorts.

Class-imbalanced data classification is a challenge of training predictive models in many fields such as the medical domain. Studies have been done in this area, and different algorithms have been proposed [56, 38, 34]. The proposed solutions mainly include sampling techniques [50, 50, 55, 60], such as oversampling the minority class or undersampling the majority class, and employing ensemble learning approaches [11].

This paper introduces our proposed framework named pBoost which was developed based on our adapted study design and contributes both clinically and technically to the literature. From the clinical perspective, we incorporated a novel case definition to identify the pediatric patient with CVL who were at a higher risk of experiencing a serious infection episode during their hospitalization time. From the technical point of view, we proposed a framework that can perform a series of tasks such as data engineering, data preprocessing and feature transformation to model hyperparameter tuning, providing multiple performance metrics and the most influential features. We further highlight clinical features which contributed the PSI prediction using data from the EMR. This study was conducted according to Emory University protocol number 19-012.

2.3 Methods

We proposed pBoost framework which performs multiple tasks; feature engineering, data preprocessing (e.g., feature transformation, missing values imputation, removing multicollinearity, etc.), optimize the classifier setting with Bayesian optimization technique, provide performance metrics and the features with the most significant effect on the model's decision-making process. The GitHub repository for pBoost framework is publicly available.

We performed a retrospective cohort study of all hospitalized patients with a CVL at a single pediatric health system. Patients were included in the cohort if they were admitted to one of three freestanding children's hospitals between January 1st, 2013 and December 31st, 2018 and had a central line documented in the system before or at the time of admission or received at least one CVL during the hospitalization. We extracted data routinely available across electronic health records systems including demographics, vital signs, laboratory values, prior diagnoses, medication administrations, microbiology results, respiratory support, CVL properties, and CVL care documentation. The full set of data collected are listed in Appendix A.

PSI T0 time is defined as the first specimen collection time associated with a blood culture after a central line is inserted, followed by four days of new antibiotic administration. For the patients who already had a CVL at the time of admission, we considered all the blood cultures on their records. For patients who had a CVL placed during the hospitalization, we only included blood cultures after their first central line was inserted (Figure 2.1-A). A patient may develop multiple PSIs during the stay at the hospital. We disregarded the first and last 12 h of the hospitalization then selected a random event time in the intermediate period to limit selection bias from specific admission/discharge workups.

Data was preprocessed, and machine learning features developed as described in Supplemental Appendix A and Figure 2.1-C. Model development was performed using XGBoost [22] and logistic regression with a L1L2 regularization also known as ElasticNet [112] (Figure 2.1-D). ElasticNet has the ability to identify more important features and penalize the less informative ones, and XGBoost is a scalable tree-based boosting technique that is popular for supervised machine learning involving highly imbalanced (i.e. low prevalence) and missing data. Additional details on the machine learning algorithms are described in Supplement Appendix A.

We calculated Area Under the Receiver Operating Characteristic curve (AUC) and Area Under the Precision-Recall Curve (AUCpr) performance metrics. AUC is a commonly used measure to present binary classification results. In the presence of a class-imbalanced dataset, AUCpr can provide a more informative insight into the classifier's performance relevant to clinician decision-making. We then calculated the true positive and true negative classification rates for each classification model.

We evaluated the importance of variables in the machine learning method by using SHaply Additive exPlanations (SHAP) [63, 62] values for pBoost which is a validated interpretability tool for machine learning models that provides the average marginal

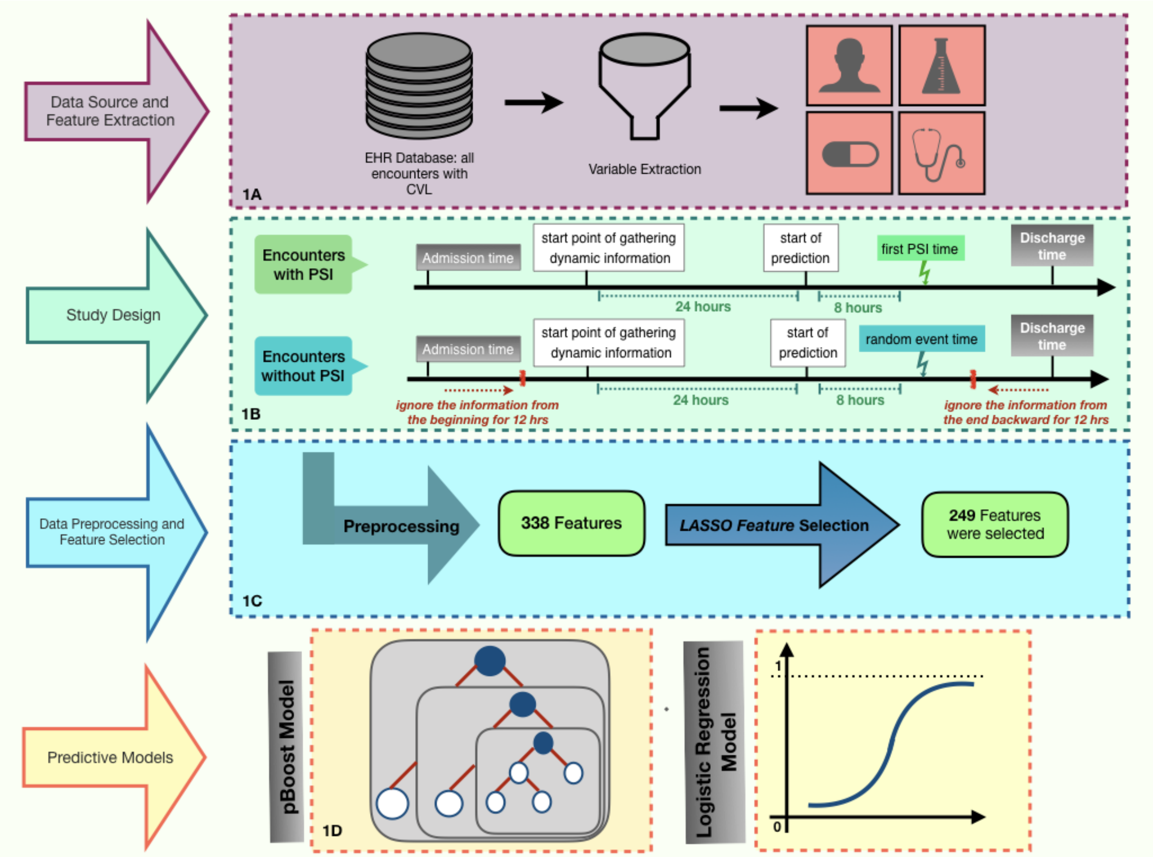

Data Source and Feature Extraction
EHR Database: all encounters with CVL
Variable Extraction
1A
Study Design
Encounters with PSI
Admission time
start point of gathering dynamic information
start of prediction
first PSI time
Discharge time
24 hours
8 hours
Encounters without PSI
Admission time
start point of gathering dynamic information
start of prediction
random event time
Discharge time
ignore the information from the beginning for 12 hrs
24 hours
8 hours
ignore the information from the end backward for 12 hrs
1B
Data Preprocessing and Feature Selection
Preprocessing
338 Features
LASSO Feature Selection
249 Features were selected
1C
Predictive Models
pBoost Model
1D
Logistic Regression Model
1
0

Figure 2.1: Summary of the methods. (1-A) A cohort of patients with CVL and demographic, clinical, and laboratory characteristics were extracted from the EMR database. (1-B) If an encounter has a PSI, the time of PSI (t_{PSI}) is marked as the event time and the prediction time is set at 8 h prior (t_{PSI}- 8 h). Minimum and maximum of the time-variant variables, such as laboratory results and vital signs, are considered from 24 h prior to the prediction time. If a patient encounter does not have a PSI, the first and last 12 h of patient information is ignored. Then a random point in time is selected as the hypothetical event time, and the prediction time is set at 8 h prior. (1-C) There were 338 features extracted in the study. To eliminate the features without a significant effect on the outcome, we applied LASSO feature selection, which led to a 249-dimensional feature space. (1-D) XGBoost and ElasticNet models were employed to predict if a patient will develop PSI during the next 8 h from the prediction time using the selected 249 features. Cross-validation and Bayesian optimization were applied in the training process of both models in order to find the best settings of the predictive models and avoid overfitting.

contribution of each feature to the prediction. Equation 2.1 demonstrates the SHAP calculations in which f is the classification model, F is the complete set of features, S is a subset of F, and feature i is the feature that we want to calculate its contribution to the outcome of the classification model. $f_{S\cup\{i\}}(x_{S\cup\{i\}})$ is the classification outcome using S and feature i while $f_S(x_S)$ is the outcome of the model withholding feature i. Therefore, the classification results with all possible combinations of feature in F, with and without feature i, is calculated and weighted by $\frac{|S|!(|F|-|S|-1)!}{|F|!}$. Then, ϕ_i will be the marginal contribution of feature i on the outcome of the classification model f and represents the SHAP value of feature i.

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \tag{2.1}$$

We compared the performance of the pBoost model to PRISM-III [78]. We calculated PRISM-III during the 24-hour interval prior to the PSI prediction start time. As demonstrated in Figure 2.1-B, the 24-hour time period before the start of the prediction is the same interval that the dynamic features for the ElasticNet and pBoost models are collected.

To assess the performance of PRISM-III in predicting PSI during the next 8 hours, we calculated the performance metrics of PRISM-III to predict PSI using different cut-off values on the training dataset. We then applied the optimal cut-off values on the testing dataset for performance comparisons.

This manuscript was prepared using the guidelines provided by Leisman et al. [53] for reporting of prediction models.

2.4 Results

We initially screened a total of 97,424 patient encounters associated with 15,704 patients, of which 65,766 encounters were excluded due to having length-of-stay less than 24 h. After applying the exclusion criteria mentioned in section 3, a total of 27,137 unique encounters were thus eligible to be included in the analysis. A total of 2,749 neonates, 4,076 infants, 5,580 toddlers and preschoolers, 6,500 children, and 8,232 adolescents met eligibility criteria. Figure A.1 in Supplement Appendix A presents the CONSORT diagram for this study.

To test the statistically significant difference between the features from PSI and non-PSI group, We applied Wilcoxon rank sum test and chi-squared test on the numerical and binary features, respectively. We observed a statistically significant difference between the median of age, weight, height, length of stay (LOS), and race in PSI and non-PSI groups, while there was no statistical difference in patients' sex between PSI and non-PSI groups (Table 2.1). The median age of patients in the PSI group was 2.9 years (IQR = [0.18, 11.8]), whereas in the non-PSI group, the median age was 6.5 years (IQR = [1.3, 13.6]). The median LOS for a patient with PSI was 30 days (IQR = [18.4, 54.2]), compared to 4.8 days (IQR = [3.1, 9.3]) for non-PSI.

Table 2.1: Cohort characteristics

	PSI	non-PSI	p-value
Age (years) (Median [25th, 75th])	2.9 [0.2, 11.8]	6.5 [1.3, 13.6]	< 0.001
Weight (Kg) (Median [25th, 75th])	13.1 [3.8, 37.6]	20.9 [9.5, 46.2]	< 0.001
Height (cm) (Median [25th, 75th])	89.9 [51.9, 142.3]	115 [72.5, 153.5]	< 0.001
Length of Stay (LOS) (Median [25th, 75th])	30 [18.4, 54.2]	4.8 [3.1, 9.3]	< 0.001
Gender			
Male (%)	45.9	45.4	0.52
Race			
Asian (%)	3.4	4	0.07
Caucasian (%)	49.7	54.9	< 0.001
African American (%)	41	35.3	< 0.001
American Indian or Alaska Native (%)	0.3	0.2	0.22
Native Hawaiian or Pacific Islander (%)	0.1	0.2	0.21
Other (%)	5.5	5.4	0.79
Insurance Status			
Commercial (%)	33.7	40.2	< 0.001
Public - Medicaid (%)	62.9	55.8	< 0.001
Public - non-Medicaid (%)	2.8	3.1	0.47
Self-pay (%)	0.6	0.9	0.02
ICU Admission (%)	70.5	41.1	< 0.001
Placed on Extracorporeal Membrane Oxygenation (%)	8.3	1.3	< 0.001
Mortality (%)	0.08	0.05	0.53

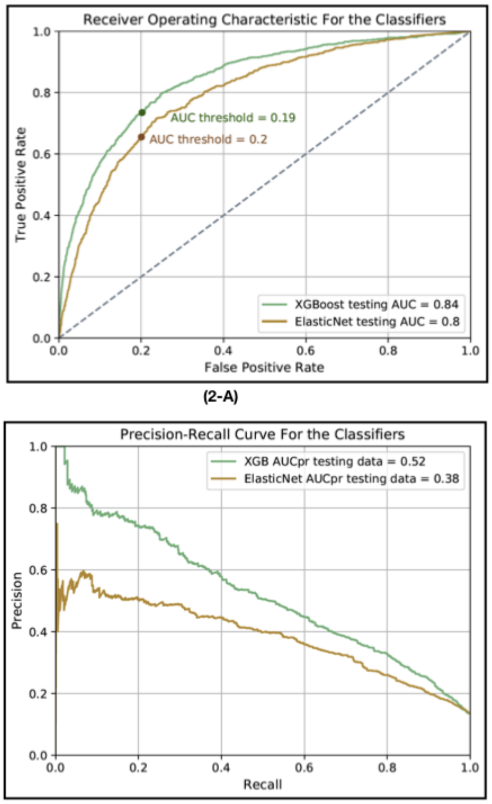
The results of the classifiers are presented in Figure 2.2 pBoost performed best in the testing dataset, with an average AUC of 0.84 (95% CI = [0.82, 0.85]) and AUCpr of 0.52 [0.51, 0.53] compared to the AUC of 0.79 [0.78, 0.79] And AUCpr of 0.38 [0.36, 0.39] for ElasticNet ($p < 0.001$). When fixing the specificity of both models at 0.80, the sensitivity of pBoost was 0.73 [0.69, 0.74] compared to sensitivity of 0.64 [0.62, 0.66] for ElasticNet ($p < 0.001$). The PPV of the pBoost model, 0.36 [0.34, 0.36], was slightly higher that the ElasticNet 0.33 [0.32, 0.34] ($p < 0.001$). However, the NPVs were nearly the same (pBoost 0.94 [0.94, 0.95], ElasticNet 0.93 [0.93, 0.94], $p < 0.001$).

Comparing the confusion matrices of pBoost and ElasticNet models in Figure 2.2, true positive cases increased while false positive and false negative cases decreased when employing pBoost instead of ElasticNet model.

Explanability: We identified which features contributed most to the prediction of PSI (Figure 2.3). The maximum value of the diastolic blood pressure in the 24 h prior to the prediction period was on average the leading predictor of PSI risk. The next most important features were height, maximum temperature, minimum Hemoglobin and minimum pulse oximetry. Multiple Complete Blood Count (CBC) components were also important for PSI prediction.

Comparison to PRISM-III: Table 2.2 demonstrates the performance of PRISM-III in predicting PSI. At a PRISM-III score of 10, the sensitivity was 0.19 [0.16, 0.22], a drop of 54% compared to pBoost, the PPV (0.30 [0.26, 0.34]) was reduced by 6%, and the NPV (0.88 [0.87, 0.90]) was 6% worse for the same comparison. Reducing the cut-off to 5 improved the sensitivity (0.48 [0.44, 0.51]); however, there was an 13% drop in PPV (0.17 [0.15, 0.19]) and 1% improvement in NPV (0.89 [0.88, 0.90]).

Sensitivity analysis of age groups: We compared PSI prevalence and model performance across all age groups. The highest and lowest PSI prevalence, 25.5% and 9.7%, were observed in neonates and children, respectively. We investigated the performance of pBoost model on each age group broken down by male and female



(2-A)

(2-B)

pBoost Model Total Testing records = 5428		True Labels		
		non-PSI	PSI	
Predicted Labels	non-PSI	TN = 3749 (TNR = 80%)	FN = 193 (FNR = 27%)	3942
	PSI	FP = 951 (FPR = 20%)	TP = 535 (TPR = 73%)	1486
		4700	728	

(2-C)

ElasticNet Model Total Testing records = 5428		True Labels		
		non-PSI	PSI	
Predicted Labels	non-PSI	TN = 3757 (TNR = 80%)	FN = 251 (FNR = 34%)	4008
	PSI	FP = 943 (FPR = 20%)	TP = 477 (TPR = 66%)	1420
		4700	728	

(2-D)

	pBoost	*ElasticNet*
AUC-ROC [95% CI]	**0.84 [0.82, 0.85]**	0.79 [0.78, 0.79]
Sensitivity [95% CI]	0.73 [0.69, 0.74]	0.64 [0.62, 0.66]
Specificity [95% CI]	0.80 [0.79, 0.80]	0.80 [0.79, 0.8]
Positive Predictive Value (PPV) [95% CI]	0.36 [0.34, 0.36]	0.33 [0.32, 0.34]
Negative Predictive Value (NPV) [95% CI]	0.94 [0.94, 0.95]	0.93 [0.93, 0.94]
F1 Score [95% CI]	0.46 [0.45, 0.47]	0.44 [0.42, 0.44]
Accuracy [95% CI]	0.78 [0.78, 0.79]	0.78 [0.77, 0.78]

(2-E)

Figure 2.2: The predictive models' performance. (2-A) The Receiver Operating Characteristic (ROC) plot for pBoost and ElasticNet models applied to the testing dataset. This plot demonstrates the trade-off between the sensitivity and specificity of the classifiers. To find the optimum AUC threshold, the specificity is fixed at 0.80 and the rest of the metrics are calculated. The selected AUC threshold is marked on each curve. (2-B) Precision-Recall curve (PRC) for pBoost and ElasticNet models applied to the testing dataset. A PRC plots the positive predictive value (precision or PPV in the y-axis) against the true positive rate (recall or sensitivity in the xaxis). In class-imbalanced data classification, it is more informative to look at both ROC and PRC to consider the trade-off between PPV and sensitivity. (2-C) The confusion matrix of applying the pBoost model on the testing dataset. This table presents the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values that are calculated based on the optimal AUC threshold marked on the ROC plots in part 2-A of this figure. The total number of PSI and non-PSI records and the total number of predicted labels of each class are mentioned in the confusion matrix. (2-D) The confusion matrix associated with the ElasticNet classifier. (2-E) Two predictive models, ElasticNet and pBoost, are employed to predict if a given patient encounter will develop PSI during the next 8 h of hospital stay. The AUC values are reported for testing subsets of the data. To make a better comparison among the results, the specificity level is fixed at 0.80 and the rest of the performance measurements are calculated subsequently. The mean and 95% confidence interval of each metric are reported.

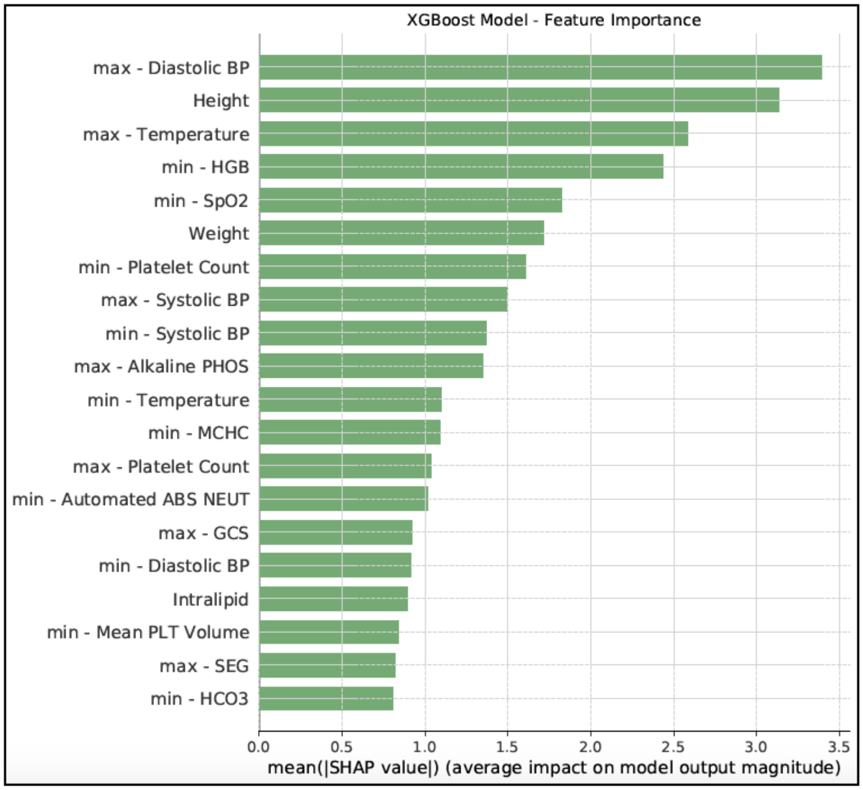


Figure 2.3: Feature importance from pBoost model This figure presents the top 20 features according to the pBoost model using SHAP values. In tree-based models, there are different criteria to sort features based on their effect on the outcome. Employing SHAP values in tree-based models, such as pBoost, is the most reliable and consistent way to calculate feature importance. SHAP values do not provide the direction of the effect, so there is no information regarding if an important feature positively or negatively affects the outcome. The min or max mentioned in the feature names, refers to the minimum or maximum of the time-variant variables, such as laboratory results or vital signs, during the 24 h prior to the PSI prediction time.

Table 2.2: Performance of PRISM-III score as the clinical benchmark to predict PSI in the next 8 hours of hospitalization. On the training dataset, different cut-off points are examined for the PRISM-III threshold such that if a PRISM-III score of a patient is equal or greater than that threshold, then the predicted value will be that patient will develop PSI in the next 8 hours of hospital stay. We selected the PRISM-III cut-off values that lead to better sensitivity and specificity values. Then, check the selected threshold performances on the testing dataset. The mean and 95% confidence interval for each metric is presented in this table.

	cut-off value = 5		cut-off value = 6		cut-off value = 7		cut-off value = 8		cut-off value = 9		cut-off value = 10	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Sensitivity Mean [95%CI]	0.50 [0.48, 0.52]	0.48 [0.44, 0.51]	0.45 [0.43, 0.47]	0.44 [0.40, 0.47]	0.35 [0.34, 0.37]	0.34 [0.30, 0.38]	0.29 [0.27, 0.31]	0.28 [0.24, 0.31]	0.23 [0.22, 0.25]	0.22 [0.19, 0.25]	0.19 [0.17, 0.20]	0.19 [0.16, 0.22]
Specificity Mean [95%CI]	0.64 [0.63, 0.64]	0.64 [0.62, 0.65]	0.68 [0.67, 0.69]	0.68 [0.67, 0.70]	0.82 [0.81, 0.82]	0.82 [0.80, 0.83]	0.86 [0.86, 0.87]	0.87 [0.86, 0.88]	0.90 [0.90, 0.91]	0.91 [0.90, 0.91]	0.93 [0.93, 0.94]	0.93 [0.92, 0.94]
PPV Mean [95%CI]	0.18 [0.17, 0.18]	0.17 [0.15, 0.19]	0.18 [0.17, 0.19]	0.18 [0.16, 0.19]	0.23 [0.22, 0.24]	0.22 [0.20, 0.25]	0.25 [0.23, 0.26]	0.24 [0.21, 0.27]	0.27 [0.25, 0.29]	0.27 [0.23, 0.30]	0.29 [0.27, 0.31]	0.30 [0.26, 0.34]
NPV Mean [95%CI]	0.89 [0.89, 0.90]	0.89 [0.88, 0.90]	0.89 [0.88, 0.89]	0.89 [0.88, 0.90]	0.89 [0.89, 0.90]	0.89 [0.88, 0.90]	0.89 [0.88, 0.89]	0.89 [0.88, 0.89]	0.88 [0.88, 0.89]	0.88 [0.87, 0.89]	0.88 [0.88, 0.89]	0.88 [0.87, 0.90]
F-1 Score Mean [95%CI]	0.26 [0.25, 0.27]	0.25 [0.23, 0.27]	0.26 [0.24, 0.27]	0.26 [0.23, 0.27]	0.28 [0.27, 0.29]	0.27 [0.24, 0.30]	0.27 [0.25, 0.28]	0.26 [0.23, 0.29]	0.25 [0.23, 0.26]	0.24 [0.21, 0.27]	0.23 [0.21, 0.25]	0.23 [0.20, 0.27]

(Tables A.3 to A.8 in Supplemental Digital Content). The pBoost model performed best on adolescents with average AUC of 0.84 [0.83, 0.85] with slight drop in performance in other age groups; an average AUC of 0.81 [0.80, 0.83] for children, 0.80 [0.78, 0.82] for toddlers and infants, and 0.74 [0.72, 0.76] for neonates.

2.5 Discussion

In this study, we developed a novel algorithm to predict the risk of PSI among hospitalized children using readily available features from the EMR. The pBoost model has a high NPV and a PPV of nearly three times the baseline PSI prevalence. We found that diastolic blood pressure, height, temperature, and CBC components were the most useful clinical variables for predicting PSI, consistent with other measures of patient acuity. The pBoost model is based on the XGBoost method, which enhances prediction of low prevalence events through boosting, where misclassified cases from one round of model training are weighted more heavily in subsequent rounds. This approach outperformed the ElasticNet model which does not have any boosting structure. Both models had statistically superior performance in predicting PSI compared to PRISM-lll alone. While PRISM-III score is a wellestablished clinical benchmark, we included more features in the pBoost framework that may help with boosting the prediction performance of the machine learning model. Moreover, a machine learning model does not have the limitation of the clinical benchmark and can learn the patterns for PSI positive and negative patient conditions regardless of their PRISM-III score. The superior performance of the pBoost framework demonstrated the potential of machine learning techniques to inform infection prevention and management.

PSI is a relatively new definition of serious infection compared to sepsis and CLABSI. However, its simplicity and ease of implementation is advantageous particularly in pediatrics. In our study, we found that patients with PSI had a longer LOS (median of 30 days vs. 4.8 days, pvalue < 0.001), were more likely to be admitted to the ICU (70.5% vs. 41.1%, p-value < 0.001), more likely to be placed on extracorporeal membrane oxygenation (8.3% vs. 1.3%, p-value < 0.001), and a higher mortality rate (0.08% vs. 0.05%, p-value $= 0.53$). We also found that PSI was more common in African Americans (41% vs. 35.3%, p-value < 0.001) and those with Medicaid insurance (0.62.9% vs. 55.8%, p-value < 0.001), consistent with health disparities seen in adult patients with sepsis [19]. Thus, efforts aimed at predicting, preventing and managing PSI are likely to improve outcomes for children.

Diastolic blood pressure had the largest contribution to the PSI prediction, consistent with prior studies of early predictors of pediatric sepsis [47]. Other studies in pediatric or adult sepsis prediction identified body temperature [48, 52], hemoglobin [87], SpO2 [52], platelet count [87] and systolic blood pressure [65] as the features with significant impacts on sepsis prediction, which were aligned with our findings from investigating important features by mean SHAP values.

The predictive performance of pBoost as measured by AUC was superior to other models of pediatric sepsis relying on EMR data alone. Le et al. [52] and Desautels et al. [27] achieved an AUC of 0.73 to predict pediatric severe sepsis four hours prior to the onset using boosted ensembles of decision trees. Several factors may contribute to higher AUC in our model; first, PSI has higher prevalence than severe sepsis and may be easier to predict. Second, we restricted our population to patients with CVL which increases the prevalence of the serious infection and may facilitate prediction. Finally, in our study we used a much broader range of features which may provide additional predictive power compared to vital signs and a small set of laboratory values.

Limitations: Our study has several important limitations. First, even though our data was derived from a large multi-hospital tertiary pediatric health system, it represents a single instance of the EMR and may reflect workflows specific to this setting. Therefore, any practices and procedures that are unique to the system may bias the outcomes of the model if applied to an external site. Second, we investigated only two methods of predictive modeling specifically to handle imbalanced (i.e. low prevalence) data. Deep learning approaches may improve on the performance achieved with XGBoost and ElasticNet. We also did not consider unstructured data, which may contain further information about the patient's acuity and infection risk. This represents a future opportunity to improve on the performance of our model. Third, our patient cohort represents a highly diverse population that is unique to the South, U.S.A, and therefore may limit the generalizability to other clinical contexts. Finally, our approach used a lookback method in which we compared performance of the model 8 h prior to known PSI events with 8 h randomly selected from encounters without PSI. This "lookback" method is a standard practice in the development and validation of many predictive models, but it nonetheless increases the prevalence of the event in our sample compared to what would be observed clinically moving forward in time. Thus, the PPVs of our models may be higher than what might be seen in prospective validation where predictions may be assessed at regular time intervals throughout a hospitalization.

Recent studies have suggested that machine learning algorithms can outperform clinical rules-based criteria [111]. In our study, the pBoost achieved a sensitivity and PPV (0.73 [0.69, 0.74] and 0.36 [0.34, 0.36]) higher than the benchmark using PRISM-lll (0.19 [0.16, 0.22] and 0.30 [0.26, 0.34], p-values < 0.001) at a cut-off of ≥ 10. This indicates that the improved performance of pBoost can meaningfully improve clinical decision making by alerting clinicians earlier, thereby allowing for more rapid intervention and by that potentially improving outcomes.

Chapter 3

Deep Learning Model to Predict Serious Infection among Children with Central Venous Lines

3.1 Abstract

Objective: Predict the onset of presumed serious infection, defined as a positive blood culture drawn and new antibiotic course of at least 4 days (PSI*), among pediatric patients with Central Venous Lines (CVLs).

Design: Retrospective cohort study. Setting: Single academic children's hospital. Patients: All hospital encounters from January 2013 to December 2018, excluding the ones without a CVL or with a length-of-stay shorter than 24 hours.

Interventions: None.

Measurements and Main Results: Clinical features including demographics, laboratory results, vital signs, characteristics of the CVLs and medications used were extracted retrospectively from electronic medical records. Data were aggregated across all hospitals within a single pediatric health system and used to train a deep learning model to predict the occurrence of PSI* during the next 48 hours of hospitalization. The proposed model prediction was compared to prediction of PSI* by a marker of illness severity (PELOD-2). The baseline prevalence of line infections was 0.34% over all segmented 48-hour time windows. Events were identified among cases using onset time. All data from admission till the onset was used for cases and among controls we used all data from admission till discharge. The benchmarks were aggregated over all 48 hour time windows [N=748,380 associated with 27,137 patient encounters]. The model achieved an area under the receiver operating characteristic curve of 0.993 (95% CI = [0.990, 0.996]), the enriched positive predictive value (PPV) was 23 times greater than the base prevalence. Conversely, prediction by PELOD-2 achieved a lower PPV of 1.5% [0.9%, 2.1%] which was 5 times the baseline prevalence.

Conclusion: A deep learning model that employs common clinical features in the electronic health record can help predict the onset of CLABSI in hospitalized children with central venous line 48 hours prior to the time of specimen collection.

3.2 Introduction

Central line-associated bloodstream infections (CLABSIs) are a major cause of healthcare-associated infections among hospitalized children and contribute to increased morbidity, length of hospital stay, and cost [73, 80]. The U.S. Centers for Disease Control and Prevention (CDC) estimates that approximately 80,000 new CLABSIs occur in the United States every year, and data show a 12%-25% increased risk of mortality in hospitalized patients who develop a CLABSI [86, 32]. Early identification of the onset of infections such as CLABSI or sepsis can prevent adverse outcomes, reduce costs, and improve the quality of care [69, 82].

While specific definitions for entities such as CLABSI and sepsis exist in pediatrics, they often have inadequate sensitivity for clinically important infections and may be difficult to generalize across electronic medical record (EMR) platforms [51, 5]. Presumed serious infection (PSI), which is used in both adult and pediatric sepsis surveillance systems, is defined as a blood culture being obtained (regardless of the result) followed by new antimicrobial agents started within 2 days of the blood culture (i.e., agents that were not being administered prior to the blood culture) that are administered for at least 4 consecutive days or until time of death or transfer to another hospital [42, 84, 98]. This PSI definition captures suspicion for infection (as identified by obtaining a blood culture) along with sufficient antimicrobial use to distinguish empirical treatment of a suspected infection from definitive treatment. Successful prediction of PSI, or sepsis in general, among hospitalized children or the adult population could expedite recognition and initiation of therapy [69].

Machine learning models have the potential to predict the onset of infection prior to clinical suspicion, allowing clinicians to take preventive measures and reduce mortality and morbidity [52, 26, 85, 79]. However, one of the main challenges in employing machine learning models in the clinical domain is that many events worthy of prediction are uncommon, also known as the extremely class-imbalanced dataset problem [88]. For example, in the pediatric cardiac intensive care unit (ICU), Alten et al. found that hospital acquired infection occurred in 2.4% of CICU encounters at a rate of 3.3/1000 CICU days [2]. To date, studies to predict CLABSI onset have mainly investigated known clinical risk factors associated with the infection and developed discriminative models based on non-temporal data [31, 75]. While these approaches may be able to predict if a CLABSI will occur during an entire hospital visit or not, their performance likely decreases when considering the next 48-72 hours of a patient's care. Real-time predictions that estimate the risk of an adverse event in a defined time window are more useful clinically, but they are more challenging to develop because the prevalence of the event in a defined time window is lower than its prevalence across an entire hospital stay [71, 104]. Currently a CLABSI prediction tool does not exist and instead providers use either subjective information or derived metrics such as severity of illness scores. In pediatrics, a commonly used severity of illness score is the PEdiatric Logistic Organ Dysfunction (PELOD) score. PELOD has been used to predict death and need or duration of intensive care unit resources [54].

Most traditional machine learning algorithms assume a balanced distribution of negative and positive samples in the data (i.e., a prevalence close to 50%). Deep learning models have the potential to overcome these limitations as they are more capable of finding patterns in extremely class-imbalanced high-dimensional data. However, deep learning models are commonly thought of as impossible to understand, overly complex, and not pragmatic. These models' lack of explainability may reduce their implementation effectiveness even with good predictive performance.

In this study, we aimed to develop a pragmatic deep learning framework that can adequately predict the onset of presumed bloodstream infection in children with a central line during the next 48 hours of their hospitalization. At each point of prediction, the model provides insights to its decision-making process by outputting the effect of the most influential features on the predicted outcome.

3.3 Material and Methods

3.3.1 Study Design

A retrospective cohort study was conducted which included all hospitalized patients with a central venous line (CVL) at a single tertiary care pediatric health system.

The inclusion criteria for patients were (1) admission to one of three freestanding children's hospitals between January 1st, 2013 and December 31st, 2018, (2) having a documented CVL at some point during the hospitalization (e.g., present and not yet removed at the time of admission or placed during the hospitalization), and (3) having length-of-stay longer than 24 hours. As described earlier, our goal is not to identify causes of presumed bloodstream infection associated with CVL, but rather predict the infection among patients with CVL. The predictive model was developed as it would be applied in clinical practice; therefore, we included both patients whose line was placed within the local health system or before admission. If CVL was placed within the local health system, information about line placement, such as sterile technique, was included. For patients whose line was not placed within the local health system, those data were not available to the model, just as they would not be available in the EHR when making a prediction in real clinical practice. This study was conducted according to Emory University protocol number 19-012.

3.3.2 Outcome Definition

We defined our primary outcome as a presumed serious infection (PSI) along with a laboratory confirmed bloodstream infection defined as a positive blood culture [42, 84]. We reviewed this definition through informal interviews with 2 pediatric infectious disease specialists, 1 pediatric critical care physician, 1 neonatologist, and 1 pediatric hematology/oncology specialist to validate its appropriateness and clinical utility. From this point, we referred to PSI with positive blood culture as PSI* for clarity reasons.

3.3.3 Feature Extraction

The extracted features from the EHR were demographics, laboratory results, vital signs, prior diagnoses, microbiology results, medications, respiratory support, CVL information, and CVL care documentation. We focused on features anticipated to be routinely recorded in the EHR across centers. The full list of extracted features and the preprocessing steps are available in Appendix A and Appendix B, respectively.

As initial deep learning techniques are often exploratory, it is true that many variables would on the surface seem unrelated. While biopathophysiologic links can indeed be created related to escalating PEEP (e.g., worsening microvascular/endothelial injury in the pulmonary vasculature potentially related to cytokine storm/inflammation as a response to a brewing infection or pulmonary edema from endovascular injury and leak and fluid delivery) – the beauty of a deep learning model approach is it reduces clinician bias that a variable (or set of variables) is or is not related to the outcome of interest. As the literature shows – many models have been able to identify constellations of variables that would go otherwise unheeded as heralds to a patient event [98, 75].

3.3.4 Window-Wise Study Design

The onset of PSI* is defined as a positive blood culture time after a CVL was inserted, succeeded by a new antibiotic administration for at least four days. Hospitalized patients in the cohort could have a CVL at the time of admission or received at least one during hospitalization time. We restricted our analysis to blood cultures with specimen collection timestamps while the patient had a CVL during hospitalization. A patient may become infected multiple times during a single hospitalization. However, for the purposes of this analysis, we censored hospitalizations after the first PSI* event for a patient if present.

To predict the onset of PSI* in a real-time setting, we used a window-wise study design (Figure 3.1). We started monitoring a patient from admission or first line insertion time, whichever was earlier. We then aimed to predict whether a patient would have a PS* in the next window of 48 hours; this prediction window was selected to give health providers enough time to intervene to potentially prevent a PSI*, for example by removing high risk CVLs or other interventions. Every 8 hours, the model would incorporate new information obtained and make another prediction for the subsequent 48 hours. The 8 hour sliding window was selected to reflect the cadence of shift changes and rounds, particularly in the ICUs at our institution. Even if the windows do not correspond specifically to shift changes and rounds, we nonetheless felt that more frequent updates would yield more relevant information for clinicians. All 48 hour windows that included a PSI* time were labeled as positive and the rest were negative.

The patient encounters were split into training (80%) and testing (20%). The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. We followed the commonly used 80-20 split in order to provide enough examples for the models to learn. Additionally, 10% of the training set was used as the validation set to optimize the model's settings and tune the model's hyperparameters. After preprocessing the data and removing collinearity, there were 135 features to feed into the prediction model. The list of 135 features and the details on preprocessing are described in Appendix B. The PSI* prevalence in the window-wise study was 0.34%, meaning that approximately 1/300 of the 48 hour time windows contained the onset of a PSI*.

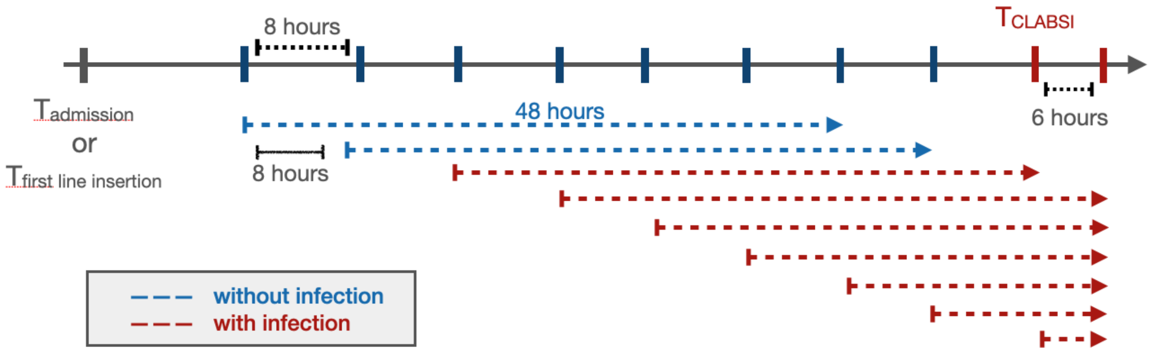


Figure 3.1: The window-wise study design. If a patient had a documented CVL that was not documented as removed at the time of admission, the start point of the analysis would be the admission time. Otherwise, the start point would be the first line insertion time. The prediction window was 48 hours with an 8 hour sliding window until the end of the patient's hospitalization or removal of the last CVL. When the onset of CLABSI occurred within a 48 hour prediction window, that window was considered positive (red), while the rest (blue) were labeled as negative. The prediction was performed at the start of each arrow.

3.3.5 Models

Real-time prediction of PSI* is an extremely class-imbalanced problem (see below). To tackle this challenge, we started with a Long Short-Term Memory (LSTM) model [40, 89], a recurrent neural network model capable of dealing with long sequences of data that has performed well for adult sepsis prediction [91]. To improve the performance of this model on an extremely class-imbalanced dataset, we hypothesized that:

Hypothesis 1: Penalizing false positives and false negatives in the optimization function (focal loss) will improve model performance. In extremely class-imbalanced modeling, the model is biased towards the majority class which in our case is not having an onset of PSI*. In machine learning models, a loss function value is a measure of how far off a model's prediction is from the actual outcome value, and the algorithms are optimized to minimize this value. Focal loss reduces the loss of well-classified examples, emphasizing the false positives and negatives [59]. We hypothesized that a focal loss function would improve performance relative to traditional methods for dealing with imbalanced data such as under-sampling the majority class.

Hypothesis 2: Incorporating an attention mechanism will improve model performance. An attention mechanism in deep learning assigns attention weights to source data at each time point, allowing the model to focus only on information relevant to the next prediction [6].

To evaluate these hypotheses, we developed and evaluated the following machine learning models: (1) a simple Bidirectional LSTM with binary cross-entropy, (2) a simple Bidirectional LSTM that was trained with an under-sampled majority class to make the labels more balanced, (3) a Bidirectional LSTM with Focal loss, and (4) a Bidirectional LSTM with Focal loss and an attention mechanism. More details on the proposed model are presented in Appendix B.

3.3.6 Performance Metrics

For each model, we calculated the Area Under the Receiver Operating Characteristics Curve (AUROC), sensitivity, specificity and accuracy. We also calculated metrics that are more informative in extremely class-imbalanced data classification models such as Area Under Precision-Recall Curve (AUPRC), positive predictive value (PPV), negative predictive value (NPV) and F-1 score. The 95% confidence interval estimation for each metric was calculated using bootstrapping.

3.3.7 Model Explainability

Decision making process of a deep learning model is often assumed to be overly complex. However, there are several ways to illuminate the decisions a model makes. It is also achievable to understand which features are the most salient in a model's prediction.

We estimated feature importance for each prediction by employing Shaply Additive exPlanations (SHAP) values, a method for explaining predictive models based on game theory [63]. SHAP values presents the contribution of each feature to the model's decision-making process and their effect size on the predicted outcome. These SHAP values can be summarized across the cohort or calculated for an individual model prediction to inform clinicians of the features influencing a specific prediction, providing model transparency and observability to the end user [68].

3.3.8 Clinical Benchmark

To make the model relevant, we compared performance against an existing model used for prediction of illness in hospitalized children. In the absence of a discrete prediction model used for prediction of line or bloodstream infections, we used the PEdiatric Logistic Organ Dysfunction 2 (PELOD-2) score. The PELOD-2 score has been validated for prediction of morbidity and mortality in hospitalized children. We calculated PELOD-2 at every prediction point, then considered different cut-off values to identify the PSI* positive windows (28). Applying the same threshold values on the testing set, we predicted the PSI* positive windows by the use of the corresponding PELOD-2 values for each prediction window.

Pediatric Risk of Mortality III (PRISM-III) has also been validated for mortality prediction in hospitalized children [78]. Calculating PRISM-III score enables the physicians to identify which patients require more urgent care and interventions. We investigated the differences in PRISM-III components across PSI* and non-PSI* time windows.

This manuscript was prepared using the guidelines provided by Leisman et al. [53] for reporting of prediction models.

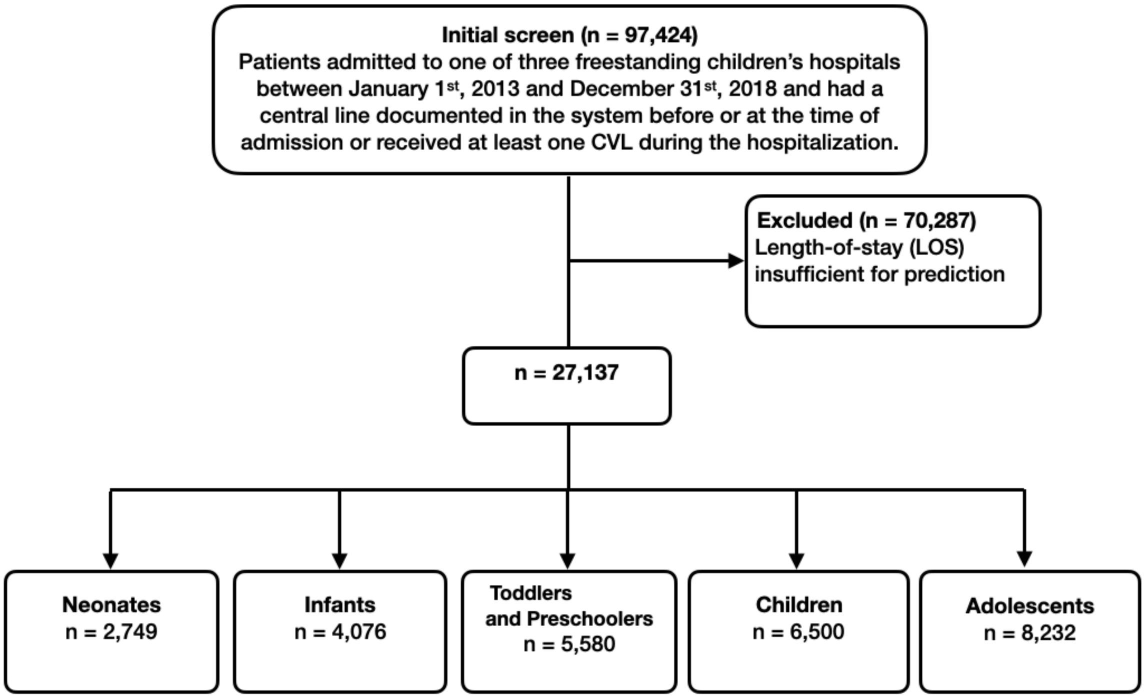

Initial screen (n = 97,424)
Patients admitted to one of three freestanding children's hospitals between January 1st, 2013 and December 31st, 2018 and had a central line documented in the system before or at the time of admission or received at least one CVL during the hospitalization.
Excluded (n = 70,287)
Length-of-stay (LOS) insufficient for prediction
n = 27,137
Neonates n = 2,749
Infants n = 4,076
Toddlers and Preschoolers n = 5,580
Children n = 6,500
Adolescents n = 8,232

Figure 3.2: Inclusion flowchart. The final number of patient visits that we used in training and testing the machine learning models were 27,137.

3.4 Results

In total, 97,424 patient encounters associated with 15,704 patients were extracted from the EHR. Of these, 70,287 encounters were excluded due to length-of-stay less than 24 hours (most likely representing appointments for patients with existing CVLs). A total of 2,749 neonates (age less than 28 days), 4,076 infants (age between 28 days and one year), 5,580 toddlers and preschoolers (age between one and five years), 6,500 children (age between five and 12 years), and 8,232 adolescents (older than 12 years) met eligibility criteria. Figure 3.2 presents the associated CONSORT diagram.

Table 3.1 presents the cohort characteristics. There was a statistically significant difference between the median age, weight, and height with PSI* patients younger and smaller. Length of stay was significantly longer in patients with PSI*African American race and Medicaid insurance were significantly more common in patients with PSI*. There was no statistically significant difference in gender between PSI* and non- PSI* groups. Moreover, statistical tests were performed to investigate if there were statistically significant differences between the components of PELOD-2 and PRISM-III between PSI* and non-PSI* groups across time windows (Appendix B).

Table 3.1: Cohort characteristics

	PSI*	non-PSI*	p-value
Age (years) (Median [25th, 75th])	3.6 [0.2, 12.6]	6.1 [1, 13.4]	< 0.001
Weight (Kg) (Median [25th, 75th])	14.3 [3.8, 40.9]	20 [8.5, 45.3]	< 0.001
Height (cm) (Median [25th, 75th])	93 [52, 149]	112 [69, 152]	< 0.001
Length of Stay (LOS) (Median [25th, 75th])	36 [23, 69]	5 [3, 13]	< 0.001
Gender			
Male (%)	44.8	45.5	0.737
Race			
Asian (%)	4.1	3.9	0.796
Caucasian (%)	46.7	54.3	< 0.001
African American (%)	43.8	35.9	< 0.001
American Indian or Alaska Native (%)	0.4	0.2	0.292
Native Hawaiian or Pacific Islander (%)	0.2	0.2	0.992
Other (%)	4.7	5.5	0.465
Insurance Status			
Commercial (%)	34.5	39.4	0.025
Public - Medicaid (%)	62.1	56.6	0.013
Public - non-Medicaid (%)	2.9	3	0.933
Self-pay (%)	0.39	0.9	0.224
ICU Admission (%)	63.9	44.7	< 0.001
Placed on Extracorporeal Membrane Oxygenation (%)	8.7	2.1	< 0.001
Mortality (%)	0.20	0.06	0.19

The results of the four predictive models are presented in Table 3.2. Our proposed model, the Bidirectional LSTM with Focal loss and attention mechanism, outperformed the rest of the models with AUROC of 99.3% [99.0%, 99.6%] and AUPRC of 13.9% [10.6%, 18.0%]. The ROC and Precision-Recall curves of all trained models are presented in Figure 3.3. Fixing the sensitivity of all models to 85% to select a threshold, our proposed model's specificity was 99.4% [99.2%, 99.5%], F-1 was 9.9% [7.1%, 13.8%] and PPV was 7.7% [5.7%, 10.3%] which is 23 times the baseline PSI* prevalence (0.34%). All performance metrics except for sensitivity and NPV were statistically different from the other models' metrics ($p < 0.001$). Moreover, the model generated 0.049 [0.044, 0.054] false alarms per patient per day. In other words, there should be 34 positive PSI* per 10,000 48 hour time windows (prevalence of 0.34%). The results of the proposed model indicated that per 10,000 predictions which lead to X number of positive predictions, 7.7% of X will be the number of PSI* windows that were correctly predicted as positive. Besides, 99.9% of non-PSI* windows were correctly predicted as negative ones. Moreover, 85% of true PSI* were predicted correctly while 15% of the true PSI* time windows were predicted as negative ones.

Table 3.2: Performance metrics of the deep learning models in predicting PSI* in the next 48 hours of hospitalization. The two numbers in the brackets present the estimated 95% confidence interval using bootstrap sampling.

	BiLSTM		BiLSTM + Under-sampling		BiLSTM + Focal Loss		BiLSTM + Focal Loss + Attention	
	Train	Test	Train	Test	Train	Test	Train	Test
AUROC Mean (%) [95%CI]	88.4 [86.7, 89.9]	89.3 [86.6, 91.5]	88.3 [86.8, 89.6]	85.7 [82.5, 88.2]	92.8 [90.8, 94.8]	91.1 [86.8, 94.9]	99.7 [99.6, 99.7]	99.3 [99.0, 99.6]
Sensitivity Mean (%) [95%CI]	85.1 [85.0, 85.2]	85.3 [78.3, 91.6]	85.1 [85.0, 85.2]	81.1 [72.7, 88.4]	85.1 [85.0, 85.2]	83.3 [75.3, 90.2]	85.1 [85.0, 85.2]	72.9 [62.8, 82.1]
Specificity Mean (%) [95%CI]	84.3 [83.6, 85.2]	84.2 [83.2, 85.2]	84.0 [83.3, 84.7]	83.2 [82.4, 83.9]	93.6 [92.5, 94.7]	93.2 [92.0, 94.3]	99.4 [99.2, 99.6]	99.4 [99.2, 99.5]
PPV Mean (%) [95%CI]	0.4 [0.3, 0.4]	0.4 [0.3, 0.5]	0.4 [0.3, 0.4]	0.3 [0.3, 0.4]	1.0 [0.8, 1.2]	0.9 [0.7, 1.1]	9.4 [6.9, 12.5]	7.7 [5.7, 10.3]
NPV Mean (%) [95%CI]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]
Accuracy Mean (%) [95%CI]	84.3 [83.6, 85.2]	84.2 [83.2, 85.2]	84.0 [83.3, 84.7]	83.2 [82.4, 83.9]	93.5 [92.5, 94.7]	93.1 [92.0, 94.3]	99.4 [99.2, 99.6]	99.3 [99.2, 99.5]
F-1 Score Mean (%) [95%CI]	0.8 [0.7, 0.9]	0.8 [0.6, 0.9]	0.8 [0.7, 0.9]	0.7 [0.6, 0.8]	3.3 [2.3, 4.2]	2.6 [1.4, 3.9]	58.0 [46.0, 70.8]	16.1 [10.4, 22.5]
AUPRC Mean (%) [95%CI]	0.4 [0.3, 0.4]	0.4 [0.3, 0.5]	0.3 [0.3, 0.4]	0.3 [0.2, 0.3]	3.9 [3.1, 4.7]	3.2 [1.9, 5.3]	80.7 [76.3, 84.6]	41.2 [30.7, 50.2]

Explainability: For the final model, we calculated SHAP values of each feature at every prediction point. Figure 3.4 presents the most important features for a specific timestamp in which the model predicted positive PSI*. For this patient, temperature had the highest effect size on the predicted outcome, followed by rinse agent, which was used to remove germs from the mouth, and platelet count.

Comparison to PELOD-2: The performance of PELOD-2 in window-wise prediction of PSI* is presented in Table 3.3. The cut-off points that yield higher performance metrics are listed. On the testing set, the best PPV was achieved at a cut-off point of PELOD-2 = 8 (1.5% [0.9%, 2.1%]) which was almost 5 times the baseline

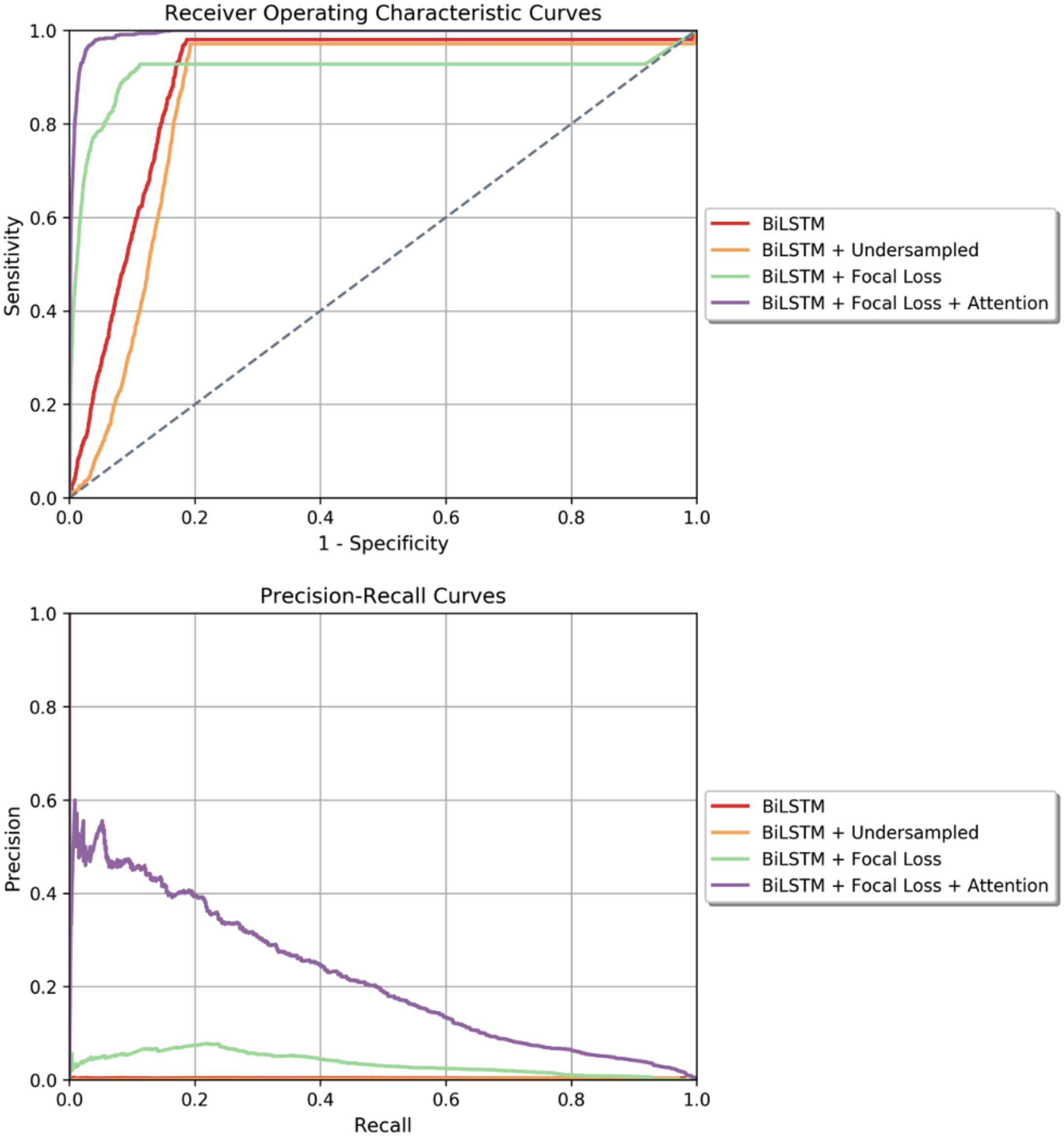


Figure 3.3: (Top) Receiver Operating Characteristics curves for all four models tested in the window-wise study. (Bottom) Precision-Recall curve for all the models tested in the window-wise study. In both plots, our proposed model which is the Bidirectional LSTM with Focal loss and attention mechanism achieved the highest area under curve.

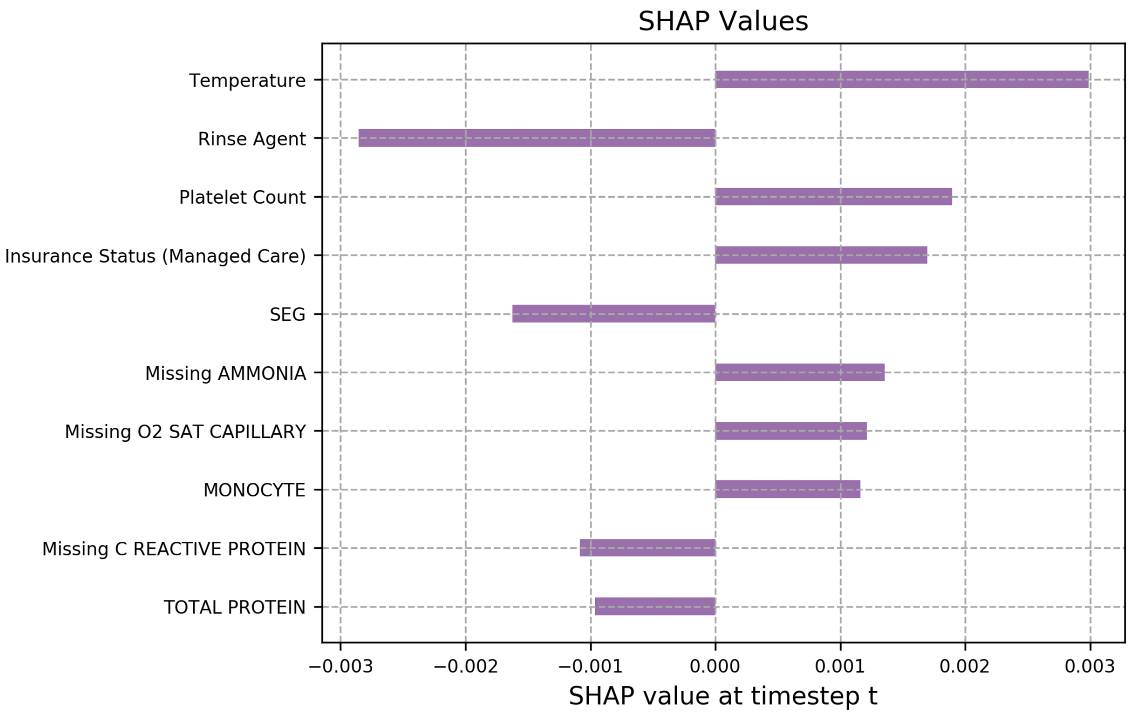


Figure 3.4: Feature importance plot based on SHAP values for an example prediction in which the model predicted the patient would develop a CLABSI within the next 48 hours.

prevalence. At this cut-off value, the sensitivity was 3.2% [1.8%, 4.5%], specificity was 99.2% [99.2%, 99.3%], F-1 was 2% [1.2%, 2.9%]. Comparing to the proposed model, there were lower values achieved for PPV (6.2% drop), sensitivity (69.7% drop), F-1 (7.9% drop) but specificity of PELOD-2 model was almost similar to the proposed model.

3.5 Discussion

Many important clinical events where accurate predictions could improve outcomes such as sepsis, deterioration, or cardiac arrest are rare, especially in pediatrics [108, 28, 103]. The prevalence of these conditions would be even lower if estimated over 48 hour time intervals during hospitalization instead of only counting the final outcome over an entire hospital stay. The techniques described in this study would likely translate to prediction of other clinical events with extreme class imbalance.

Table 3.3: Performance of PELOD-2 score in predicting PSI* in the next 48 hours of hospitalization. Different thresholds are selected for the PELOD-2 values. If a score exceeded the threshold, the predicted outcome for that patient would be developing PSI* during the next 48 hours of hospitalization. The two numbers in the brackets present the estimated 95% confidence interval using bootstrap sampling.

	PELOD-2 Threshold=4		PELOD-2 Threshold=6		PELOD-2 Threshold=8		PELOD-2 Threshold=10	
	Train	Test	Train	Test	Train	Test	Train	Test
Sensitivity Mean (%) [95%CI]	46.0 [44.0, 47.8]	52.0 [48.4, 55.9]	6.6 [5.6, 7.6]	7.4 [5.5, 9.5]	2.0 [1.5, 2.6]	3.2 [1.8, 4.5]	0.7 [0.4, 1.0]	0.6 [0.0, 1.2]
Specificity Mean (%) [95%CI]	83.2 [83.1, 83.2]	82.7 [82.5, 82.8]	97.2 [97.2, 97.3]	97.1 [97.0, 97.2]	99.2 [99.2, 99.2]	99.2 [99.2, 99.3]	99.8 [99.8, 99.8]	99.9 [99.8, 99.9]
PPV Mean (%) [95%CI]	0.9 [0.9, 1.0]	1.1 [1.0, 1.2]	0.8 [0.7, 0.9]	0.9 [0.7, 1.2]	0.8 [0.6, 1.1]	1.5 [0.9, 2.1]	1.2 [0.6, 1.8]	1.5 [0.0, 3.1]
NPV Mean (%) [95%CI]	99.8 [99.8, 99.8]	99.8 [99.8, 99.8]	99.7 [99.7, 99.7]	99.6 [99.6, 99.7]	99.7 [99.7, 99.7]	99.6 [99.6, 99.7]	99.7 [99.7, 99.7]	99.6 [99.6, 99.7]
Accuracy Mean (%) [95%CI]	83.0 [82.9, 83.1]	82.5 [82.4, 82.7]	96.9 [96.9, 97.0]	96.8 [96.7, 96.9]	98.9 [98.9, 98.9]	98.9 [98.8, 98.9]	99.5 [99.5, 99.5]	99.5 [99.5, 99.5]
F-1 Score Mean (%) [95%CI]	1.8 [1.7, 1.9]	2.1 [1.9, 2.4]	1.4 [1.2, 1.7]	1.7 [1.2, 2.1]	1.2 [0.9, 1.5]	2.0 [1.2, 2.9]	0.8 [0.5, 1.3]	0.9 [0.0, 1.8]

We developed a novel algorithm to predict a presumed serious infection in a hospitalized pediatric patient within 2 hospital days. Besides having a decent predictive performance, our proposed model employed SHAP values which explained the effect of the salient features on the risk of a PSI* event. Moreover, the SHAP values present the most influential features specific to a patient in a given time; therefore, these values can dynamically change through time as the condition of a patient changes. SHAP values give insight to the model's decision-making process by providing transparency and observability to the end-user of the features most important to model prediction. Insight into the model's focus for a specific prediction allows the end user to calibrate trust in the prediction.

Predictive models intended for use in clinical environments must recognize the complex adaptive systems in which they will be implemented [68, 57]. The sensitivity and PPV of the model can inform the appropriate time in workflow where the model would be most useful. Our model demonstrates strong enrichment (i.e., the PPV is 23 times higher than the baseline prevalence of PSI*) while maintaining good sensitivity, but the PPV is nonetheless quite low – only 1 of every 13 predictions developed a PSI* in the subsequent 48 hours. This apparent low PPV is in large part due to the window-wise design which lowers the apparent prevalence of PSI* relative to using an entire encounter as the unit of analysis. Thus, we anticipate this model would be more likely to be used as a non-interruptive monitoring system (e.g., displayed on patient lists) that can segregate out low-risk patients (NPV 0.999) while informing clinicians' estimate of the risk of PSI* in order to make decisions about line maintenance and interventions. Similarly, the model could direct attention for teams reviewing vascular access across a unit or a hospital to improve the efficiency of PSI* prevention efforts.

In our study cohort, PSI* was more common in African American patients and those with Medicaid insurance. While this analysis was not designed to describe disparities or their sources, this finding was nonetheless consistent with health disparities seen in adult sepsis patients [19]. Model performance was not significantly different by patient race or insurance status (Appendix B). We also performed sensitivity analysis based on patient age and included the results in Appendix B.

Our study has important strengths and limitations. We also had several limitations. First, our data was associated with a single pediatric health system and may reflect the particular structure and patient mix of this setting. While we extracted EHR features expected to be available across systems, the external application of our model on other health systems may be biased. Nonetheless, limiting to structured EHR data likely reduces the technical barriers to implementation in a real-time system. Second, our model was developed and evaluated based on a retrospective cohort. While we attempted to simulate prospective implementation using a window-wise design, predictive performance may deteriorate when implemented in real time. Third, we have not evaluated how these predictions would supplement clinical decision-making when clinicians determine to remove a CVL or change their interventions. Thus, it is possible that implementation at this or even a higher level of predictive performance may not change outcomes. Fourth, we included patients with CVLs placed prior to admission. While inclusion of CVLs placed prior to admission may lower predictive performance since the model has fewer data available, we nonetheless felt it important to include as this reflects the decision-making clinicians must make in reality about all CVLs whether placed locally or not. Finally, we benchmarked our comparison versus a standard of illness score. While not intended for the prediction of infections, PELOD, along with other scores such as the Pediatric Risk of Mortality (PRISM) and Pediatric Index of Mortality (PIM) scores are currently the only standard that exist to identify the risks of morbidity and mortality in hospitalized children. Thus, it would not be expected for these scores to have strong predictive performance for PSI* associated with CVL. Nonetheless, we demonstrate our model's additional value when applied to this use case compared to existing severity scores.

We only included structured data in our analyses while unstructured data are known to have benefits when used in predictive models. Including text data or waveform data in subsequent iterations may improve our prediction outcomes.

Chapter 4

A Machine Learning Pipeline for Integrating Structured and Unstructured Data for Timely Prediction of Bloodstream Infection among Children with Central Venous Lines

4.1 Abstract

Background: Hospitalized children with central venous lines (CVLs) are at higher risk of hospital acquired infections. Information in electronic health records (EHR) can be employed in training deep learning models to predict the onset of these infections. We investigated the effect of incorporating clinical notes in addition to structured EHR data to predict serious bloodstream infections, defined as a positive blood culture followed by at least four days of new antimicrobial agent administration, among hospitalized children with CVLs.

Methods: Structured EHR information and clinical notes were extracted for a retrospective cohort including all hospitalized patients with CVLs at a single tertiary care pediatric health system from 2013-2018. Deep learning models were trained to determine the added benefit of incorporating the information embedded in clinical notes in predicting serious bloodstream infection.

Results: A total of 24,351 patient encounters met inclusion criteria. The best-performing model restricted to structured EHR data had a specificity of 0.951 and positive predictive value (PPV) of 0.056 when sensitivity was set to 0.85. The addition of contextualized word embeddings improved the specificity to 0.981 and PPV to 0.113.

Conclusions: Integrating clinical notes with structured EHR data improved the prediction of serious bloodstream infections among pediatric patients with CVLs.

4.2 Introduction

Children with central venous lines (CVLs) are at higher risk of the adverse outcomes associated with hospital acquired infections such as central line-associated bloodstream infection (CLABSI) and sepsis. The U.S. Centers for Disease Control and Prevention estimates that approximately 80,000 new CLABSIs occur in the United States every year, and hospitalized patients who develop CLABSI have a 12%-25% increased risk of mortality [86, 32].

The increasing use of electronic health records (EHRs) in the healthcare domain along with advanced computational techniques lead to the opportunities to create reliable and generalizable population-level monitoring systems which incorporate routinely captured clinical data without the need to conduct resource-intensive chart reviews [106, 98, 52, 66]. In recent years, a number of studies have been conducted on the application of advanced analytics of structured EHR data to improve detection and prediction of the adverse outcomes in the hospital [82, 71, 79]. In our own work, we used structured EHR data to predict presumed serious infections (PSI) and serious bloodstream infections in hospitalized children [98, 97].

Clinical notes written by health providers are rich sources of a patient's health status through hospitalization time. While this information has previously been inaccessible to predictive models, more recent natural language processing (NLP) techniques show promise in harnessing the information embedded in unstructured EHRs for aiding clinical decisions [43, 1, 110]. Incorporating structured and unstructured EHR data can boost predictive performance and lead to more accurate results. For example in adult sepsis prediction, Amrollahi et al. integrated structured and unstructured EHR data to predict the onset of sepsis among ICU patients [3]. The results showed an improvement in the predictive model's performance compared to only using the structured EHR data. Similarly, Liang et al. incorporated clinical notes to train a disease classifier to predict a clinical diagnosis for pediatric patients [58]. However, these approaches have not been applied to serious bloodstream infections in hospitalized children.

In this study, we investigated the added benefit of integrating structured EHR data with unstructured data gleaned from clinical notes in predicting serious bloodstream infection, defined as a positive blood culture coupled with at least 4 days of new intravenous antibiotics, among hospitalized children with CVLs. We propose a data fusion approach and predictive model that can be employed prospectively in the pediatric ward to predict the risk of a serious bloodstream infection developing during the next 48 hours of the hospitalization.

4.3 Material and Methods

4.3.1 Study Population

Electronic health records, including structured and unstructured data, were extracted for a retrospective cohort of all hospitalized patients with a central venous line (CVL) at a single tertiary care pediatric health system. The inclusion criteria were admission to one of three freestanding children's hospitals between January 1st, 2013 and December 31st, 2018, having a documented CVL at some point during the hospitalization, having length-of-stay longer than 24 hours and having recorded clinical notes. A complete list of structured information extracted from EHRs is included in Appendix A. This study was approved by the Emory University Institutional Review Board (protocol number 19-012).

4.3.2 Outcome Definition

PSI was initially proposed as a part of pediatrics sepsis surveillance definition by Hsu et al. [42] and has previously been validated [107]. Among pediatric patients, PSI was defined as a blood culture drawn and new antibiotic course of at least 4 days (fewer if patients die or are transferred to hospice or another acute care hospital). The minimum of 4 days of antibiotic administration was selected to minimize the false positives from patients for whom the suspected infection was not confirmed and had the empirical treatment stopped. Our primary outcome of serious bloodstream infection was defined as a PSI along with a laboratory confirmed bloodstream infection defined as a positive blood culture [42, 84]. We reviewed this definition through informal interviews with 2 pediatric infectious disease specialists, 1 pediatric critical care physician, 1 neonatologist, and 1 pediatric hematology/oncology specialist to validate its appropriateness and clinical utility. From this point, we referred to PSI with positive blood culture as serious bloodstream infection (SBI).

4.3.3 Data Preprocessing

We used a window-wise study design, as presented in Figure 4.1, to predict the onset of SBI in a real-time setting. The start point of the study was the admission time or line insertion time, whichever was earlier. We aimed to predict whether the patient would develop SBI in the next 48-hour window; the 48-hour prediction window provides enough time for health providers to intervene and potentially prevent a SBI event. In the proposed study design, the SBI prediction was done every 24 hours using the most recent information. The 24-hour sliding window was selected to ensure that the recorded clinical notes of a patient were updated. If a 48-hour sliding window included an onset of SBI, that window was considered as a positive one. Overall, the prevalence of the positive windows was 0.35% which indicated an extremely imbalanced data problem. Stratified sampling was used to split patient encounters to training (80%) and testing (20%) sets. Moreover, 10% of the training patient encounters were employed as the validation set to optimize the hyperparameters of the models.

Structured EHR Data. Initially, the structured data included 252 features. The numerical features were transformed, imputed and standardized. The categorical features were one-hot-encoded. We removed multicollinearity with a threshold of 0.8. Finally, there were 129 features from the structured data to include in the analysis. Appendix C includes more details on the preprocessing steps along with a list of the selected features.

Unstructured EHR Data. All the provider notes recorded for a patient during the same time-window were concatenated. To reduce the effect of the redundant

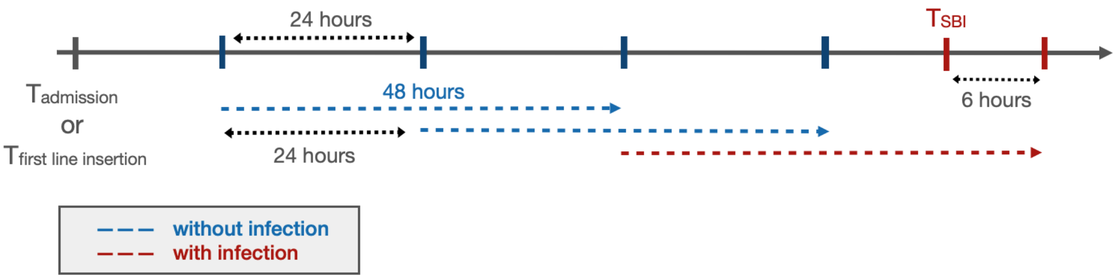


Figure 4.1: Window-Wise Study Design. If a patient had a documented CVL at the time of admission, the start point of the analysis would be the admission time. Otherwise, the start point would be the first line insertion time. The prediction window was 48 hours with a 24 hour sliding window until the end of the patient's hospitalization or removal of the last CVL. When the onset of SBI occurred within a 48 hour prediction window, that window was considered positive (red), while the rest (blue) were labeled as negative. The prediction was performed at the start of each arrow. "CC BY 4.0"

parts of the clinical notes, we selected the sections with more discriminative information such as history of present illness, impression and plan, patient active problem list, medical decision making, etc. After that, common text preprocessing steps were applied in which all text was transformed to lower case and extra white spaces, punctuations and numbers were removed. Finally, the clinical notes were matched with the corresponding structured data through the text recording timestamps.

4.3.4 Feature Extraction from Clinical Notes

There are two main approaches to incorporate a pre-trained language model in the predictive models; first, fine-tuning a pre-trained language model for down-stream tasks, second, calculating the contextualized word embeddings and feeding them as features to a classification or regression model. We followed the latter approach as it empowered the integration of structured data and clinical notes.

The BERT model has yielded remarkable performance in the clinical domain compared to ELMo and non-contextual embeddings [93]. Recent studies have demonstrated that using a domain-specific model achieves better performance compared to nonspecific embeddings; therefore, we employed the Clinical BERT model, which was pre-trained on approximately two million clinical notes in MIMIC-III dataset [46], to

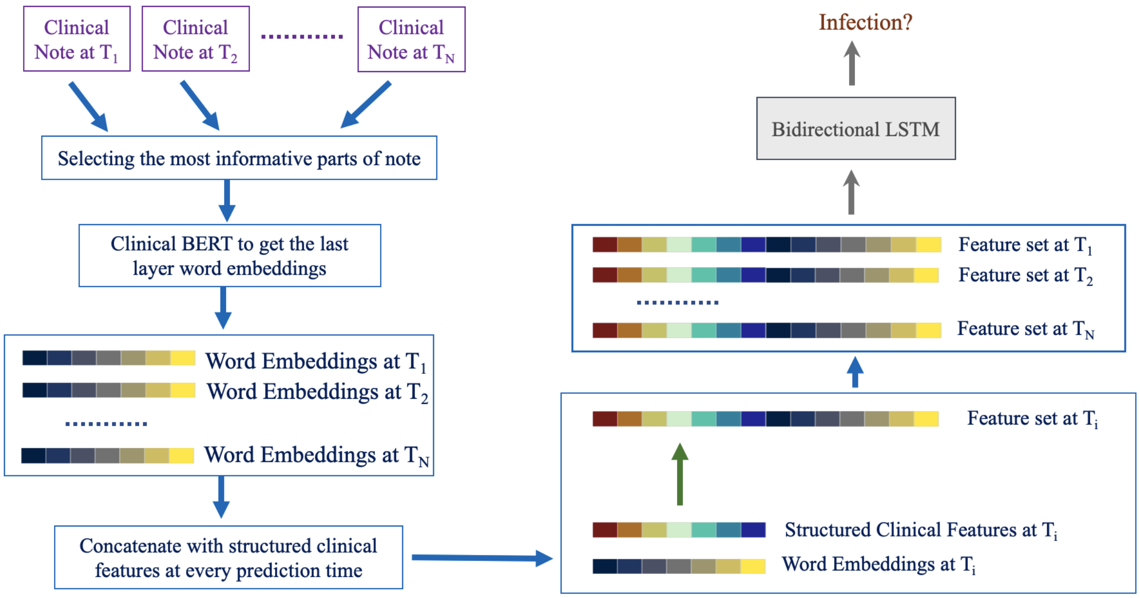


Figure 4.2: Data Fusion Diagram. The most informative sections of the clinical notes recorded for a patient at the time of prediction were selected and provided to the Clinical BERT model to calculate the contextualized word embeddings using the last hidden layer of the model. Then, the 768 dimensional contextualized word embeddings were concatenated with the 129 dimensional features from the structured EHR at every prediction point. The Bidirectional LSTM model with the attention mechanism and Focal loss incorporated the 897 dimensional input to predict if a SBI will occur during the next 48 hours of this patient's hospitalization. "CC BY 4.0"

acquire the contextualized word embeddings for the clinical notes in our cohort [1]. To assess the performance of the contextual word embeddings from the Clinical BERT model, we also extracted text features through the term frequency-inverse document frequency (TF-IDF) method.

Figure 4.2 demonstrates our approach to integrate the clinical notes with the rest of the structured clinical features to train the predictive models.

4.3.5 Predictive Models

Model Structure: We employed Bidirectional Long Short-Term Memory (BiLSTM) model as BiLSTMs can look at the information prior and successor of a given word in the note which is closer to human reading abilities and yields strong performance in the NLP domain [3].

Loss Function: There are two types of observations in a classification task; hard and easy. The hard observations are defined as the ones that confound the predictive model. These are the examples that the model should focus on to improve its overall performance. The extreme class-imbalanced problem in this study (prevalence of 0.35%) required a strategy to assign more weight to the minority class observations while taking the easy/hard examples into consideration to ultimately improve the true positive and true negative predictions. We employed Focal loss as a solution to this obstacle [59]. Focal loss is a loss function to lessen the weight of easy examples while intensifying the penalization in the case of an incorrect classification of hard examples.

Attention Mechanism: Attention mechanism in deep learning was motivated by how humans pay attention to different regions of an image or correlate words in a sentence [6]. When it comes to a class-imbalanced classification task, it is crucial to attend to the more prominent parts of the input sequence to achieve better results. Since we had long sequences of structured and unstructured data, the attention mechanism was incorporated to train the model to further attend to the more relevant parts of the input and explain the relationship between words in the context.

Training Process: We trained five models with the following input features to evaluate the benefit of integrating structured EHRs with clinical notes; (1) structured data, (2) extracted features from unstructured data with TF-IDF, (3) extracted contextualized word embeddings from unstructured data using Clinical BERT, (4) structured data along with the TF-IDF features, (5) structured data and the contextualized word embeddings. We trained all the models with a batch size of 128, Adam optimizer, and dropout regularization. The hyperparameters of the models (e.g., learning rate, dropout rate, number of neurons for each layer, etc.) were tuned by Bayesian optimization method. Appendix C includes the details on model training, structure and optimization

4.3.6 Statistical Analysis

To check the statistically significant difference of features' values between SBI and non-SBI groups, Wilcoxon rank-sum test for numerical features and Chi-squared test for categorical features were applied. Moreover, the estimated 95% confidence interval of the models' performance metrics were calculated through bootstrapping method.

This manuscript was prepared using the guidelines provided by Leisman et al. [53] for reporting of prediction models.

4.4 Results

For this study, 97,424 patient encounters associated with 15,704 patients were extracted from the EHR. Among these patient encounters, there were outpatient appointments and hospital outpatient department visits for patients with existing CVLs; therefore, 73,073 patients encounters were excluded from the cohort due to length-of-stay less than 24 hours or not having recorded clinical notes. After applying these exclusion criteria, a total number of 2,733 neonates (age less than 28 days), 5,383 infants (age between 28 days and one year), 4,286 toddlers and preschoolers (age between one and five years), 5,625 children (age between five and 12 years), and 6,324 adolescents (older than 12 years) were included in the analysis. Figure 4.3 demonstrates the associated CONSORT diagram.

The demographic and clinical characteristics of the patients in our study are listed in Table 4.1. SBI patients were younger (median = 3.1 years vs. 5.8 years, $p < 0.001$), with lower weights (13.6 Kg vs. 19.3 Kg, $p < 0.001$), shorter heights (90.3 cm vs. 110.2 cm, $p < 0.001$), more African Americans (44.3% vs. 36.3%, $p < 0.001$), and less Caucasian (47.1% vs. 54.1%, $p = 0.002$). Overall, SBI patients had higher hospital length-of-stay (36.7 days vs. 6.1 days, $p < 0.001$), had higher ICU admissions (65.4% vs. 48.2%, $p < 0.001$), $p < 0.001$), and had a higher rate of having Medicaid health

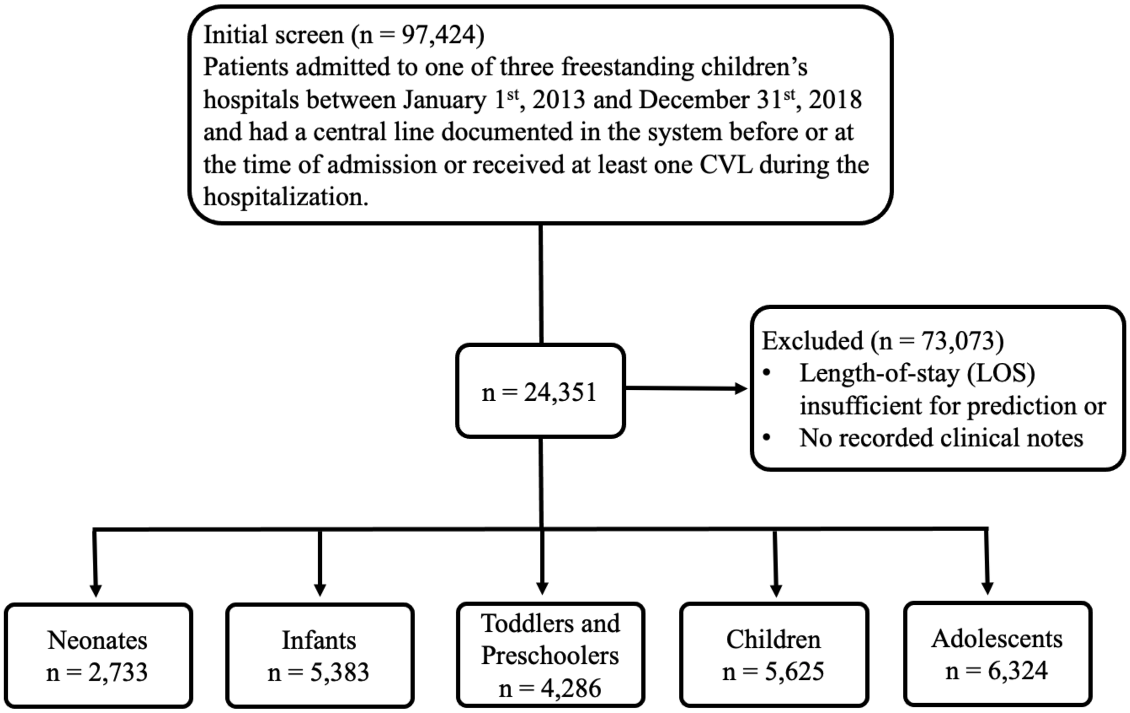

Initial screen (n = 97,424)
Patients admitted to one of three freestanding children's hospitals between January 1st, 2013 and December 31st, 2018 and had a central line documented in the system before or at the time of admission or received at least one CVL during the hospitalization.
n = 24,351
Excluded (n = 73,073)
• Length-of-stay (LOS) insufficient for prediction or
• No recorded clinical notes
Neonates n = 2,733
Infants n = 5,383
Toddlers and Preschoolers n = 4,286
Children n = 5,625
Adolescents n = 6,324

Figure 4.3: Patient Encounter Inclusion Flowchart. The final number of patient visits that were employed in training and testing the machine learning models was 24,351. "CC BY 4.0"

insurance (62.5% vs. 57.1%, $p = 0.02$) while having a lower rate in Commercial health insurance (34.2% vs. 38.9%, $p = 0.03$). The mortality rate was higher among SBI patients but there were no statistically significant differences for this feature among the two groups (0.2% vs. 0.06%, $p = 0.22$). Other features were comparable between SBI and non-SBI groups ($p > 0.05$).

Statistical tests were performed to assess statistical significance between the components of PEdiatric Logistic Organ Dysfunction (PELOD-2) score and Pediatric RISk of Mortality (PRISM-III) Score between SBI and non-SBI groups across time windows [54, 78]. PELOD-2 was primarily designed to describe the severity of organ dysfunction and PRISM-III was developed to predict the risk of mortality among the pediatric population. The results are presented in Appendix C.

To assess the benefit of incorporating the clinical notes in SBI prediction, five predictive models with different inputs were trained. The performance metrics are

Table 4.1: Cohort characteristics

	SBI	non-SBI	p-value
Age (years) (Median [25th, 75th])	3.1 [0.2, 12.1]	5.8 [0.8, 13.3]	< 0.001
Weight (Kg) (Median [25th, 75th])	13.6 [3.6, 38.6]	19.3 [7.7, 44.6]	< 0.001
Height (cm) (Median [25th, 75th])	90.3 [51, 149]	110.2 [66, 152]	< 0.001
Length of Stay (LOS) (Median [25th, 75th])	36.7 [23.2, 71.3]	6.1 [3.3, 14]	< 0.001
Gender			
Male (%)	45.3	45.8	0.82
Race			
Asian (%)	3.5	3.9	0.64
Caucasian (%)	47.1	54.1	0.002
African American (%)	44.3	36.3	< 0.001
American Indian or Alaska Native (%)	0.4	0.2	0.27
Native Hawaiian or Pacific Islander (%)	0.2	0.2	0.97
Other (%)	4.5	5.3	0.44
Insurance Status			
Commercial (%)	34.2	38.9	0.03
Public - Medicaid (%)	62.5	57.1	0.02
Public - non-Medicaid (%)	3.1	3.1	0.93
Self-pay (%)	0.2	0.8	0.12
ICU Admission (%)	65.4	48.2	< 0.001
Mortality (%)	0.2	0.06	0.22

presented in Table 4.2. The model which coupled the structured clinical features with word embeddings from Clinical BERT model outperformed the rest of the models with highest specificity of 0.981 with 95% CI = [0.980, 0.982], positive predictive value (PPV) of 0.113 [0.09, 0.137], negative predictive value (NPV) of 0.999 [0.999, 0.999], accuracy of 0.980 [0.978, 0.981], F-1 score of 0.195 [0.159, 0.231] and area under the precision-recall curve (AUPRC) of 0.282 [0.188, 0.366]. Figures 4.4 and 4.5 demonstrate the associated receiver operative characteristics and precision-recall curves for all the five models applied on the testing dataset, respectively.

Using the word representation from the last four hidden layers of the Clinical BERT model (instead of only using the last hidden layer) did not improve the models' performance while it added to the computational costs.

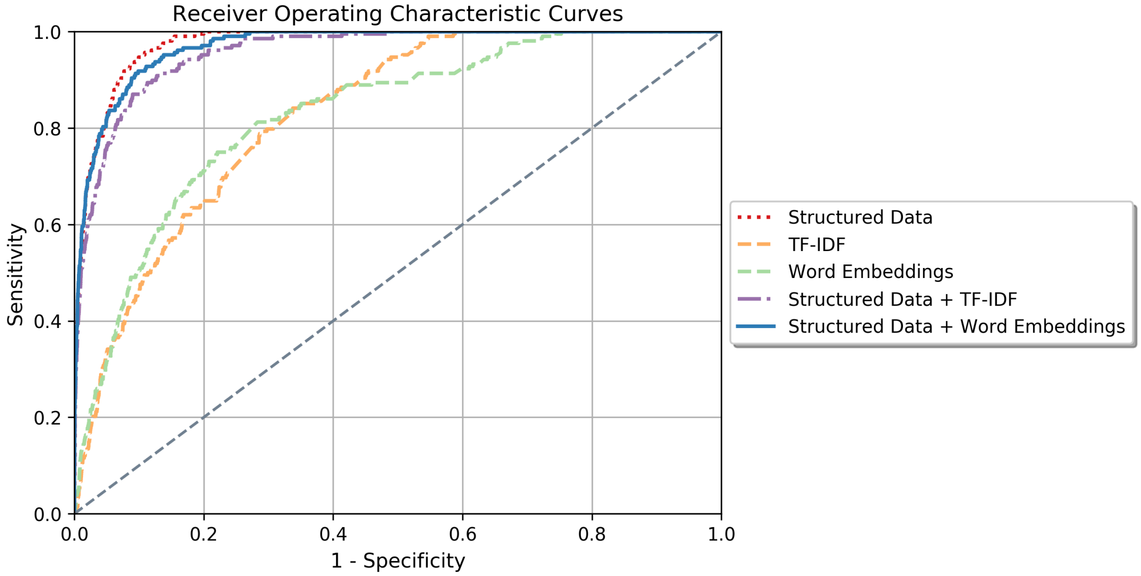


Figure 4.4: Receiver Operating Characteristics curves (ROC curves) for all the models tested in this study. Incorporating the structured information in EHR achieved the highest area under the curve. “CC BY 4.0”

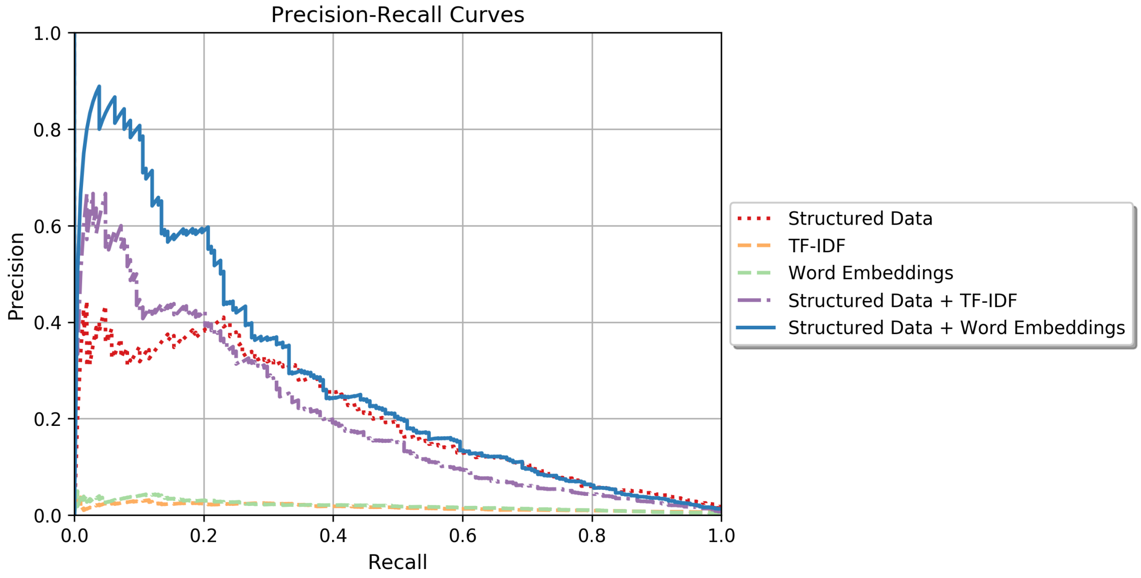


Figure 4.5: Precision-Recall curves (PRC curves) for all the models tested in this study. Incorporating the structured information in EHR integrated with the contextualized word embeddings from the Clinical BERT model achieved the highest area under the curve. “CC BY 4.0”

Table 4.2: Performance metrics of the deep learning models in predicting SBI in the next 48 hours of hospitalization. The two numbers in the brackets present the estimated 95% confidence interval using bootstrap sampling.

	Numerical Data		TF-IDF		Word Embeddings		Numerical Data + TF-IDF		Numerical Data + Word Embeddings	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
AUROC Mean (%) [95%CI]	0.979 [0.975, 0.982]	0.975 [0.969, 0.982]	0.848 [0.834, 0.858]	0.829 [0.788, 0.858]	0.859 [0.841, 0.874]	0.830 [0.784, 0.867]	0.989 [0.988, 0.992]	0.958 [0.944, 0.970]	0.989 [0.986, 0.991]	0.970 [0.961, 0.979]
Sensitivity Mean (%) [95%CI]	0.850 [0.817, 0.881]	0.814 [0.744, 0.883]	0.850 [0.810, 0.879]	0.841 [0.767, 0.908]	0.850 [0.818, 0.878]	0.821 [0.750, 0.890]	0.850 [0.828, 0.880]	0.586 [0.511, 0.671]	0.850 [0.819, 0.884]	0.673 [0.589, 0.751]
Specificity Mean (%) [95%CI]	0.951 [0.950, 0.952]	0.951 [0.948, 0.953]	0.668 [0.664, 0.672]	0.662 [0.648, 0.669]	0.698 [0.694, 0.701]	0.693 [0.685, 0.699]	0.982 [0.982, 0.983]	0.981 [0.979, 0.982]	0.982 [0.981, 0.983]	0.981 [0.980, 0.982]
PPV Mean (%) [95%CI]	0.058 [0.051, 0.064]	0.056 [0.044, 0.068]	0.009 [0.008, 0.010]	0.009 [0.007, 0.011]	0.010 [0.009, 0.011]	0.010 [0.007, 0.012]	0.145 [0.130, 0.157]	0.099 [0.082, 0.123]	0.142 [0.128, 0.157]	0.113 [0.09, 0.137]
NPV Mean (%) [95%CI]	0.999 [0.999, 0.999]	0.999 [0.999, 0.999]	0.999 [0.999, 0.999]	0.999 [0.999, 0.999]	0.999 [0.999, 0.999]	0.999 [0.999, 0.999]	0.999 [0.999, 0.999]	0.998 [0.998, 0.999]	0.999 [0.999, 0.999]	0.999 [0.999, 0.999]
Accuracy Mean (%) [95%CI]	0.951 [0.950, 0.952]	0.950 [0.948, 0.953]	0.669 [0.665, 0.673]	0.662 [0.649, 0.669]	0.698 [0.695, 0.702]	0.693 [0.686, 0.700]	0.982 [0.981, 0.982]	0.979 [0.978, 0.980]	0.982 [0.981, 0.982]	0.980 [0.978, 0.981]
F-1 Score Mean (%) [95%CI]	0.108 [0.097, 0.119]	0.105 [0.084, 0.126]	0.018 [0.016, 0.020]	0.018 [0.014, 0.022]	0.019 [0.017, 0.021]	0.019 [0.015, 0.023]	0.248 [0.226, 0.265]	0.169 [0.142, 0.203]	0.243 [0.222, 0.266]	0.195 [0.159, 0.231]
AUPRC Mean (%) [95%CI]	0.314 [0.266, 0.359]	0.202 [0.142, 0.279]	0.019 [0.017, 0.022]	0.016 [0.012, 0.024]	0.043 [0.030, 0.058]	0.020 [0.013, 0.028]	0.556 [0.523, 0.596]	0.201 [0.145, 0.250]	0.646 [0.606, 0.693]	0.282 [0.188, 0.366]

4.5 Discussion

In this study, we evaluated the effect of coupling clinical notes with the structured clinical features (e.g., demographic, physiological, and laboratory test results, etc.) in predicting the onset of SBI, defined as a culture drawn associated with a positive test result followed by at least four days of new antimicrobial agent administration, among pediatric patients with CVLs. The proposed deep learning model predicts if a hospitalized patient with CVL will develop SBI during the next 48 hours of hospitalization. Our model had a PPV of 0.113, which is 32 times greater the baseline prevalence of SBI across the 48-hour time windows, and a very high NPV 0.999 which presents the strength of the model in ruling out the patients with lower risk of the infection. Incorporating the clinical notes improved the specificity (0.951 vs. 0.981, $p < 0.001$), PPV (0.056 vs. 0.113, $p < 0.001$), accuracy (0.950 vs. 0.980, $p < 0.001$), F-1 score (0.105 vs. 0.195, $p < 0.001$) and AUPRC (0.202 vs. 0.282, $p < 0.001$) compared to the model which employed only the information from structured EHRs. However, incorporating clinical notes did result in a reduction in sensitivity in the test data set from 0.81 when using structured data only down to 0.67 when adding word embeddings. Because SBI is a rare event in this window-wise data set, both models nonetheless maintain an NPV of 0.999.

The predictive performance of our proposed model and study design outperformed the performance of the prior models trained to predict CLABSI. Most of these models were based on a retrospective case control study and did not incorporate the temporal information in EHRs. Training a Random Forest predictive model based on non-temporal data, Beeler et al. obtained AUROC of 0.87 in predicting CLABSI among adult, pediatric and neonatal patients [7]. Sung et al. trained a CLABSI prediction model using Gradient Boosting Trees which attained AUROC of 0.77 among pediatric cohorts receiving cancer medications [95]. Our model and study design had characteristics that may have contributed to achieving better predictive performance; first, we only included the pediatric patients with a documented CVL at the time of admission or at some point during hospitalization. Second, we extracted and incorporated an extensive set of features recorded in EHR. Third, we used the information embedded in the clinical notes through a state-of-the-art NLP framework. Finally, we trained a deep learning model capable of including the temporal information while dealing with low prevalence classification problem by using a loss function specifically designed for extreme class-imbalanced classification and attention mechanism to focus on the most predictive parts of the input sequence at every prediction point.

Previous studies have been done to investigate the added benefit of integrating clinical notes to the structured EHR data in predicting patient clinical outcomes. Amrollahi et al. utilized structured and unstructured EHR information to model capable of timely prediction of sepsis which outperformed the model trained only on the structured data [3]. In a similar study, Goh et al. developed an artificial intelligence algorithm incorporating the two data modalities and concluded that the model performance improved after integrating the clinical notes to the input features [36]. In another study, Horng et al. conducted a research to demonstrate the incremental benefit of using free text data in addition to vital sign and demographic data to identify patients with suspected infection in the emergency department [41].

Limitations Our study has several limitations. First, while we intentionally extracted EHR features that are routinely recorded across systems, the external application of the proposed model, which was trained within a single pediatric health system, on other health systems may be biased. Second, the deep learning field is very dynamic and new models are introduced every day; therefore, the model structure that we applied to tackle low prevalence classification problems and extract contextualized word embeddings may not reflect all the capacity of deep learning application in predicting this adverse outcome. Third, in some cases, the clinical notes are updated with delay. This delay in recording clinical notes may affect the performance of the model. Finally, the Clinical BERT model used for extracting contextualized word embeddings has a limitation in the number of words it can take in each of the recorded clinical notes at every prediction point. This limitation requires extensive text preprocessing to only include the parts describing the patient's health condition at the moment and exclude history and administrative sections.

Chapter 5

A Novel Technique for Developing a Natural Language Processing Algorithm to Identify Intimate Partner Violence in a Hospital Setting

5.1 Abstract

Importance: Hospital settings need a sensitive screening tool to identify intimate partner violence cases among the visits to the emergency department.

Objective: To develop an algorithm using natural language processing to identify cases of intimate partner violence among emergency department encounters.

Design: Observational cohort study.

Setting: Unstructured clinical provider, nursing and social worker notes were extracted from hospital electronic health records. The recorded clinical notes and patient narratives were screened for a set of 23 situational terms, derived from the literature on intimate partner violence, along with an additional set of 49 extended situational terms, extracted from known intimate partner violence cases. We compared the effectiveness of the proposed model with the capability of ICD-9/10 codes in detecting such cases.

Participants: In total, 1,064,735 patient encounters (405,303 patients) who visited the emergency department of a level one trauma hospital from January 2012 to August 2020 were included in the analysis.

Main Outcomes and Measures: The outcome was identification of an intimate partner violence-related encounter.

Results: In this study, we utilized the information embedded in unstructured electronic health records to develop a natural language processing algorithm that employs the recorded clinical notes to identify the intimate partner violence visits to the emergency department. Employing a set of 23 situational terms along with 49 extended situational terms, the algorithm successfully identifiedy 7,399 intimate partner violence-related visits representing 5,975 patients; the algorithm achieved 99.5% precision in detecting positive cases.

Conclusions and Relevance: Using a set of pre-defined intimate partner violence-related situational terms, we successfully developed a novel natural language processing algorithm using unstructured data capable of identifying intimate partner violence visits with high precision.

5.2 Introduction

Intimate partner violence (IPV) is defined as sexual, physical, psychological, or economic violence that occurs between current or former intimate partners [16]. Nearly 30% of women globally have experienced IPV making it a serious public health concern [74]. IPV is a significant contributor to violence related injury and a leading cause of femicide, the intentional killing of women based solely on their gender [70]. Although men may experience IPV, women are disproportionately affected [17]. In the United States (U.S.) one in four women and one in nine men have experienced a severe form of IPV at some point in their lifetime [105].

Individuals who experience IPV experience both short and long-term adverse health outcomes such as chronic pain, substance use, and mental health disorders [25, 12, 102, 29]. Victims of IPV may seek care for IPV related injuries in health care settings which makes recognition and intervention in these facilities critical [81, 30]. Yet, IPV is profoundly underdiagnosed in healthcare settings, limiting identification and response efforts. A number of screening tools have been successfully developed to detect IPV, however, screening tools are not yet universally implemented [14, 101]. Emerging efforts have focused on using machine learning to aid in detection of conditions including non-accidental trauma and IPV [4, 8, 76]. Khurana et al. proposed a machine learning algorithm that utilizes radiologic findings of high-risk injuries (ex injury location and patterns specific to IPV) to identify patients who are at high risk of IPV [49]. Using the 2016 South African Demographic and Health Survey dataset, Amusa et al. developed a machine learning model using country-specific self-reported survey data to capture common characteristics contributing to IPV [4].

Information captured in the electronic health records (EHRs) including, but not limited to, clinical notes, radiology reports, and imaging tests has been widely used to predict adverse outcomes for specific medical conditions. These data have also been used to a limited extent to detect IPV based on radiologic findings IPV [21]. In this study, we propose a novel natural language processing (NLP)-based algorithm using data embedded in EHR notes to detect Emergency Department (ED) encounters for IPV.

5.3 Methods

5.3.1 Study Population

We extracted EHR data for all ED encounters at an urban U.S. based level-one trauma center from January 2012 to August 2020. These unstructured data included chief complaint alongside provider, nursing and social worker notes (ED triage, ED notes, ED provider, ED progress, history and physical, consults, assessment and plan, addendum, miscellaneous); structured data including International Classification of Diseases (ICD-9 or ICD-10) codes, procedure and billing codes, admission diagnosis, disposition, patient status, and date of birth were also extracted. This study was conducted according to the Emory University Institutional Review Board Study Protocol 00000432.

5.3.2 Detecting IPV Cases

We attempted to use the structured data to identify IPV-related visits to the ED; these attempts were followed by use of the unstructured data to develop our NLP-based algorithm for identifying the IPV-related visits to the ED. Here we describe our three iterative approaches to identify IPV-related visits. Figure 5.1 demonstrates a summary of the different approaches in this analysis.

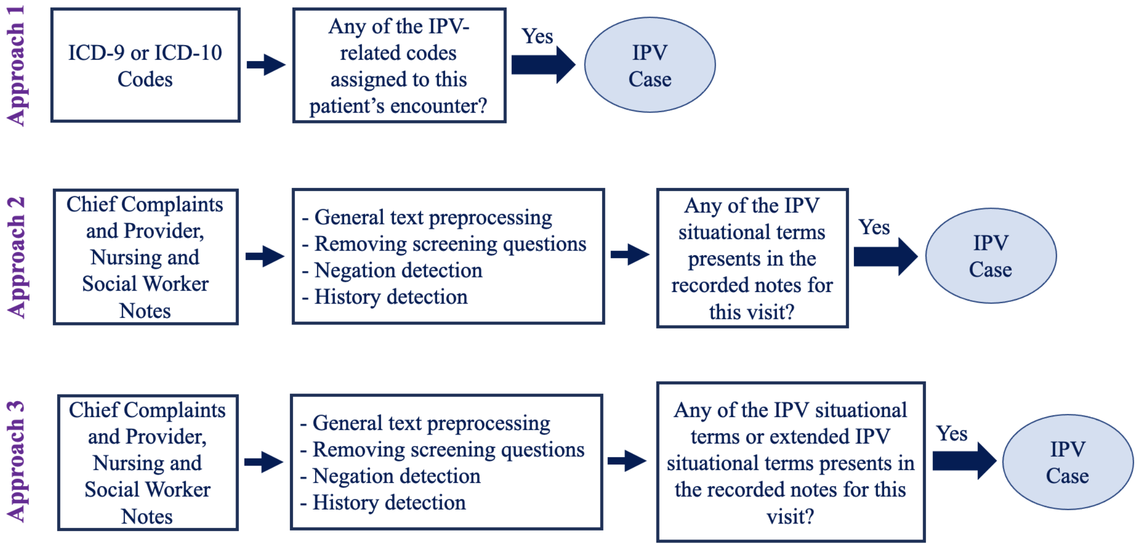

Approach 1
ICD-9 or ICD-10 Codes
Any of the IPV-related codes assigned to this patient's encounter?
Yes
IPV Case
Approach 2
Chief Complaints and Provider, Nursing and Social Worker Notes
- General text preprocessing
- Removing screening questions
- Negation detection
- History detection
Any of the IPV situational terms presents in the recorded notes for this visit?
Yes
IPV Case
Approach 3
Chief Complaints and Provider, Nursing and Social Worker Notes
- General text preprocessing
- Removing screening questions
- Negation detection
- History detection
Any of the IPV situational terms or extended IPV situational terms presents in the recorded notes for this visit?
Yes
IPV Case

Figure 5.1: Three methods for developing a natural language processing algorithm to identify intimate partner violence in a hospital setting. "CC BY 4.0"

Approach 1: ICD-9/ICD-10 Codes

In the first approach, IPV-related ICD-9 (2012-September 2015) and ICD-10 (October 2015-August 2021) codes were identified by the study team (Table 5.1). In our analysis, if at least one of the ICD-9/ICD-10 codes appeared in an encounter, the visits were identified as a case of IPV.

Table 5.1: ICD-9 and ICD-10 used to identify cases of intimate partner violence.

ICD-9 Codes		ICD-10 Codes	
Code	Diagnosis (Dx) Name	Code	Diagnosis (Dx) Name
995.83	Adult sexual abuse	T76.21XA	Adult sexual abuse, suspected, initial encounter
995.83	Adult rape	T76.51XA	Adult forced sexual exploitation, suspected, initial encounter
995.82	Adult emotional abuse	T76.11XA	Adult physical abuse, suspected, initial encounter
995.81	Adult physical abuse	T74.11XA	Adult physical abuse, confirmed, initial encounter
995.8	Adult abuse	T74.21XA	Adult sexual abuse, confirmed, initial encounter
E967.0	Perpetrator	T74.51XA	Adult forced sexual exploitation, confirmed, initial encounter
E967.9	Perpetrator	T71.9XXA	Asphyxiation due to unspecified cause, initial encounter
994.7	Asphyxiation and strangulation	T71.163A	Asphyxiation due to hanging, assault, initial encounter
-	-	T71.193A	Asphyxiation due to mechanical threat to breathing due to other causes, assault, initial encounter

Approach 2: IPV Situational Terms

IPV is stigmatized and often undisclosed by those experiencing it; health care providers may also have varying levels of awareness and comfort in dealing with IPV. As a result ICD-9/ICD-10 codes are inconsistently used and frequently undercoded. Therefore, additional IPV-related situational terms were utilized to identify patients experiencing IPV. In our second approach, IPV situational terms were derived from existing IPV literature, including validated terms from IPV risk assessment instruments and from clinician expertise [33, 13, 94]. If any of the situational terms were captured in a clinical note, the visit was classified as IPV. The 23 IPV situational terms included: *"domestic violence, intimate partner violence, spouse abuse, battered woman, domestic abuse, spousal abuse, intimate partner abuse, battered, violence against women, domestic assault, domestic dispute, problems with spouse or partner, maltreatment by spouse or partner, neglect and abandonment by spouse or partner, assault by husband, assault by partner, assault by wife, assault by spouse, assault by boyfriend, assault by girlfriend, assault by significant other, referral to partnership against domestic violence, resources or shelter for domestic violence."*

Approach 3: IPV Extended Situational Terms

Finally, additional IPV-related terms were identified through review of notes from confirmed IPV encounters. These extended terms included specific descriptions of various forms of physical abuse (i.e. attack, strike, strangle) also derived from the literature [74, 33, 13, 94]. If any of the situational or extended situational terms were captured in a clinical note, the visit was classified as IPV. The 49 IPV extended situational terms included: *"intimate partner homicide, femicide, intimate partner death, spousal homicide, ipv, dv, domestic violence resources, assault by so, assault by domestic partner, assault by ex, assault by bf, assault by gf, strangle by boyfriend, strangle by girlfriend, strangle by wife, strangle by husband, strangle by spouse, strangle by domestic partner, strangle by partner, strangle by significant other, strangle by so, strangle by ex, strangle by bf, strangle by gf, strike by boyfriend, strike by girlfriend, strike by wife, strike by husband, strike by spouse, strike by domestic partner, strike by partner, strike by significant other, strike by so, strike by ex, strike by bf, strike by* *gf, attack by boyfriend, attack by girlfriend, attack by wife, attack by husband, attack by spouse, attack by domestic partner, attack by partner, attack by significant other, attack by so, attack by ex, attack by bf, attack by gf, violence against women."*

5.3.3 Data Preprocessing

Members of the study team completed a manual review of charts identified as positive IPV cases in real-time relative to each approach. During the application of approaches 2 and 3, a number of text based scenarios identified in unstructured clinical notes led to false positive IPV cases. As a result, several data preprocessing steps were required to prepare the data prior to application of the algorithm. These include general and task-specific text preprocessing steps along with negation and history detection.

General and Task-Specific Preprocessing

The following text-based scenarios led to false positives: 1) auto-populated IPV screening questions (whether completed or blank) and 2) auto-populated past medical, obstetric or psychiatric history reflecting a history of IPV unrelated to the identified encounter. As a result, task-specific text preprocessing was required for these scenarios. Prepositions and time indications were removed from the text to make clinical notes consistent. For example, "assaulted last night by her husband" was changed to "assault by husband." Additionally, general text preprocessing steps were performed including transforming all text to lowercase, and removing numbers, extra white spaces and words with less than two characters.

Negation Detection

Encounters in which the patient denied a history of IPV were incorrectly labeled as IPV given the inclusion of IPV terminology. To omit these false positives, we applied a negation detection algorithm which is a simplified version of NegEx software [18]. In this approach, negation words and terminating tokens are defined. When a negation word is detected, any word between the negation word and the next terminating token is negated. For example, if the text includes: “Patient denies drug, alcohol use and intimate partner violence”, *denies* is the negation word and *period* is the termination token. Therefore, applying the negation detection algorithm results in “Patient denies drug_neg, alcohol_neg use_neg and_neg intimate_neg partner_neg violence_neg.” As a result, such cases were excluded from situational and extended IPV terms and thus not labeled as IPV. Table 5.2 includes a list of negation words as well as termination tokens in our analysis designed according to the literature [41].

Table 5.2: Negation words and terminations tokens for a natural language processing algorithm to identify cases of intimate partner violence in a hospital setting.

Negation Words	Termination Tokens
”denies”, ”denied”, ”deny”, ”no”, ”non”, ”not”, ”without”, ”unable”	”?”, ”.”, ”-”, ”;”, ”:”, ”+”, ”and”, ”but”, ”complains”, ”did”, ”except”, ”has”, ”per”, ”pt”, ”reports”, ”secondary”, ”states”

History Detection

The algorithm initially detected encounters in which a patient had a history of IPV as described in the text (separate from the auto-populated history). Following a similar approach as in negation detection, encounters with a history of IPV included in the text were not labeled as IPV as this was not the reason for the ED encounter. IPV history detection tokens are listed in Table 5.3. For example, “Patient reports a history of IPV during previous pregnancy but not currently” was not labeled as IPV. Punctuations were removed at the end of this step.

Table 5.3: History words and termination tokens for a natural language processing algorithm to identify cases of intimate partner violence in a hospital setting.

History Words	Termination Tokens
“history of”, “hx of”, “h/x of”, “ho of”, “h/o of”, “hx”, “h/x”, “h/o”, “ho”	”?”, ”.”, ”-”, ”;”, ”:”, ”+”, ”and”, ”but”, ”complains”, ”did”, ”except”, ”has”, ”per”, ”pt”, ”reports”, ”secondary”, ”states”

5.3.4 NLP Algorithm Application

To validate the performance of the proposed NLP algorithm, manual chart reviews were conducted for the identified IPV cases. Manual review was conducted for nearly 25% of the identified IPV cases.

5.4 Results

During the study period (January 2012 - August 2020) there were 1,064,735 ED encounters (405,303 patients). To identify IPV visits, all ICD-9 and ICD-10 codes and data from unstructured notes were used to investigate the performance of the three aforementioned approaches.

5.4.1 Approach 1: ICD-9/ICD-10 Codes

The first approach using only ICD-9 and ICD-10 codes resulted in the identification 1,404 IPV visits representing 1,299 patients over a nine year time period. Based on clinician expertise and anecdotal experiences at the hospital site, this number of cases was significantly lower than expected given the duration of time. Additionally, during manual review of these cases, a number of encounters were found to be unrelated to IPV. Some cases were indicative of elder abuse, reflecting the inaccuracy of relying exclusively on ICD-9 and ICD-10 codes as these codes are often used inconsistently or inappropriately. Notably, during the manual review of positive IPV cases identified through approach 3 and confirmed by manual review, a number of true IPV encounters did not have an associated IPV ICD-9 or ICD-10 code verifying that these codes are under- or inappropriately utilized.

5.4.2 Approaches 2 and 3: IPV Situational Terms Extended Situational Terms

In our next approach, we defined a set of 23 IPV-related situational terms. If any of these terms appeared in a visit's recorded clinical notes, the visit was labeled as IPV. The second approach using IPV situational terms yielded 6,437 IPV visits reflecting 5,280 patients. Building on this approach, we added more mechanism-related terminology (i.e. attack, strike, strangle) to the set of the 23 terms referred to as IPV extended situational terms. The third approach using IPV extended situational terms identified 7,399 IPV-related visits representing 5,975 patients. Relative to the use of ICD codes, these approaches had significantly improved accuracy in identifying true IPV cases, with extended situational terms identifying more positive IPV cases without a notable difference in identifying false positives. For Approach 3, manual chart reviews were conducted for a random subset of 1,798 (25%) identified cases to validate the results of the algorithm. Considering the 1,798 identified cases, 1,790 (99.5%) were confirmed IPV visits, only five (0.3%) reported a history of IPV or domestic violence, two (0.1%) were incorrectly labeled as IPV, and there was a concern of IPV for only one (0.1%) visit.

Figure 5.2 presents the result of applying the three approaches, ICD-9/ICD-10 codes, IPV situational terms and IPV extended situational terms, to identify IPV cases.

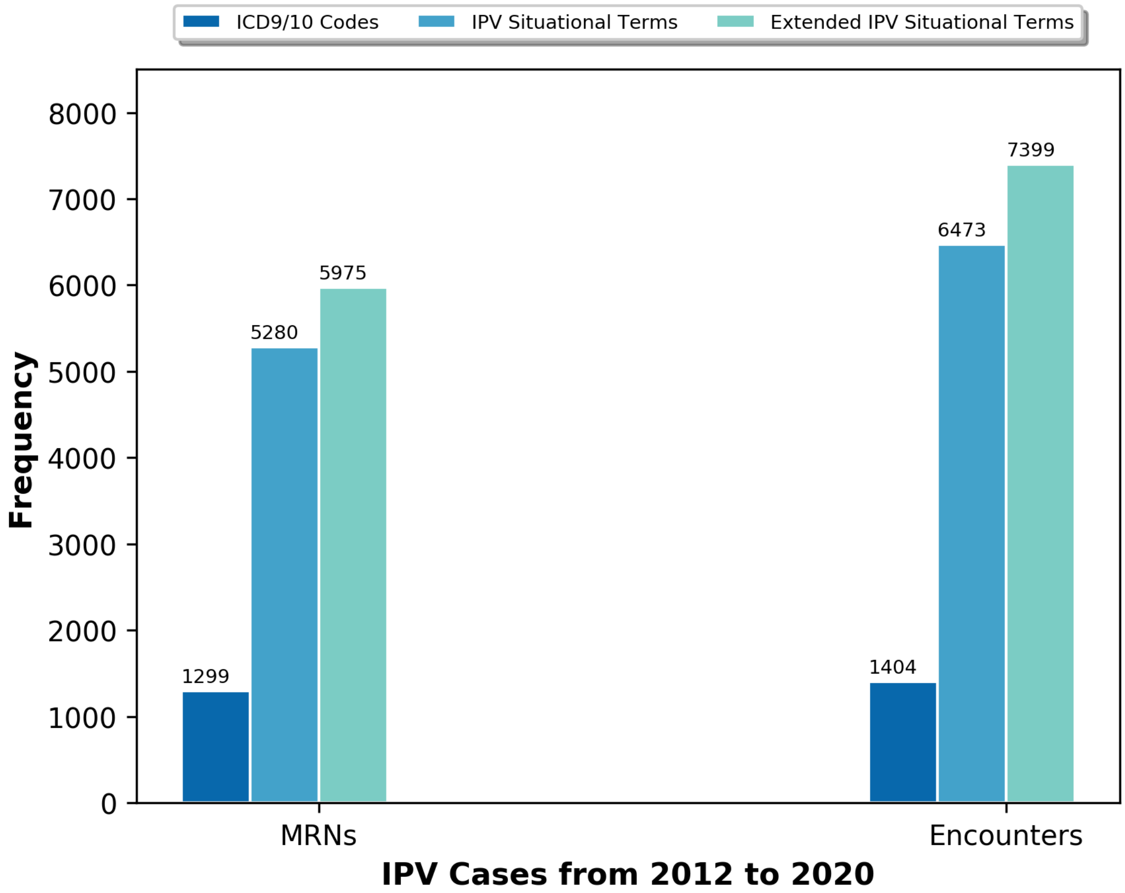


Figure 5.2: The number of identified IPV cases using the three approaches. “CC BY 4.0”

5.5 Discussion

We successfully developed a novel NLP algorithm using EHR data to identify IPV encounters in a hospital setting. We explored three different NLP approaches using: 1) ICD-9/ICD-10 codes, 2) a set of 23 IPV-related situational terms, and 3) a set of 49 IPV-related extended situational terms. Among the three approaches incorporated in this study, the use of ICD-9/ICD-10 codes alone identified the fewest IPV visits over a 9-year time interval. IPV visits were significantly under-coded and in some cases, IPV-related codes were used for non-IPV related visits. This approach is not sufficient for the accurate and meaningful identification of IPV-related visits. The second and third approaches using unstructured EHR data appropriately identified a larger number of IPV encounters. While the second approach led to identifying more IPV cases, we used expert knowledge to expand the IPV situational terms and added the extended situational terms to capture additional IPV cases among the visits to ED. As a result, the third approach using extended situational terms generated the largest number of true IPV visits achieving a 99.5% precision.

In a study conducted by Chen et al., the authors generated an NLP predictive algorithm using radiology reports from confirmed IPV cases [21]. IPV labels were identified using IPV injury patterns and predictive words from radiologic findings. This study differed from ours in that it relied only on radiologic findings to develop an algorithm rather than clinical notes. The information obtained in clinical notes provides greater context and IPV specific terminology, and is more inclusive of individuals who may not undergo radiologic imaging. Thus, our algorithm may be able to detect more cases by utilizing a more expansive source of clinian information. Similar to our study, Blosniche et al. used clinical notes to identify transgender related terminology in order to better identify transgender patients [10]. The methodology differed in that they first used transgender based ICD codes to identify patients and then used clinical notes from these encounters to identify transgender-related terms. However, the study similarly demonstrates that clinician notes can be an important source of data for labeling encounters that are otherwise difficult to identify or are socially stigmatized. It should also be noted that the purpose of the study was different in that it sought to identify a population (transgender patients) rather than a condition or experience (IPV).

Unstructured EHR data with free-text formatting provides a rich source of information related to the circumstances of medical visits and related health sequelae. The data provided in clinical notes can be an important source of information with which to identify the social and contextual factors surrounding IPV-related visits, as well as providing an opportunity to appropriately identify IPV visits. The main challenge in using this type of data is the unstructured nature of notes, which makes extracting information a complicated task. As a result, we applied extensive preprocessing steps to ready these data for the screening process. Sequentially building our algorithm grounded first in ICD codes, and then complemented by both situational and extended terms enabled greater specificity in identifying IPV cases when compared to the use of ICD codes alone; the search and use of relevant terms in clinical notes was key to the success of this approach. Future efforts to improve our algorithm could incorporate active learning to identify a greater number of IPV encounters [109]. This method is a process of prioritizing the data which needs to be labeled in order to improve the overall performance of a predictive model.

People experiencing IPV often seek care in hospital settings. Therefore the early and appropriate detection of and response to such cases is critical in disrupting the cycle of abuse including IPV related morbidity and mortality. The novel NLP-based algorithm described in this manuscript is an innovative tool to identify victims of IPV with accuracy. The algorithm can be utilized in health care settings to identify victims of IPV for surveillance and intervention purposes. For example, the extent to which COVID-19 has impacted IPV related health seeking behaviors in the U.S. is still largely unknown. As identification of IPV in health systems is challenging, application of this algorithm could assist with understanding the impact of movement related restrictions during the COVID-19 pandemic on IPV related encounters. Additionally, this algorithm could be used to develop predictive modeling allowing for the detection of those at risk of IPV. Early detection during hospital encounters could aid in novel injury prevention strategies, ensuring those at risk have access to support and social services.

Limitations: Limitations: This study has limitations. First, the NLP algorithm utilizes the recorded clinical notes and patient narratives; therefore, the model cannot detect IPV cases if the patient or health provider did not mention or document any of the IPV-related terms. Second, we applied extensive text preprocessing before moving on to search for IPV situational terms. However, if a patient or provider stated the history of IPV in a way that is not captured by our history detection algorithm, the proposed NLP algorithm would identify that case as IPV. Third, the set of IPV terms that we incorporated are limited. If a patient uses terminology outside the set of pre-defined IPV situational terms, the algorithm will not identify the encounter. On the other hand, some terms may be used in a non-IPV context. For example, domestic dispute can be used in IPV encounters but can also refer to a conflict among members of a family (e.g., mother and child) and generate false positives. While our final approach demonstrated superiority over and above the use of ICD codes alone or the use of situational terms it admittedly still missed some cases. As conversations about the use of NLP and other technologies continue, debate over what degree of sensitivity is reasonable for a model such as ours is warranted. From our perspective, missing any cases is unacceptable. Any future models should endeavor for even greater sensitivity and precision to ensure that opportunities to interrupt IPV are not missed.

Chapter 6

Conclusion

6.1 Summary and Contributions

The work presented in this thesis addressed the issue of extreme class-imbalanced data classification in hospital settings. Theoretically, this thesis focuses on improving the predictive performance of the models trained on complex data to predict rare events in a dynamic environment.

In particular, the work in this thesis focuses on two cases of rare event classification in the healthcare domain: 1) predicting serious infection among hospitalized children with CVLs and 2) identifying IPV cases among the visits to the ED of a hospital system.

6.1.1 Serious Infection Prediction among Hospitalized Children

Hospitalized children with CVLs are at high risk of morbidity and mortality from hospital-acquired infections (HAI), including CLABSIs. Definitions have been proposed for CLABSI in the pediatric domain, but they commonly have inadequate sensitivity for clinically important infections and may be difficult to generalize across

EHR platforms [51, 5]. Many serious infections in hospitalized children are likely preventable through interventions that prevent them or identify them early to initiate antimicrobial therapy. On the other hand, excessive use of antimicrobials can lead to adverse events and worsening antimicrobial resistance; therefore, training predictive models which can help the health providers in identifying patients at the highest risk for this type of infection can help clinicians better achieve the balance between early intervention and antimicrobial overuse. In Chapters 2 to 4, we proposed and validated a surrogate definition for CLABSI, which can be inferred from EHR information and eliminates the need for extensive chart reviews. Then, we trained predictive models to identify the onset of the serious infection, which is considered among the rare patient adverse outcomes.

In Chapter 2, we proposed a new definition for bloodstream infection among hospitalized children with CVLs. We defined this type of infection as at least one blood culture draw followed by at least four consecutive days (or fewer if the patient dies or is transferred out) of antimicrobial agents that were not administered in the week before the blood culture draw. This definition was adopted from a pediatric sepsis definition [42] which has been validated [107]. The positive labels generated by this definition were validated through patient chart reviews. We also designed a *lookback* study to evaluate the performance of an optimized tree-based ensemble model in predicting the onset of the infection during hospitalization time.

In Chapter 3, we designed a window-wise study to mimic how patients are monitored while hospitalized and identify the patients at higher risk of bloodstream infection at the beginning of each shift in ICU (every 8 hours). In this study design, we presented a predictive model that incorporated the EHR's temporal information to boost its performance. We further applied attention mechanism to help the model focus on the more informative parts of the information, and a Focal loss function which is designed for extreme class-imbalanced data classification.

We evaluated the added benefit of data fusion in predicting bloodstream infection in Chapter 4. We integrated the unstructured clinical notes recorded at the time of admission and throughout the hospitalization by employing a pre-trained NLP model to extract the information embedded in the free-text data [1]. Preserving the same model structure as Chapter 3, which contained attention layer and Focal loss, and coupling the two data modalities augmented the prediction task. The improvement after data fusion has been observed in other studies as well [3, 36, 41].

After achieving decent performance from the predictive model in 3, teams from Children's Healthcare of Atlanta (CHOA), Epic and the Biomedical Informatics Department at Emory University have been working towards implementing a near-real time, prospective surveillance system, as described in Chapter 3, in order to identify the patients at higher risk of infection at the beginning of each ICU shifts at CHOA.

6.1.2 Identifying IPV Cases among the Visits to ED

IPV is a pervasive social challenge with severe health and demographic consequences. People experiencing IPV may seek care in hospital settings. Despite the urgency of this critical public health issue, IPV continues to be profoundly under-diagnosed and is considered a persistent hidden epidemic. While IPV is rarely observed in a hospital setting, the early and appropriate detection of and response to such cases is critical in disrupting the cycle of abuse, including IPV-related morbidity and mortality.

In Chapter 5, we proposed an NLP-based algorithm to identify IPV cases among the visits to ED of a hospital. Our approach incorporated the information derived from the literature in the field (including validated terms from systematic reviews) along with the knowledge from clinician expertise to finalize a set of 23 IPV situational terms and 49 IPV extended situational terms to screen the recorded provider notes and patient narratives in ED for any signals of such a violent act. Given the severity of IPV, timely identification of individuals injured by IPV proves crucially important for providing support and social services for this vulnerable population. The novel NLP-based algorithm described in Chapter 5 is an innovative tool to identify victims of IPV with 99.5% sensitivity.

6.2 Limitations

The work presented in this thesis has some limitations. The methods presented in Chapters 2 to 5 were developed and tested offline. The methods were not tested online at the point of care because this work focused on the development of the functionalities rather than in the implementation of them. Nevertheless, the obtained results were designed and developed taking implementation feasibility into consideration.

Besides, the models presented in this work are based on extracted EHR information which brings the following challenges and limitations:

6.2.1 Serious Infection Prediction among Hospitalized Children

In Chapters 2 to 4, we employed data associated with a single pediatric health system which may reflect the particular structure and patient mix of this setting. Although we extracted the EHR features expected to be routinely recorded across the hospitals, the external application of our model on other health systems may be biased.

While simple to infer from the EHR data, using an easily computable definition for CLABSI that is not equivalent to the NHSN definition limited validation of the novel definition used.

Moreover, our models proposed in Chapters 3 and 4 were trained and evaluated based on a retrospective cohort. While we attempted to simulate prospective implementation using a window-wise design, predictive performance may deteriorate when implemented in real time.

6.2.2 Identifying IPV Cases among the Visits to ED

In Chapter 5, we proposed an NLP algorithm utilizing a set of pre-defined IPV situational terms and look for the signs of an IPV situation; therefore, the model cannot detect IPV cases if the patient or health provider did not mention or document any of the IPV-related terms.

While our approach achieved high sensitivity, it admittedly still missed some cases. We applied extensive text preprocessing before moving on to search for IPV situational terms. However, if a patient or provider stated the history of IPV in a way that is not captured by our history detection algorithm, the proposed NLP algorithm would identify that case as IPV.

Furthermore, the set of IPV terms that we incorporated are limited. So, if a patient uses terminology outside the set of pre-defined IPV situational terms, the algorithm will not be able to capture that case which leads to having false negatives.

6.3 Future work

This work helps support the development of generalizable approaches to address class-imbalanced classification problem in the clinical settings.

In Chapter 4, our results demonstrated that data fusion approach can work as a remedy to the class-imbalanced classification problem. Additionally, we will extend our methods on data fusion and incorporate other data modalities such as waveform data in subsequent iterations to improve the prediction outcomes.

Similarly for IPV case identification in Chapter 5, we will incorporate other inputs from patients such as recorded voice and facial expressions captured during the first patient-healthcare providers' encounter in ED to improve the accuracy of the IPV detection algorithm and prevent missing any cases.

As conversations about the use of NLP and other technologies continues debate over what degree of sensitivity is reasonable for a model such as ours is warranted. From our perspective, missing any cases is unacceptable. Given that, another future step of our work is to apply active learning to identify additional IPV cases among the visits to ED [109]. This method is a special case of machine learning in which a learning algorithm can interactively query a user to label new observations with the desired outputs.

The focus of the work presented on IPV case identification in Chapter 5 have been on assigning labels of IPV and non-IPV to the ED visits. After improving the accuracy of this algorithm, we will extend or work by training a predictive model that can utilize the labels learned from the retrospective data and act as a screening tool to detect IPV cases from the recorded health providers notes and patient narratives in ED. We will transfer the knwoledge learned in predicting serious infection in pediatric cohort to identify cases of IPV among adult population in ED and employ similar predictive model structure with attention mechanism and Focal loss. Any future models should endeavor for even greater sensitivity to ensure that opportunities to interrupt IPV are not missed.

Appendix A

A.1 Features

The complete list of the features extracted from CHOA database is mentioned in the following. To preserve the models' reproducibility, we selected the features that are available across all healthcare systems; therefore, we incorporated patient and encounter level, central line property and microbiology information, current medications, antibiotic administration, laboratory results and measured vital signs.

Features extracted from CHOA clinical database:

- Patient-Level data: Name, DOB, MRN, Sex, Race, Ethnicity, Zip Code, Gestational Age, Birth Weight, Primary Language.
- Encounter-Level data: CSN, Admission Date/Time, Discharge Date/Time, Department(s), All diagnoses prior to admission date, All Diagnoses prior to discharge date, Admission Diagnoses, Hospital Diagnoses, Flags (Oncology Diagnoses, BMT Diagnoses, Transplant Diagnoses, Short Gut Syndrome, NEC, DiGeorge, SCID, Downs, Heterotaxy), Insurance status, Admission Weight, Admission Height, Caregiver cognitive factors including ability to read
- ICU Admission-Level data: Time of transfer into ICU, Time of transfer out of ICU, Flags (PICU, CICU, NICU, Technology-dependent ICU)
- Co-morbidities variable: Presence and number of pediatric complex care condi-

tions (see R package or original citation)

- PRISM-3 at 1st PICU admission, PIM-2 at 1st PICU admission, SNAPPE-2 at NICU admission
- ECMO/Bypass variable: ECMO Start Time, ECMO End Time, Cardiac Bypass Start Time, Cardiac Bypass End Time
- Bone Marrow Transplant Level Data Variable: HCT Type, Transplant Date, Presence of Acute GVHD, Presence of Chronic GVHD
- Line Properties Variable: Line Insertion Date, Line Type, SITE Location, Gauge, Needle Length, Patient Prep, Insertion Bundle Complete?, Site Prep, Patient Tolerance, Inserted By, Discharged with Line/Drain/Tube, Removal Date/Time
- Line Timestamped Data Variable: Line/Site/Dressing WNL, Status, Line Exception, Line intervention, Needle Type, Needle -Manufacturer, Needle Gauge, Needle Length (decimal), Needle Length (fraction), Site exception, Site intervention, Dressing type, Dressing exception, Dressing intervention
- Endotracheal Tube Properties Variable: Placement Date, Placement Time, Airway Type, Size, Number of Attempts, Removal Date/Time
- Medication-Level Data Variable: Med name, Med dose, Med route, Med start date/time, Med stop date/time, Therapeutic Class, Pharmacy class, Pharmacy Subclass, Chemotherapy?, Antibiotic?, Fluid Bolus?, Dextrose Concentration if Fluid or TPN?, TPN?, Intralipid?, TPA for catheter clearance?, Blood Product? (PRBCs, Platelets, FFP, Cryo, Factor 8, G-CSF, IVIG), Sedation drip, Vasopressors/Inotropes, Systemic Steroid, Systemic Hydrocortisone, Systemic Immunosuppressant, Opioid pain medication, Paralytic, Diuretic, Insulin drip, Insulin intermittent

- Microbiology Data Variable: Culture Collection Date/Time, Culture Source (Specimen Description), Culture Source (Special Requests), Blood Culture Date/Time growth noted, Blood Culture Result – Gram Stain, Blood Culture Result – Species, Blood Culture Result – Susceptibilities, Respiratory Viral Panel, Respiratory Culture, Urine Culture, Stool PCR, Wound Culture, Eye culture
- Non-Micro Lab Data Variables: WBC, RBCS, HGB, HCT, MCV, MCH, MCHC, RDW, PLATELET COUNT, MEAN PLT VOLUME, AUTOMATED ABS NEUT, SEG, BAND, LYMPHOCYTE, MONOCYTE, METAMYELOCYTE, MYELOCYTE, Ammonia, Arterial pH, Arterial pO2, Arterial pCO2, Arterial O2 sat, Arterial Base Deficit, Venous pH, Venous pO2, Venous pCO2, Venous O2 sat, Venous Base Deficit, Capillary pH, Capillary pO2, Capillary pCO2, Capillary SaO2, Capillary Base Deficit, Lactate, Troponin, BNP, Cortisol, Na – lab, Na – gas, K – lab, K - gas, Cl – lab, HCO3 – lab, HCO3 – gas, BUN, Cr, BUN/Cr, Glucose - lab, Glucose - gas, Ca – lab, Ionized Ca – gas, Magnesium, Phosphorus, AST, ALT, Albumin, Total Protein, Alk Phos, GGT, Bilirubin, INR, PT, PTT, Anti-Xa, Fibrinogen, D-Dimer, AT3, Activated Clotting Time, CRP, ESR, CPK
- Presence and time of radiology studies: Chest X-ray, CT (any), Ultrasound (any), Abdominal X-Ray
- Immunization Data Variable: Immunizations received prior to hospital admission date, Immunizations received during hospital encounter
- Vital Signs Variable: Core Temperature, Temperature, Heart Rate, Respiratory rate, Arterial line BP, Non-invasive BP, SpO2, ET CO2, CVP, Capillary Refill
- Other Time-Stamped Data Variable: Tooth Brushing (Mouth Care), Rinse Agent, Bath (including CHG Bath), High touch surface clean protocol, Linen change, Patient Behaviors, Parent/Caregiver Behaviors, Parent/Caregiver Involvement, Braden Q score, CAPD scores, Current SBS (State Behavioral

Score), Desired SBS, WAT score, NIRS Left, NIRS Right, Oxygen Mode, FiO2 (%), Oxygen Flow (lpm), Type of Mechanical Ventilation, CPAP Pressure, Bilevel Pressures (IPAP/EPAP), Oxygenation Index, PaO2/FiO2, Type of Mechanical Ventilation, Ventilator Mode, Ventilator Rate, Set PIP, Measured PIP, PEEP, Tidal Volume Set, Tidal Volume Exhaled, Mean Airway pressure, Minute Ventilation, Spontaneous Rate, Inspiratory Time, Rise Time (slope), Pressure Support, Sensitivity, APRV Pressure High, APRV Pressure Low, APRV Time High, APRV Time Low, APRV Ventilator Rate, APRV Mean Airway Pressure, APRV Dump Volume, APRV Minute Volume, HFOV Hertz, HFOV DeltaP/Amplitude, HFOV Mean Airway Pressure, HFOV Inspiratory Time (%), Nitric Oxide Start/Stop, Inhaled Nitric Oxide, Urine Output (mL), Urinary Frequency, Bladder Scan, Emesis, Emesis (Frequency), Stool (measured), Stool frequency, Bladder Pressure, Output CVVH Ultrafiltrate, HD Positive Ultrafiltrate, HD Negative Ultrafiltrate, PD Positive Ultrafiltrate, PD Negative Ultrafiltrate, Time on, Time off, Time of order placement, Procedure name, Procedure start time, Procedure end time, Mucus fistula placement date, Fluoroscopy of mucus fistula.

A.2 Inclusion Flowchart

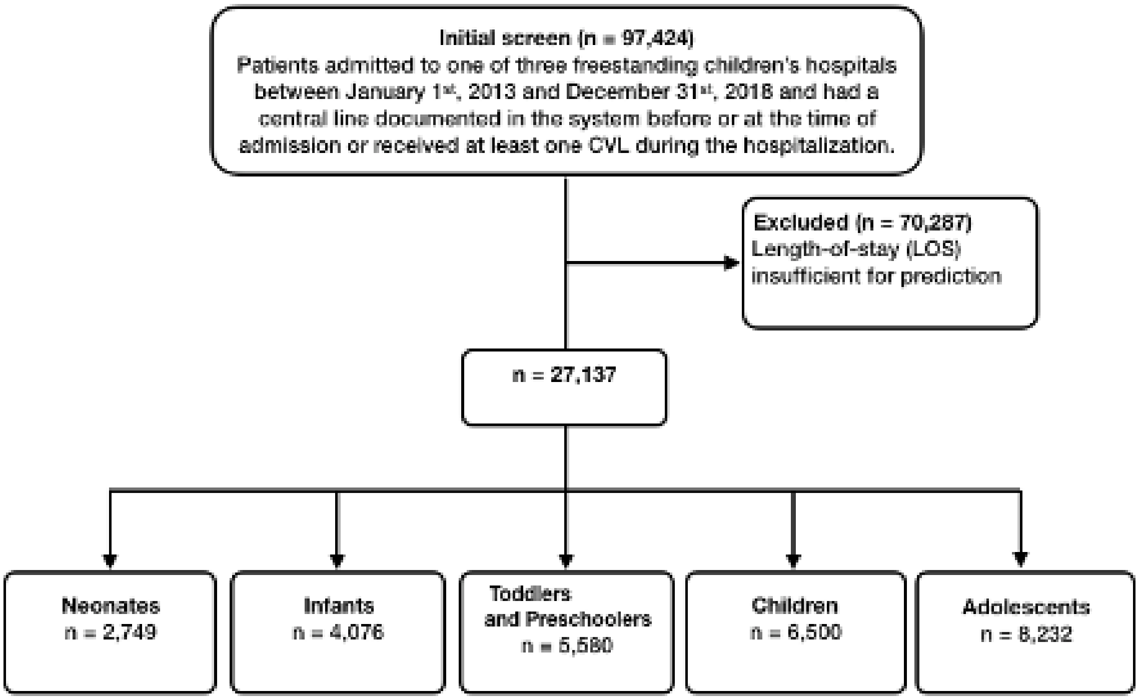

Initial screen (n = 97,424)
Patients admitted to one of three freestanding children's hospitals between January 1st, 2013 and December 31st, 2018 and had a central line documented in the system before or at the time of admission or received at least one CVL during the hospitalization.
Excluded (n = 70,287)
Length-of-stay (LOS) insufficient for prediction
n = 27,137
Neonates
n = 2,749
Infants
n = 4,076
Toddlers
and Preschoolers
n = 5,580
Children
n = 6,500
Adolescents
n = 8,232

Figure A.1: Inclusion flowchart. The final number of patient visits that we used in training and testing the machine learning models were 27,137. "CC BY 4.0"

A.3 Data Preprocessing

After combining all the information to the input variables, the data were preprocessed for modeling. Orders for laboratory tests are non-random, and their ordering frequency or absence may convey information about provider suspicion for infection. In order to control for such bias, we included an indicator flag for each laboratory value, such that any missing value could be incorporated within the model. This ensures that during the model training, 'missing not at random' variables are appropriately flagged and controlled across both PSI and non-PSI groups [61].

Moreover, we took the following steps for the preprocessing part:

- To reduce the effect of the outliers, we calculated the 0.5 and 99.5 percentiles of the features. These two values performed as the lower and upper bounds, respectively. Any value below the lower bound was replaced with the lower bound and any value above the upper bound was replaced with the upper bound.

- Numerical features were transformed to push their distribution more towards the normal distribution. Log or square root transformation was applied at this point.
- Missing information was imputed by the associated mean values.
- Features were scaled by subtracting the feature's mean value and dividing by the corresponding standard deviation.

Binary and categorical variables were one-hot encoded, to measure distinct aspects of these variables and to ensure that these categorical variables are appropriately distinguished from continuous measures.

Performing all the preprocessing steps lead to having a 338-dimensional feature space. Finally, LASSO feature selection was applied to remove features with negligible impact on the outcome. The final predictive models' input included 249 features which are listed in the following.

Laboratory Results: Minimum values of these tests: O2 Saturation Venous, Albumin, Alkaline Phosphatase, ALT SGPT, Ammonia, Arterial POC PH, AST SGOT, Atypical Reactive Lymphocyte, Automated ABS Neutrophil, BAND, BNP, BUN, Calcium, Capillary POC PCO2, Capillary POC PO2, Chloride, Creatinine, Eosinophils, Erythrocyte Sedimentation Rate, HCO3, Hemoglobin, International Normalized Ratio, Magnesium, MCH, MCHC, MCV, Mean Platelet Volume, Monocyte, Myelocyte, Platelet Count, POC Calcium Ionized, POC Glucose, POC Lactic Acid, Potassium, PTT, RBCS, Sodium, Total Protein. Maximum values of these tests: Albumin, Alkaline Phosphatase, Antithrombin Assay, Arterial Base Excess, Arterial POC PCO2, Arterial POC PO2, AST SGOT, Atypical Reactive Lymphocyte, Automated ABS Neutrophil, BAND, Bilirubin Total, CAP Base Excess, Capillary POC

PCO2, Capillary POC PH, Capillary POC PO2, Chloride, Cortisol, Creatinine, D Dimer Units, Eosinophils, Glucose, HCO3, Magnesium, MCH, MCHC, Mean Platelet Volume, Monocyte, Myelocyte, Phosphorus, Platelet Count, POC Calcium Ionized, POC Lactic Acid, POC Sodium, Potassium, Prothrombin time, RBCS, RDW, SEG, Sodium, Venous POC PO2, White blood cells. Missing flags for these tests: O2 Saturation Capillary, O2 Saturation Venous, Albumin, Alkaline Phosphatase, ALT SGPT, Ammonia, Antithrombin Assay, Arterial Base Excess, Arterial POC PCO2, Arterial POC PH, Arterial POC PO2, AST SGPT, Atypical Reactive Lymphocyte, Automated ABS Neutrophil, BAND, Bilirubin Total, BNP, BUN, C Reactive Protein, Calcium, CAP Base Deficit, CAP Base Excess, Capillary POC PCO2, Capillary POC PH, Capillary POC PO2, Chloride, Cortisol, Creatine phosphokinase, Creatinine, D Dimer Units, Eosinophils, Erythrocyte Sedimentation Rate, Fibrinogen, Gamma GGT, Glucose, HCO3, Hemoglobin, International Normalized Ratio, Magnesium, MCH, MCHC, MCV, Mean Platelet Volume, Metamyelocyte, Monocyte, Myelocyte, Phosphorus, Platelet Count, POC Calcium Ionized, POC Glucose, POC Lactic Acid, POC Potassium, POC Sodium, Potassium, Prothrombin time, PTT, RBCS, RDW, SEG, Sodium, Total Protein, Troponin, Venous POC PCO2, Troponin, Venous POC PCO2, Venous POC PH, Venous PO2, White blood cells. **Vital Signs:** Capillary Refill Minimum of these features: Core Temperature, Temperature, Heart Rate, Respiratory Rate, Arterial Line Blood Pressure, SpO2, ET CO2, CVP, Systolic Blood Pressure, Diastolic Blood Pressure, Glascow Coma Score. Maximum of these features: Temperature, Heart Rate, Respiratory Rate, Arterial Line Blood Pressure, SpO2, ET CO2, CVP, Systolic Blood Pressure, Diastolic Blood Pressure, Glascow Coma Score. **Pediatric Risk Scores:** PRISM-III score, PELOD-2 score **Demographics:** Age < 28 days, 1 year < Age < 4 years, Age < 12 years, Gestational age, Birth weight, Gender, Race (White, Black or African American, Non-Hispanic or Latino, Weight, Height, Insurance status (CMO Medicaid, Commercial, Managed

Care, Medicaid, Self-pay, Shared Service, Tricare) **Medications:** Chemotherapy, Fluid Bolus, TPN, Intralipid, TPA for catheter clearance, Sedation Drip Grouper, Sedation Drip Bolus, Vasopressors Inotropes Grouper, Systemic Steroid Grouper, Systemic immunosuppressant, Opioid pain medication, Paralytic, Diuretic, Insulin drip, Insulin intermittent, Mouth care, Rinse agent, Antimicrobial bath. **Line Properties and Line Insertion Information:** Gauge, Line type (Central Venous Line, Peripherally Inserted Central Catheter, Portacath, Power Port, Vascath/Permacath), Needle length (1 inch, 3/4 inch), Patient preparation (Central Line Insertion Bundle Completed, Child Life Consult, Comfort measures, Distraction, EMLA, Lidocaine, Position of comfort, Pre-medicated, Anesthesia Record Information, Sucrose pacifier, Not applicable due to patient condition, no patient preparation), Insertion Bundle Complete (Central Line Insertion Bundle Completed, Full Body Drape, Indwelling Urinary Catheter Insertion Bundle Completed, Sterile Gown, Sterile Prep Drape, No insertion bundle), Line Site Prep Used (Alcohol, Betadine, EMLA, LMX, Lidocaine), Patient Tolerance (Agitated, Calm, Cooperative, Crying/consolable, Tolerated well), Discharged With Line Drain Tube (Yes, No)

A.4 The predictive models

The 249-dimensional input was split into training (80%) and testing (20%) subsets. Since the data was class-imbalanced, we used stratified sampling to split the data in a way that the proportion of the minority class observations were the same in both subsets of the data. The training set was used to train the machine learning models and optimized the settings through tuning the models' hyperparameters.

The pBoost framework is developed using the proposed study design, a series of preprocessing tasks and eXtreme Gradient Boosting algorithm, also known as XGBoost [22]. After that, the performance of pBoost was compared with the logistic

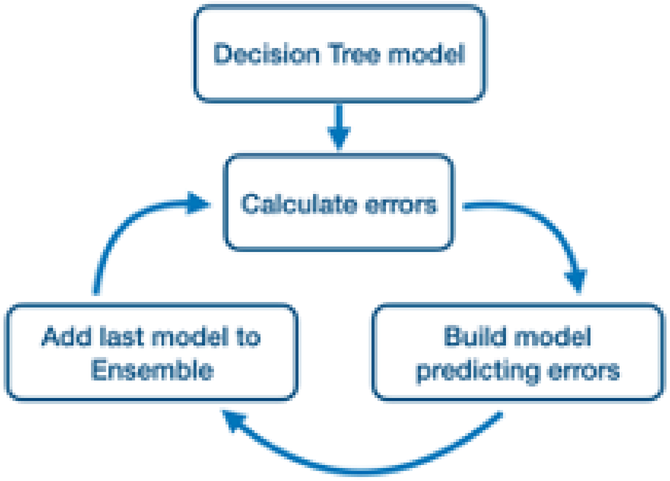


Figure A.2: Steps in training an XGBoost model. XGBoost is a boosting approach based on additive trees. At each iteration during the model training, a new tree will be added with the intention of correctly classifying the misclassified samples. “CC BY 4.0”

regression model with L1L2 regularization. The descriptions of the machine learning algorithms are presented in the following

eXtreme Gradient Boosting algorithm (XGBoost) The eXtreme Gradient Boosting algorithm, also known as XGBoost, is a boosting approach based on additive trees [22]. This model underweights the strong learners that classify correctly and overweights the ones that misclassify the target class. XGBoost is a variant of gradient boosting algorithms which is computationally efficient [35]. This model is capable of discriminating the positive class from the negative in extremely imbalanced data classification problems and achieved decent performance results in machine learning competitions [23]. XGBoost algorithm is based on additive trees meaning that at every training iteration the model fixes what it has learned and adds one new tree at a time. Then, the new tree tries to correctly classify the observations that were incorrectly classified in the previous training rounds (Figure A.2).

In machine learning algorithms, models have parameters and hyperparameters. The difference between these two groups is that the parameters are internal to the model and will be set based on the data but hyperparameters are external to the model and will not be estimated by the data. To achieve the best performance of a classification model, the hyperparameters should be tuned to have the optimized settings that lead to the desirable performance metrics. In this work, XGBoost hyperparameters such as depth of the trees and learning rate are optimized by employing Bayesian optimization algorithm [90] and the best setting of the model that achieves the highest AUC value is selected. The list of the optimized hyperparameters and their values are listed in the following.

- Learning rate = 0.23
- Maximum depth of the trees = 44
- Scale of positive (minority) class = 259.8
- Alpha (L1 regularization term on weights) = 0.69
- Lambda (L2 regularization term on weights) = 900.9
- Gamma (Minimum loss reduction required to make a further partition on a leaf node of the tree) = 0.054
- Subsample ration of the training instances = 0.66
- Subsample ratio of columns when constructing each tree = 0.76
- Subsample ratio of columns for each level = 0.31
- Maximum delta step we allow each leaf output to be = 1.66
- Minimum sum of instance weight needed in a child = 5

Regularized Logistic Regression with L1L2 Regularization (ElasticNet) Logistic regression model is an adapted version of linear regression to perform classification. By imposing a regularization term, this model can minimize the effect of the features with less significant effect on the classification outcome. There are three regularized version of logistic regression model: L1 (LASSO), L2 (Ridge), and L1L2 (ElasticNet).

In this study, we incorporated ElasticNet which is a weighted average of LASSO and Ridge. This model can penalize the less significant features by imposing an L1L2 penalty. The model's hyperparameters were tuned by cross-validation.

The analyses of this research were conducted in MATLAB 2019b and Python 3.6.

A.5 PRISM-III Score

Table A.1: This table presents the mean and standard deviation of PRISM-III score components. T-test was applied to test if there is a statistically significant difference between the mean of these features in PSI and non-PSI groups.

	PSI	non-PSI	p-value
Systolic Blood Pressure (mm Hg)			
Infants (mean [std])	78.3 [16]	81.6 [14.3]	< 0.001
Children (mean [std])	106.5 [15.6]	106.7 [15]	0.09
Diastolic Blood Pressure (mm Hg) (mean [std])	55.8 [15.5]	57.7 [14.8]	< 0.001
Heart Rate (beats per minute)			
Infants (mean [std])	154.1 [20.9]	148.4 [18.9]	< 0.001
Children (mean [std])	121.7 [25.7]	104.5 [24.5]	< 0.001
Respiratory Rate (breaths per minute)			
Infants (mean [std])	43.1 [17.1]	44.6 [14.6]	< 0.001
Children (mean [std])	25.7 [9.8]	23.5 [8.2]	< 0.001
PaO2/FiO2 (mean [std])	195.1 [449]	199.1 [580]	0.46
PaCO2 in torr (mm Hg) (mean [std])	45.7 [10.8]	44.4 [9]	< 0.001
Glasgow Coma Score (mean [std])	12.1 [3.7]	12.3 [3.6]	< 0.001
PT/PTT (mean [std])	0.44 [0.12]	0.44 [0.11]	0.31
Total bilirubin (mg/dL) (mean [std])	1.8 [3.4]	1.6 [2.9]	< 0.001
Potassium (mEq/L) (mean [std])	4 [0.7]	4.2 [0.7]	< 0.001
Calcium (mg/dL) (mean [std])	8.7 [0.8]	8.9 [0.8]	< 0.001
Glucose (mg/dL) (mean [std])	111.1 [43.5]	103.5 [37.8]	< 0.001
Bicarbonate in (mEq/L) (mean [std])	26.7 [5.9]	26.7 [5.2]	0.99

A.6 PELOD-2 Score

Table A.2: This table presents the mean and standard deviation of PELOD-2 score numerical components and percentages of the binary component which is the use of invasive ventilation. T-test and chi-square test were applied to test if there was a statistically significant difference between the features in PSI and non-PSI groups.

	PSI	non-PSI	p-value
Glasgow coma score (mean [std])	12.1 [3.7]	12.3 [3.6]	< 0.001
Lactatemia (mmol/L) (mean [std])	3.2 [7.6]	2.3 [4.5]	< 0.001
Mean arterial pressure (mmHg)			
Age in month <1 (mean [std])	65.7 [16.5]	68.6 [16.1]	< 0.001
1 < Age in month < 11 (mean [std])	74.6 [21.2]	81.6 [21.6]	< 0.001
12 < Age in month < 23 (mean [std])	92.3 [18.5]	95.2 [19.6]	< 0.001
24 < Age in month < 59 (mean [std])	97.7 [19.5]	98.6 [16.8]	0.2
60 < Age in month < 143 (mean [std])	101.5 [17.3]	102.7 [21.5]	0.23
Age in month >= 144 (mean [std])	113.7 [20.7]	115.7 [25.1]	< 0.001
Creatinine (mol/L)			
Age in month <1 (mean [std])	0.61 [0.79]	0.43 [0.49]	< 0.001
1 < Age in month < 11 (mean [std])	0.36 [0.28]	0.3 [0.23]	< 0.001
12 < Age in month < 23 (mean [std])	0.33 [0.38]	0.29 [0.23]	< 0.001
24 < Age in month < 59 (mean [std])	0.29 [0.21]	0.31 [0.42]	0.22
60 < Age in month < 143 (mean [std])	0.35 [0.4]	0.40 [0.62]	< 0.001
Age in month >= 144 (mean [std])	0.73 [1]	0.75 [1.26]	0.11
PaO2/FiO2 (mmHg) (mean [std])	195.1 [449]	199.1 [580]	0.46
PaCO2 (mmHg) (mean [std])	45.7 [10.8]	44.4 [9]	< 0.001
Invasive ventilation (% Yes)	11.1%	9.9%	0.21
WBC (10^9/L) (mean [std])	11 [15.4]	10.3 [8.2]	< 0.001
Platelet (10^9/L) (mean [std])	230.2 [175.6]	300.5 [175.4]	< 0.001

A.7 Sensitivity Analysis

In this study, age of the patients is divided into five different groups:

- neonates: 0 - 28 days of age; 2,749 encounters and 25.5% PSI prevalence
- infants: 21 days - 1 year of age; 4,076 encounters and 17.8% PSI prevalence
- toddlers and preschoolers: 1 - 5 years of age; 5,580 encounters and 12.4% PSI prevalence
- children: 5-12 years; 6,500 encounters and 9.7% PSI prevalence
- adolescents: older than 12 years. 8,232 encounters and 10.7% PSI prevalence

The number of records and the percentage of PSI patients for each group are mentioned in Table A.3. The chi-squared proportion test was applied to test if the difference in the proportion of PSI patients in male and female groups were statistically significant. The associated p-value is listed in the last row of Table A.3.

The highest and lowest PSI prevalence, 25.5% and 9.7%, were observed in neonates and children, respectively. We investigated the performance of pBoost model on each age group broken down by male and female (Tables A.3 to A.8). The pBoost model performed best on adolescents with average AUC of 0.84 [0.83, 0.85] with slight drop in performance in other age groups; an average AUC of 0.81 [0.80, 0.83] for children, 0.80 [0.78, 0.82] for toddlers and infants, and 0.74 [0.72, 0.76] for neonates.

Table A.3: The performance metrics of the pBoost model applied to all testing records and considering male and female patients separately.

	All Age Categories		
	Male	Female	All
AUROC (Mean [95% CI])	0.83 [0.82, 0.84]	0.81 [0.79, 0.81]	0.83 [0.82, 0.84]
Sensitivity (Mean [95% CI])	0.71 [0.68, 0.74]	0.66 [0.63, 0.69]	0.70 [0.67, 0.73]
Specificity (Mean [95% CI])	0.80 [0.79, 0.80]	0.80 [0.79, 0.80]	0.80 [0.79, 0.80]
PPV (Mean [95% CI])	0.38 [0.36, 0.39]	0.32 [0.31, 0.33]	0.35 [0.34, 0.36]
NPV (Mean [95% CI])	0.94 [0.94, 0.95]	0.94 [0.94, 0.95]	0.94 [0.94, 0.95]
F-1 Score (Mean [95% CI])	0.49 [0.47, 0.51]	0.43 [0.41, 0.44]	0.47 [0.46, 0.49]
Accuracy (Mean [95% CI])	0.78 [0.78, 0.79]	0.78 [0.77, 0.78]	0.78 [0.78, 0.79]
p-value of chi-squared test	0.2		

Table A.4: The performance metrics of the pBoost model applied to all neonates in testing records and considering male and female neonates patients separately.

	Neonates		
	Male	Female	All
AUROC (Mean [95% CI])	0.73 [0.70, 0.76]	0.75 [0.73, 0.78]	0.74 [0.72, 0.76]
Sensitivity (Mean [95% CI])	0.51 [0.43, 0.58]	0.54 [0.46, 0.61]	0.53 [0.47, 0.58]
Specificity (Mean [95% CI])	0.79 [0.76, 0.80]	0.79 [0.76, 0.80]	0.79 [0.78, 0.80]
PPV (Mean [95% CI])	0.50 [0.45, 0.54]	0.44 [0.41, 0.48]	0.47 [0.45, 0.50]
NPV (Mean [95% CI])	0.80 [0.77, 0.82]	0.85 [0.83, 0.87]	0.83 [0.81, 0.84]
F-1 Score (Mean [95% CI])	0.50 [0.44, 0.56]	0.49 [0.44, 0.54]	0.50 [0.46, 0.54]
Accuracy (Mean [95% CI])	0.71 [0.68, 0.73]	0.73 [0.71, 0.75]	0.73 [0.71, 0.74]
p-value of chi-squared test	0.17		

Table A.5: The performance metrics of the pBoost model applied to all infants in testing records and considering male and female infants patients separately.

	Infants		
	Male	Female	All
AUROC (Mean [95% CI])	0.84 [0.82, 0.86]	0.75 [0.73, 0.78]	0.80 [0.78, 0.82]
Sensitivity (Mean [95% CI])	0.72 [0.65, 0.79]	0.56 [0.47, 0.64]	0.64 [0.59, 0.69]
Specificity (Mean [95% CI])	0.79 [0.75, 0.80]	0.79 [0.76, 0.80]	0.79 [0.77, 0.80]
PPV (Mean [95% CI])	0.41 [0.37, 0.43]	0.34 [0.30, 0.37]	0.38 [0.36, 0.40]
NPV (Mean [95% CI])	0.93 [0.92, 0.95]	0.90 [0.89, 0.92]	0.92 [0.91, 0.93]
F-1 Score (Mean [95% CI])	0.52 [0.48, 0.56]	0.42 [0.37, 0.47]	0.48 [0.44, 0.51]
Accuracy (Mean [95% CI])	0.77 [0.75, 0.79]	0.75 [0.73, 0.77]	0.77 [0.75, 0.78]
p-value of chi-squared test	0.83		

Table A.6: The performance metrics of the pBoost model applied to all toddlers in testing records and considering male and female toddler patients separately.

	Toddlers and Preschoolers		
	Male	Female	All
AUROC (Mean [95% CI])	0.81 [0.79, 0.83]	0.80 [0.78, 0.82]	0.80 [0.78, 0.82]
Sensitivity (Mean [95% CI])	0.64 [0.57, 0.70]	0.64 [0.57, 0.71]	0.64 [0.59, 0.69]
Specificity (Mean [95% CI])	0.79 [0.76, 0.80]	0.79 [0.77, 0.80]	0.79 [0.78, 0.80]
PPV (Mean [95% CI])	0.31 [0.28, 0.33]	0.29 [0.26, 0.31]	0.30 [0.28, 0.32]
NPV (Mean [95% CI])	0.94 [0.93, 0.95]	0.94 [0.93, 0.95]	0.94 [0.93, 0.95]
F-1 Score (Mean [95% CI])	0.41 [0.38, 0.45]	0.40 [0.36, 0.43]	0.41 [0.38, 0.43]
Accuracy (Mean [95% CI])	0.77 [0.75, 0.78]	0.77 [0.75, 0.78]	0.78 [0.76, 0.78]
p-value of chi-squared test	0.69		

Table A.7: The performance metrics of the pBoost model applied to all children in testing records and considering male and female children patients separately.

	Children		
	Male	Female	All
AUROC (Mean [95% CI])	0.83 [0.82, 0.85]	0.80 [0.78, 0.82]	0.81 [0.80, 0.83]
Sensitivity (Mean [95% CI])	0.75 [0.68, 0.82]	0.63 [0.56, 0.69]	0.69 [0.64, 0.73]
Specificity (Mean [95% CI])	0.78 [0.72, 0.80]	0.78 [0.75, 0.80]	0.79 [0.77, 0.80]
PPV (Mean [95% CI])	0.29 [0.25, 0.31]	0.22 [0.20, 0.25]	0.26 [0.24, 0.28]
NPV (Mean [95% CI])	0.96 [0.95, 0.97]	0.95 [0.95, 0.96]	0.96 [0.95, 0.97]
F-1 Score (Mean [95% CI])	0.42 [0.38, 0.45]	0.33 [0.30, 0.36]	0.38 [0.36, 0.40]
Accuracy (Mean [95% CI])	0.78 [0.73, 0.80]	0.77 [0.74, 0.79]	0.78 [0.76, 0.79]
p-value of chi-squared test	0.41		

Table A.8: he performance metrics of the pBoost model applied to all adolescents in testing records and considering male and female adolescent patients separately.

	Adolescents		
	Male	Female	All
AUROC (Mean [95% CI])	0.85 [0.84, 0.87]	0.82 [0.81, 0.84]	0.84 [0.83, 0.85]
Sensitivity (Mean [95% CI])	0.76 [0.70, 0.81]	0.69 [0.64, 0.73]	0.74 [0.70, 0.77]
Specificity (Mean [95% CI])	0.79 [0.76, 0.80]	0.79 [0.76, 0.80]	0.79 [0.78, 0.80]
PPV (Mean [95% CI])	0.36 [0.33, 0.38]	0.26 [0.24, 0.28]	0.31 [0.30, 0.33]
NPV (Mean [95% CI])	0.96 [0.95, 0.96]	0.96 [0.95, 0.96]	0.96 [0.95, 0.96]
F-1 Score (Mean [95% CI])	0.49 [0.46, 0.51]	0.38 [0.35, 0.40]	0.44 [0.42, 0.46]
Accuracy (Mean [95% CI])	0.79 [0.77, 0.80]	0.78 [0.75, 0.79]	0.79 [0.77, 0.80]
p-value of chi-squared test	0.04		

Appendix B

B.1 Data Preprocessing

At the time of the study, the data from 2013 to 2018 was provided by the institution, so we could not acquire 2019 and 2020 data.

After gathering all the windows from the patients' hospitalization, the data was preprocessed and ready for model training. Initially, there were 252 features in the data. No information from discharge was passed to the input of the predictive model. The following preprocessing steps were done:

- Data capping: to reduce the effect of the outliers, 0.5 and 99.5 percentiles of each feature were calculated and set as the upper and lower bounds for the values. Any value above the upper bound was replace with the upper bound and any value below the lower bound was replaced with the lower bound.
- Transformation: the numerical variables were log or square root transformed to push their distribution more towards the normal distribution.
- Imputation: the missing values were imputed with the median value of the corresponding feature.
- Standardization: each feature was scaled by subtracting the corresponding feature's mean value and dividing by the associated standard deviation.

We applied one-hot-encoding to the binary and categorical variables, to measure distinct aspects of these variables and to ensure that these categorical variables are appropriately distinguished from continuous measures. Missing laboratory tests could not be at random, so they may convey information [61]. To capture that information, we added a binary flag for each laboratory value. The flag is one when the value is missing and zero otherwise.

To remove collinearity, we set a threshold of 0.8 for pair-wise correlation among features. If the pairwise correlation between two features exceeded the threshold, we removed one of them and kept the other one in the input. After removing collinearity with the threshold of 0.8, the feature space reduced to 135 dimensions. The selected features are listed in the following.

Laboratory Results: O2 Saturation Capillary, O2 Saturation Venous, Albumin, Alkaline Phosphatase, ALT SGPT, Ammonia, Antithrombin Assay, Arterial Base Excess, Arterial POC PCO2, Arterial POC PH, Arterial POC PO2, AST SGOT, Atypical Reactive Lymphocyte, Automated Absolute Neutrophil, BAND, Bilirubin Total, BNP, Blood Urea Nitrogen, C-Reactive Protein, Calcium, CAP Base Deficit, CAP Base Excess, Capillary POC PCO2, Capillary POC PH, Capillary POC PO2, Chloride, Cortisol, Creatine Phosphokinase, Creatinine, D-dimer Units, Eosinophils, Erythrocyte Sedimentation Rate, Fibrinogen, Gamma GGT, Glucose, HCO3, Hemoglobin, International Normalized Ratio, Magnesium, MCH, MCHC, MCV, Mean Platelet Volume, Metamyelocyte, Monocyte, Myelocyte, Phosphorus, Platelet Count, POC Calcium Ionized, POC Glucose, POC Lactic Acid, POC Potassium, POC Sodium, Potassium, PTT, Red Cell Distribution Width, SEG, Sodium, Total Protein, Troponin, Venous POC PCO2, Venous POC PH, White Blood Cells, Missing Saturation Capillary, Missing O2 Saturation Venous, Missing Albumin, Missing Ammonia, Missing Arterial Base Excess, Missing Atypical Reactive Lymphocyte, Missing Automated Absolute Neutrophil, Missing BAND, Missing BNP, Missing C-

Reactive Protein, Missing D-dimer Units, Missing Eosinophils, Missing Fibrinogen, Missing HCO3, Missing Magnesium, Missing Metamyelocyte, Missing Phosphorus, Missing POC Lactic Acid. **Vital Signs:** Core Temperature, Temperature, Heart Rate, Respiratory rate, Arterial line BP, SpO2, ET CO2, CVP, Systolic BP, Diastolic BP, Capillary Refill, GCS, PaO2/FiO2. **Mechanical Ventilation:** Was the patient on mechanical ventilation?. **Demographics:** Age $<$ 28 days, 29 days $<$ Age $<$ 1 year, 1 year $<$ Age $<$ 4 year, 5 year $<$ Age $<$ 11 year, Age $>$ 12 years, Gestational Age, Birth Weight, Gender, Race (Asian), Race (White), Race (Black or African American), Race (American Indian or Alaska Native), Race (Native Hawaiian or Other Pacific Islander), Ethnicity (Hispanic or Latino), Ethnicity (Non-Hispanic or Latino), Admission Weight, Insurance Status (CMO Medicaid), Insurance Status (Commercial), Insurance Status (Managed Care), Insurance Status (Medicaid), Insurance Status (Medicare), Insurance Status (Out of State Medicaid), Insurance Status (Self-pay), Insurance Status (Shared Service), Insurance Status (Tricare), Caregiver Cognitive Factors Flag. **Medications:** Chemotherapy, Fluid Bolus, TPN, Intralipid, Sedation Drip Grouper, Sedation Drip Bolus, Vasopressors Inotropes Grouper, Systemic Steroid Grouper, Opioid pain medication, Diuretic, Rinse Agent, Antimicrobial Bath. **Line Properties and Line Insertion Information:** Gauge, Line Type (Apheresis Port Dual)

B.2 Training/Validation/Testing Data Splits

All the models and analysis were performed in Python 3.6. To avoid data leakage, we split the data to train and test sets based on patient encounters; therefore, if a patient encounter was selected to be in the training set cohort, there were no information leakage to the testing set.

The PSI* prevalence across all 48-hour time windows was 0.34% which implied an extreme class-imbalanced classification problem. To preserve the same prevalence in training and testing sets, we used stratified sampling method to split data to training set (80%) and testing set (20%). There were multiple hyperparameters in our model which required to be optimized. So, we split the training patient encounters to 90% and 10% subsets using stratified sampling and incorporated the smaller set as the validation cohort in the hyperparameter optimization process.

B.3 The Input Structure

We employed a bidirectional Long Short-term Memory (LSTM) model to predict if there would be a PSI* event during the next 48 hours of hospitalization. We aimed to use the model every 8 hours to reflect the shift change; therefore, we need the inputs to be sequences of feature values which was gathered every 8 hours during a patient's hospitalization time.

The Bidirectional LSTM model takes a 3-dimensional input in this format: (number of patient encounters, number of timesteps, number of features) Each patient had a specific number of windows for prediction as this number increased with higher length-of-stay. By default, the number of features is fixed but the number of timesteps can vary. But the model needs a fixed value for the second dimension. To set the number of timesteps, we considered the maximum number of timesteps that a patient had in our training cohort which was 168. To have the same sequence length, we zero padded the information of the patients who had less that 168 timesteps in their hospital stay. LSTM-based models are designed in a way that they can skip these padded timesteps so that it will not hurt the model's outcome.

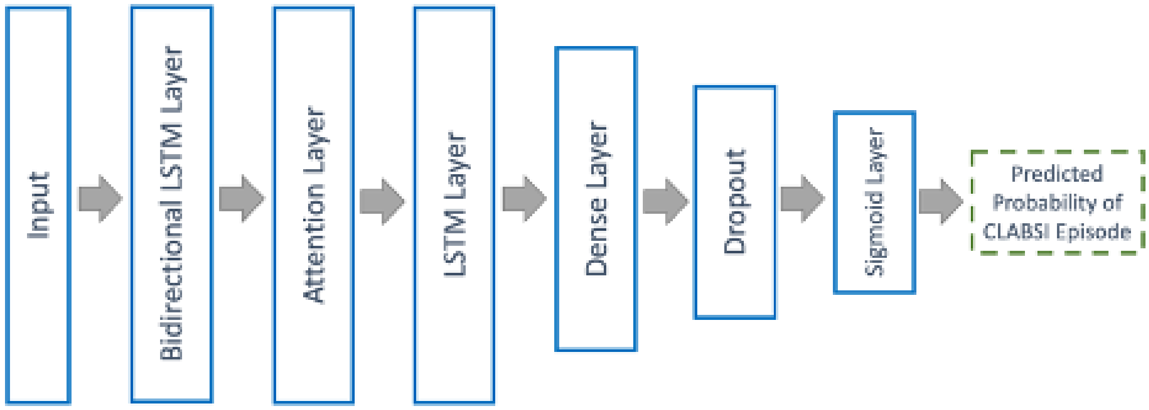


Figure B.1: The proposed model structure. "CC BY 4.0"

B.4 Model Specifications

The following figure presents the proposed model structure. We trained a bidirectional LSTM model with Focal loss and attention mechanism which used a batch size of 128. The hyperparameters of the model were optimized by employing Bayesian optimization method with 100 epochs and an early stopping criterion if there was no improve in the model's performance after 5 iterations. A list of the hyperparameters along with their optimized value are listed in the following. We did not change the default parameters of Focal loss (alpha=0.25, gamma=2).

- Adam optimizer with learning rate = 0.1
- Dropout regularization = 0.5
- Number of hidden units in the bidirectional LSTM model = 512
- Number of hidden units in the unidirectional LSTM model = 8
- Hidden units of the dense layer prior to the classification layer = 8

A Sigmoid layer was added for the final classification task.

B.5 Confidence Interval of Performance Metrics

Bootstrap method with 1000 repetition was used to calculate an estimation of the 95% confidence interval for the performance metrics.

B.6 PELOD-2 Score

Table B.1: This table presents the mean and standard deviation of PELOD-2 score components. The comparisons were done across the time windows. T-test and chi-square test were applied to test if there was a statistically significant difference between the feature values in windows with PSI* and windows without PSI* in patients' hospitalization time.

	Time windows with PSI*	Time windows without PSI*	p-value
Glasgow coma score (mean [std])	11.9 [3.7]	12.3 [3.6]	< 0.001
Lactatemia (mmol/L) (mean [std])	4.1 [7.1]	2.2 [4.4]	< 0.001
Mean arterial pressure (mmHg)			
Age in month <1 (mean [std])	71.9 [11.7]	68.9 [15]	0.026
1 < Age in month < 11 (mean [std])	75 [19.4]	68 [18]	< 0.001
12 < Age in month < 23 (mean [std])	74.7 [30.7]	66.4 [20.8]	0.026
24 < Age in month < 59 (mean [std])	72.2 [26.9]	74.6 [20.5]	0.311
60 < Age in month < 143 (mean [std])	85.7 [18.2]	86.9 [19.7]	0.559
Age in month >= 144 (mean [std])	111.1 [23.3]	107.7 [23.7]	0.008
Creatinine (mol/L)			
Age in month <1 (mean [std])	0.34 [0.23]	0.38 [0.36]	0.108
1 < Age in month < 11 (mean [std])	0.49 [0.41]	0.49 [0.66]	0.912
12 < Age in month < 23 (mean [std])	0.32 [0.14]	0.37 [0.31]	0.259
24 < Age in month < 59 (mean [std])	0.39 [0.47]	0.31 [0.26]	< 0.001
60 < Age in month < 143 (mean [std])	0.24 [0.09]	0.27 [0.14]	0.04
Age in month >= 144 (mean [std])	0.58 [1.02]	0.58 [0.97]	0.796
PaO2/FiO2 (mmHg) (mean [std])	25.6 [61.9]	34 [103.7]	0.002
PaCO2 (mmHg) (mean [std])	47.7 [10.7]	45 [9.6]	< 0.001
Invasive ventilation (% Yes)	38.1	34.1	0.002
WBC (10^9/L) (mean [std])	8.2 [9.2]	10.4 [8.2]	< 0.001
Platelet (10^9/L) (mean [std])	166.8 [154.7]	295.3 [182.9]	< 0.001

B.7 PRISM-III score's components

Table B.2: This table presents the mean and standard deviation of PRISM-III score components. The comparisons were done across the time windows. T-test was applied to test if there is a statistically significant difference between the mean of these features in windows with PSI* and windows without PSI* in patients' hospitalization time.

	With PSI*	Without PSI*	p-value
Systolic Blood Pressure (mm Hg)			
Infants (mean [std])	92.8 [16.9]	93.1 [15.8]	0.633
Children (mean [std])	104.7 [13.4]	105.8 [13.8]	0.013
Diastolic Blood Pressure (mm Hg) (mean [std])	57.2 [14.6]	57.6 [15]	0.128
Heart Rate (beats per minute)			
Infants (mean [std])	143 [24]	135 [22]	< 0.001
Children (mean [std])	115 [23]	106 [22]	< 0.001
Respiratory Rate (breaths per minute)			
Infants (mean [std])	39.8 [14.8]	37.3 [13.7]	< 0.001
Children (mean [std])	24.1 [7.6]	22.8 [6.3]	< 0.001
PaO2/FiO2 (mean [std])	25.6 [61.9]	34 [103.7]	0.002
PaCO2 in torr (mm Hg) (mean [std])	47.7 [10.7]	45 [9.6]	< 0.001
Glasgow Coma Score (mean [std])	11.9 [3.7]	12.3 [3.6]	< 0.001
PT/PTT (mean [std])	0.43 [0.12]	0.44 [0.11]	0.553
Total bilirubin (mg/dL) (mean [std])	2 [3.9]	1.6 [2.9]	< 0.001
Potassium (mEq/L) (mean [std])	4 [0.69]	4.1 [0.72]	< 0.001
Calcium (mg/dL) (mean [std])	8.7 [0.8]	8.9 [0.8]	< 0.001
Glucose (mg/dL) (mean [std])	104.8 [37.4]	103.1 [37.9]	0.021
Bicarbonate in (mEq/L) (mean [std])	27.9 [5.9]	27.1 [5.4]	< 0.001

B.8 Model performance on different patient race categories

Patients with race category of white

- Training set: PSI* prevalence = 0.31%, number of encounters = 11894, number of 48-hour windows = 292465
- Testing set: PSI* prevalence = 0.35%, number of encounters = 2917, number of 48-hour windows = 91717

Patients with race category of black

- Training set: PSI* prevalence = 0.4%, number of encounters = 7707, number of 48-hour windows = 223528
- Testing set: PSI* prevalence = 0.36%, number of encounters =1951, number of 48-hour windows = 71937

Patients with race category of other

- Training set: PSI* prevalence = 0.3%, number of encounters = 2108, number of 48-hour windows = 51565
- Testing set: PSI* prevalence = 0.41%, number of encounters = 560, number of 48-hour windows = 17168

Table B.3: This table presents the performance of the proposed model on patients with different race categories in training and testing subsets of the data.

	Race (White)		Race (Black)		Race (Other)	
	Train	Test	Train	Test	Train	Test
AUROC (%)	99.6 [99.5, 99.7]	99.5 [99.3, 99.6]	99.5 [99.4, 99.7]	99.4 [99.1, 99.6]	99.7 [99.4, 99.9]	98.7 [96.4, 99.9]
Sensitivity (%)	84.3 [79.5, 88.5]	76.1 [66.2, 85.5]	85.4 [80.2, 90.2]	77.9 [66.4, 87.8]	85.8 [73.9, 96.1]	75.6 [51.3, 97.1]
Specificity (%)	99.3 [99.3, 99.4]	99.3 [99.3, 99.4]	99.1 [99.1, 99.2]	99.1 [99.0, 99.1]	99.4 [99.3, 99.4]	99.4 [99.3, 99.5]
PPV (%)	7.1 [6.6, 7.6]	7.0 [6.0, 8.1]	8.1 [7.6, 8.7]	6.3 [5.2, 7.4]	6.9 [5.6, 8.2]	8.9 [5.8, 11.9]
NPV (%)	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]
Accuracy (%)	99.3 [99.3, 99.4]	99.3 [99.3, 99.3]	99.1 [99.1, 99.2]	99.1 [99.0, 99.1]	99.3 [99.3, 99.4]	99.4 [99.3, 99.5]
F-1 Score (%)	13.0 [12.2, 13.9]	12.9 [11.0, 14.7]	14.8 [13.8, 15.8]	11.7 [9.7, 13.6]	12.7 [10.5, 15.0]	15.9[10.7, 21.1]
AUPRC (%)	34.5 [27.4, 42.2]	20.2 [12.4, 29.6]	36.4 [29.4, 43.3]	21.2 [12.2, 32.3]	46.8 [30.8, 63.5]	43.4 [18.1, 67.7]

A one-way ANOVA test was performed and achieved a p-value of 0.29 for training and 0.86 for testing datasets. The results indicated no statistically significant difference between the mean of the model's predicted probabilities for each race category.

B.9 Model performance on different patient insurance categories

Patients with Commercial insurance

- Training set: PSI* prevalence = 0.34%, number of encounters = 8772, number of 48-hour windows = 200683
- Testing set: PSI* prevalence = 0.37%, number of encounters = 2105, number of 48-hour windows = 60620

Patients with Public-Medicaid insurance

- Training set: PSI* prevalence = 0.35%, number of encounters = 12078, number of 48-hour windows = 347905
- Testing set: PSI* prevalence = 0.35%, number of encounters = 3124, number of 48-hour windows = 114446

Patients with Public-Medicare insurance

- Training set: PSI* prevalence = 0.41%, number of encounters = 665, number of 48-hour windows = 15407
- Testing set: PSI* prevalence = 0.27%, number of encounters = 156, number of 48-hour windows = 4707

Patients with Self-pay insurance

- Training set: PSI* prevalence = 0.19%, number of encounters = 194, number of 48-hour windows = 3563

- Testing set: PSI* prevalence = 0.64%, number of encounters = 43, number of 48-hour windows = 1049

Table B.4: This table presents the performance of the proposed model on patients with different insurance status in training and testing subsets of the data.

	Commercial		Public-Medicaid		Public-Medicare		Self-pay	
	Train	Test	Train	Test	Train	Test	Train	Test
AUROC (%)	99.6 [99.4, 99.7]	99.4 [98.7, 99.8]	99.6 [99.5, 99.7]	99.3 [99.1, 99.5]	99.8 [99.6, 99.9]	98.9 [97.8, 99.9]	99.9 [99.9, 99.9]	99.9 [99.8, 99.9]
Sensitivity (%)	80.1 [73.8, 85.7]	79.1 [69.1, 88.8]	87.4 [83.0, 91.6]	75.6 [65.9, 84.2]	89.8 [77.0, 99.9]	55.3 [14.3, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]
Specificity (%)	99.4 [99.4, 99.5]	99.4 [99.4, 99.5]	99.1 [99.1, 99.2]	99.1 [99.0, 99.1]	99.4 [99.3, 99.5]	99.4 [99.2, 99.6]	99.5 [99.4, 99.7]	99.7 [99.4, 99.9]
PPV (%)	8.1 [7.3, 8.9]	8.6 [7.3, 9.9]	7.2 [6.7, 7.6]	6.1 [5.2, 6.9]	9.7 [7.6, 11.8]	4.4 [1.1, 8.9]	5.5 [4.0, 7.6]	25.0 [14.3, 46.7]
NPV (%)	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]
Accuracy (%)	99.4 [99.4, 99.5]	99.4 [99.4, 99.5]	99.1 [99.1, 99.2]	99.1 [99.0, 99.1]	99.4 [99.3, 99.5]	99.4 [99.2, 99.6]	99.5 [99.4, 99.7]	99.7 [99.4, 99.9]
F-1 Score (%)	14.7 13.3, 16.1]	15.5 [13.3, 17.8]	13.2 [12.5, 14.0]	11.3 [9.6, 12.8]	17.5 [14.0, 21.1]	8.2 [2.0, 16.3]	10.5 [7.7, 14.1]	39.4 [25.0, 63.6]
AUPRC (%)	36.0 [27.4, 44.6]	21.6 [13.4, 32.0]	37.2 [30.6, 43.5]	24.5 [16.1, 34.4]	31.7 [15.6, 53.4]	7.8 [1.6, 25.4]	73.6 [62.2, 90.7]	49.7 [30.2, 90.2]

A one-way ANOVA test was performed and achieved a p-value of 0.09 for training and 0.13 for testing datasets. The results indicated no statistically significant difference between the mean of the model's predicted probabilities for each insurance status category.

B.10 Sensitivity Analysis

Table B.5 in the following presents the performance of the proposed model on different patients' age categories. These categories were defined as:

- Neonates: age < 28 days

- Infants: 29 days < age < 1 year

- Toddlers and Preschoolers: 1 year < age < 4 years

- Children: 5 years < age < 11 years

- Adolescents: 12 years < age

In deep learning models, as the positive cases increase, we can expect higher PPV value which is the same story in our analysis. In this study cohort, the prevalence of PSI* was slightly different across the five aforementioned age categories which influenced some of the performance metrics. Specifically, PPV and F-1 score increased as the PSI* prevalence increased. On the other hand, NPV, which is the power of the model to rule out the negative cases, was not affected by the changes in the prevalence as there were sufficient NP cases for the model to learn.

Table B.5: This table presents the performance of the proposed model on patients with different age categories in training and testing subsets of the data. The numbers in the brackets are the estimated 95% confidence interval calculated by bootstrapping method.

	Neonates		Infants		Toddlers		Children		Adolescents	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
PSI* (%)	0.32	0.27	0.39	0.27	0.42	0.44	0.32	0.38	0.30	0.45
AUROC (%)	99.4 [99.2, 99.6]	99.4 [99.1, 99.6]	99.5 [99.3, 99.6]	99.3 [98.8, 99.7]	99.7 [99.5, 99.8]	99.5 [99.2, 99.8]	99.7 [99.6, 99.9]	99.5 [99.2, 99.8]	99.6 [99.4, 99.7]	99.2 [98.5, 99.7]
Sensitivity (%)	94.9 [91.3, 97.7]	92.9 [82.3, 99.9]	87.1 [80.0, 93.1]	84.1 [67.0, 96.8]	79.1 [70.4, 87.1]	75.2 [59.7, 89.7]	86.0 [77.3, 93.7]	68.8 [48.3, 86.3]	77.9 [69.3, 85.4]	69.6 [56.3, 83.0]
Specificity (%)	98.1 [97.9, 98.2]	98.0 [97.9, 98.2]	98.8 [98.7, 98.8]	98.7 [98.6, 98.8]	99.5 [99.5, 99.5]	99.5 [99.5, 99.6]	99.5 [99.5, 99.5]	99.4 [99.4, 99.5]	99.6 [99.6, 99.6]	99.5 [99.5, 99.6]
PPV (%)	6.5 [6.1, 6.9]	5.4 [4.6, 6.1]	7.1 [6.4, 7.7]	4.6 [3.7, 5.4]	9.3 [8.2, 10.4]	10.1 [7.8, 12.5]	7.4 [6.5, 8.2]	6.5 [4.5, 8.4]	8.3 [7.1, 9.4]	9.7 [7.6, 11.7]
NPV (%)	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]	99.9 [99.9, 99.9]
Accuracy (%)	98.1 [97.9, 98.1]	98.0 [97.9, 98.2]	98.8 [98.7, 98.8]	98.7 [98.6, 98.8]	99.5 [99.4, 99.5]	99.5 [99.5, 99.6]	99.5 [99.5, 99.5]	99.4 [99.4, 99.5]	99.6 [99.6, 99.6]	99.5 [99.5, 99.6]
F-1 Score (%)	12.1 [11.4, 12.8]	10.2 [8.7, 11.5]	13.1 [11.9, 14.2]	8.7 [7.0, 10.2]	16.7 [14.8, 18.6]	17.8 [13.8, 21.9]	13.6 [12.0, 15.1]	11.9 [8.4, 15.2]	15.0 [13.0, 16.9]	17.0 [13.5, 20.4]
AUPRC (%)	26.5 [19.0, 34.1]	14.2 [8.4, 22.7]	34.7 [25.5, 44.0]	14.4 [6.6, 26.0]	45.0 [34.8, 55.0]	46.5 [26.8, 66.5]	50.5 [38.7, 61.6]	35.4 [16.2, 56.9]	35.4 [24.9, 46.6]	18.4 [9.3, 30.6]

Appendix C

C.1 Clinical Notes

The following clinical note types and their associated recording timestamps were extracted from Children's Healthcare of Atlanta database:

Notes, Progress Nursing, Progress Respiratory, CIRU Progress Notes, Procedures, Transfer Summary, Consults, Interval HP Note, Aflac Oncall Note, HP, Brief Op Note, OR PreOp, OR Op Note, Transplant Evaluation, Pre-Sedation HP, OR Anesthesia, OR PostOp, OR Surgeon, Rehab IP Progress/Treatment Notes, Code Documentation, ED Provider Notes, ED Supplemental Provider Note

C.2 Data Preprocessing

C.2.1 Structured Data

At the time of the study, the data from 2013 to 2018 was provided by the institution, so we could not acquire 2019 and 2020 data.

After gathering all the 48 hour time windows from the patients' hospitalization, the data was preprocessed and ready for model training. Initially, there were 252 features in the structured EHR data. No information from discharge was passed to the input of the predictive model. The following preprocessing steps were done:

- Data capping: to reduce the effect of the outliers, 0.5 and 99.5 percentiles of each feature were calculated and set as the upper and lower bounds for the values. Any value above the upper bound was replace with the upper bound and any value below the lower bound was replaced with the lower bound.
- Transformation: the numerical variables were log or square root transformed to push their distribution more towards the normal distribution.
- Imputation: the missing values were imputed with the median value of the corresponding feature.
- Standardization: each feature was scaled by subtracting the corresponding feature's mean value and dividing by the associated standard deviation.

We applied one-hot-encoding to the binary and categorical variables, to measure distinct aspects of these variables and to ensure that these categorical variables are appropriately distinguished from continuous measures. Missing laboratory tests could not be at random, so they may convey information (1). To capture that information, we added a binary flag for each laboratory value. The flag is one when the value is missing and zero otherwise.

To remove collinearity, we set a threshold of 0.8 for pair-wise correlation among features. If the pairwise correlation between two features exceeded the threshold, we removed one of them and kept the other one in the input. After removing collinearity with the threshold of 0.8, the feature space reduced to 129 dimensions. The selected features are listed in the following. **Laboratory Results:** O2 Saturation Capillary, O2 Saturation Venous, Albumin, Alkaline Phosphatase, ALT SGPT, Ammonia, Antithrombin Assay, Arterial Base Excess, Arterial POC PCO2, Arterial POC PH, Arterial POC PO2, AST SGOT, Atypical Reactive Lymphocyte, Automated Absolute Neutrophil, BAND, Bilirubin Total, BNP, Blood Urea Nitrogen, C-Reactive Protein, Calcium, CAP Base Deficit, CAP Base Excess, Capillary POC PCO2, Capillary POC PH, Capillary POC PO2, Chloride, Cortisol, Creatine Phosphokinase, Creatinine, D-dimer Units, Eosinophils, Erythrocyte Sedimentation Rate, Fibrinogen, Gamma GGT, Glucose, HCO3, Hemoglobin, International Normalized Ratio, Magnesium, MCH, MCHC, Mean Platelet Volume, Metamyelocyte, Monocyte, Myelocyte, Phosphorus, Platelet Count, POC Calcium Ionized, POC Glucose, POC Lactic Acid, POC Potassium, POC Sodium, Potassium, PTT, Red Cell Distribution Width, SEG, Sodium, Total Protein, Troponin, Venous POC PCO2, Venous POC PH, White Blood Cells, Missing Saturation Capillary, Missing O2 Saturation Venous, Missing Albumin, Missing Ammonia, Missing Arterial Base Excess, Missing Atypical Reactive Lymphocyte, Missing Automated Absolute Neutrophil, Missing BAND, Missing BNP, Missing C-Reactive Protein, Missing D-dimer Units, Missing Eosinophils, Missing Fibrinogen, Missing HCO3, Missing Magnesium, Missing Metamyelocyte, Missing Myelocyte, Missing Phosphorus, Missing POC Lactic Acid. **Vital Signs:** Core Temperature, Temperature, Heart Rate, Respiratory rate, Arterial line BP, SpO2, ET CO2, CVP, Systolic BP, Diastolic BP, Capillary Refill, GCS, PaO2/FiO2. **Mechanical Ventilation:** Was the patient on mechanical ventilation?. **Demographics:** Age $<$ 28 days, 29 days $<$ Age $<$ 1 year, 1 year $<$ Age $<$ 4 year, 5 year $<$ Age $<$ 11 year, Age $>$ 12 years, Gestational Age, Birth Weight, Gender, Race (Asian), Race (White), Race (Black or African American), Race (American Indian or Alaska Native), Race (Native Hawaiian or Other Pacific Islander), Ethnicity (Hispanic or Latino), Ethnicity (Non-Hispanic or Latino), Admission Weight, Insurance Status (CMO Medicaid), Insurance Status (Commercial), Insurance Status (Managed Care), Insurance Status (Medicaid), Insurance Status (Medicare), Insurance Status (Out of State Medicaid), Insurance Status (Self-pay), Insurance Status (Shared Service), Insurance Status (Tricare), Caregiver Cognitive Factors Flag. **Medications:** Chemotherapy, TPN, Sedation Drip Bolus, Systemic Steroid Grouper, Opioid pain medication, Diuretic. **Line Properties and Line Insertion Information:** Gauge, Line Type (Apheresis Port Dual).

C.2.2 Unstructured Data

Common text preprocessing steps (e.g., removing stop words, punctuations, extra white space, and numbers and transforming to lowercase) were done on the recorded clinical notes. To acquire the contextualized word embeddings from the recorded notes, we used Clinical BERT which is a transformer-based model. The Clinical BERT model accepts text with the maximum length of 128 words. To account for this limitation, we extracted the specific parts of the clinical notes that could be more informative in training the predictive models; therefore, we limited the number of input words for the Clinical BERT model while keeping the most discriminative parts of the notes in the modeling process. Based on the recommendation from a pediatrician, the following sections were extracted from the clinical notes:

impression and plan, impression plan, assessment plan, interval history, iterim history, history of present illness, brief hpi, hpi, subjective, patient active problem list, active hospital problems diagnosis, diagnosis code, impression/problems, medical decision making, mdm, review of systems, ros

After selecting the aforementioned sections in the clinical notes, the number of words in each clinical note had a distribution with median of 57 and mean of 127. The final word embedding dimension was 768.

C.3 Training/Validation/Testing Subsets of Data

All the models and analysis were performed in Python 3.6. To avoid data leakage, we split the data to train and test sets based on patient encounters; therefore, if a patient encounter was selected to be in the training set cohort, there were no information leakage to the testing set.

The SBI prevalence across all 48-hour time windows was 0.35% which implied an extreme class-imbalanced classification problem. To preserve the same prevalence in training and testing sets, we used stratified sampling method to split data to training set (80%) and testing set (20%). There were multiple hyperparameters in our model which required to be optimized. So, we split the training patient encounters to 90% and 10% subsets using stratified sampling and incorporated the smaller subset as the validation cohort in the hyperparameter optimization process.

C.4 The Input Structure

We employed a Bidirectional Long Short-term Memory (BiLSTM) model to predict if there would be a SBI event during the next 48 hours of hospitalization. We aimed to use the model every 24 hours to be confident that the clinical notes were updated; therefore, we need the inputs to be sequences of feature values which was gathered every 24 hours during a patient's hospitalization time.

The BiLSTM model takes a 3-dimensional input in this format: (number of patient encounters, number of timesteps, number of features)

Each patient had a specific number of windows for prediction as this number increased with higher length-of-stay. By default, the number of features is fixed but the number of timesteps can vary. But the model needs a fixed value for the second dimension. To set the number of timesteps, we considered the maximum number of timesteps that a patient had in our training cohort which was 56. To have the same sequence length, we zero padded the information of the patients who had less that 56 timesteps in their hospital stay. LSTM-based models are designed in a way that they can skip these padded timesteps so that it will not hurt the model's outcome.

The number of features was 897 (129 features from the structured EHR and 768 contextualized word embedding features from the clinical notes).

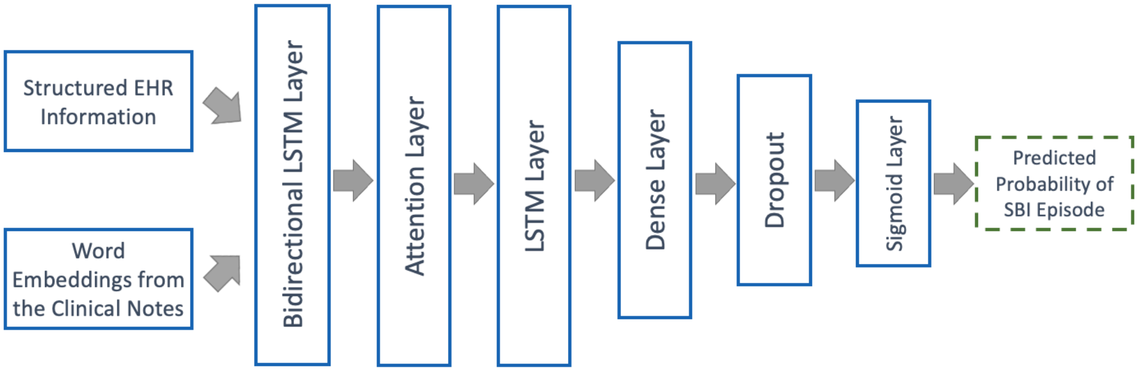

Structured EHR Information
Word Embeddings from the Clinical Notes
Bidirectional LSTM Layer
Attention Layer
LSTM Layer
Dense Layer
Dropout
Sigmoid Layer
Predicted Probability of SBI Episode

Figure C.1: The proposed model structure. "CC BY 4.0"

C.5 Model Specifications

The following figure presents the proposed model structure. We trained a bidirectional LSTM model with Focal loss and attention mechanism which used a batch size of 128. The hyperparameters of the model were optimized by employing Bayesian optimization method with 100 epochs and an early stopping criterion if there was no improve in the model's performance after 5 iterations. A list of the hyperparameters along with their optimized value are listed in the following. We did not change the default parameters of Focal loss (alpha=0.25, gamma=2).

- Adam optimizer with learning rate = 0.001
- Dropout regularization = 0.1
- Number of hidden units in the bidirectional LSTM model = 512
- Number of hidden units in the unidirectional LSTM model = 8
- Hidden units of the dense layer prior to the classification layer = 32

A Sigmoid layer was added for the final classification task. Figure C.1 presents the model structure.

C.6 Confidence Interval of Performance Metrics

Bootstrap method with 1000 repetition was used to calculate an estimation of the 95% confidence interval for the performance metrics.

C.7 PELOD-2 Score

Table C.1: This table presents the mean and standard deviation of PELOD-2 score components. The comparisons were done across the time windows. T-test and chi-square test were applied to test if there was a statistically significant difference between the feature values in SBI time windows and non-SBI time windows during patients' hospitalization time.

	SBI Time Windows	Non-SBI Time Windows	p-value
Glasgow coma score (mean [std])	11.8 [3.8]	12.3 [3.6]	0.003
Lactatemia (mmol/L) (mean [std])	4 [7]	2.2 [4.5]	< 0.001
Mean arterial pressure (mmHg)			
Age in month <1 (mean [std])	70.8 [11]	68.9 [15.1]	0.4
1 < Age in month < 11 (mean [std])	75.5 [19.7]	68.4 [18]	< 0.001
12 < Age in month < 23 (mean [std])	72.3 [32.4]	66.4 [20.9]	0.4
24 < Age in month < 59 (mean [std])	72.4 [27.6]	74.8 [20.5]	0.5
60 < Age in month < 143 (mean [std])	87.7 [19.7]	87 [19.8]	0.8
Age in month >= 144 (mean [std])	110.6 [22.1]	107.9 [23.8]	0.2
Creatinine (mol/L)			
Age in month <1 (mean [std])	0.33 [0.19]	0.39 [0.36]	0.2
1 < Age in month < 11 (mean [std])	0.48 [0.38]	0.49 [0.65]	0.9
12 < Age in month < 23 (mean [std])	0.31 [0.11]	0.37 [0.31]	0.5
24 < Age in month < 59 (mean [std])	0.37 [0.43]	0.31 [0.26]	0.04
60 < Age in month < 143 (mean [std])	0.24 [0.09]	0.27 [0.14]	0.1
Age in month >= 144 (mean [std])	0.56 [0.87]	0.58 [0.98]	0.5
PaO2/FiO2 (mmHg) (mean [std])	24.8 [55.3]	33.5 [106.7]	0.08
PaCO2 (mmHg) (mean [std])	48 [10.7]	45 [9.6]	< 0.001
Invasive ventilation (% Yes)	25.8%	33.5%	0.9
WBC (10^9/L) (mean [std])	8.6 [9.9]	10.5 [8.6]	< 0.001
Platelet (10^9/L) (mean [std])	173.9 [156.3]	297.8 [182.4]	< 0.001

C.8 PRISM-III Score

Table C.2: This table presents the mean and standard deviation of PRISM-III score components. The comparisons were done across the time windows. T-test was applied to test if there is a statistically significant difference between the mean of these features in SBI time windows and non-SBI time windows during patients' hospitalization time.

	SBI Time Windows	Non-SBI Time Windows	p-value
Systolic Blood Pressure (mm Hg)			
Infants (mean [std])	93.3 [16.6]	93.3 [15.9]	0.9
Children (mean [std])	105.7 [13.3]	106.6 [13.8]	0.3
Diastolic Blood Pressure (mm Hg) (mean [std])	57.8 [14.7]	58.1 [15.3]	0.6
Heart Rate (beats per minute)			
Infants (mean [std])	144.3 [23.6]	136.9 [21.9]	< 0.001
Children (mean [std])	118.9 [22.3]	107.8 [22.2]	< 0.001
Respiratory Rate (breaths per minute)			
Infants (mean [std])	39.8 [15.2]	37.5 [13.7]	< 0.001
Children (mean [std])	24.3 [7.6]	23 [6.5]	< 0.001
PaO2/FiO2 (mean [std])	24.8 [55.3]	33.5 [1.6.7]	0.08
PaCO2 in torr (mm Hg) (mean [std])	48 [10.7]	45 [9.6]	< 0.001
Glasgow Coma Score (mean [std])	11.8 [3.8]	12.3 [3.6]	0.003
PT/PTT (mean [std])	0.43 [0.11]	0.44 [0.11]	0.57
Total bilirubin (mg/dL) (mean [std])	2.1 [4.1]	1.6 [3]	< 0.001
Potassium (mEq/L) (mean [std])	4 [0.7]	4.1 [0.7]	< 0.001
Calcium (mg/dL) (mean [std])	8.8 [0.8]	8.9 [0.8]	< 0.001
Glucose (mg/dL) (mean [std])	105.2 [38.3]	102.9 [37.5]	0.07
Bicarbonate in (mEq/L) (mean [std])	27.9 [5.9]	27.1 [5.4]	< 0.001

Appendix D

D.1 List of Abbreviations

ABS NEUT	Absolute Neutrophil
ALT	Alanine Aminotransferase
APRV	Airway Pressure Release Ventilation
AST	Aspartate Aminotransferase
AT3	Antithrombin
AUPRC	Area Under Precision-Recall Curve
AUROC	Area Under Receiver Operating Characteristics
BiLSTM	Bidirectional Long Short-Term Memory
BMT	Bone Marrow Transplant
BP	Blood Pressure
BUN	Blood Urea Nitrogen
CAPD	Cornell Assessment of Pediatric Delirium
CHG	Chlorhexidine Gluconate
CHOA	Children's Healthcare of Atlanta
CI	Confidence Interval
CLABSI	Central Line-Associated Bloodstream Infection

CPK	Creatine Phosphokinase
CPAP	Continuous Positive Airway Pressure
CRP	C-reactive Protein
CSN	Contact Serial Number
CVL	Central Venous Line
CVP	Central Venous Pressure
CVVH	Continuous Veno-Venous Hemofiltration
DOB	Date of Birth
ECMO	Extracorporeal Membrane Oxygenation
ED	Emergency Department
EHR	Electronic Health Record
EMR	Electronic Medical
ESR	Erythrocyte Sedimentation Rate
FFP	Fresh Frozen Plasma
FN	False Negative
FP	False Positive
GCS	Glasgow Coma Scale
GGT	Gamma-Glutamyl Transferase
GVHD	Graft-Versus-Host Disease
HCT	Hematocrit
HFOV	High Frequency Oscillatory Ventilation
HGB	Hemoglobin
HP	History and Physical
ICD-9	International Classification of Diseases, Ninth Revision
ICD-10	International Classification of Diseases, Tenth Revision

INR	International Normalised Ratio
IPV	Intimate Partner Violence
IQR	Interquartile Range
IVIG	Intravenous Immune Globulin
LOS	Length of Stay
LSTM	Long Short-Term Memory
MCH	Mean Corpuscular Hemoglobin
MCHC	Mean Corpuscular Hemoglobin Concentration
MCV	Mean Corpuscular Volume
MRN	Medical Record Number
NEC	Necrotizing Enterocolitis
NICU	Newborn Intensive Care Unit
NIRS	Near Infrared Spectroscopy
NLP	Natural Language Processing
NPV	Negative Predictive Value
OR	Operating Room
PCR	Polymerase Chain Reaction
PELOD	PEdiatric Logistic Organ Dysfunction
PHOS	Phosphatase
PICU	Pediatric Intensive Care Unit
PLT	Platelet
PPV	Positive Predictive Value
PRBC	Packed Red Blood Cell
PRC	Precision-Recall Curve
PRSIM	Pediatric Risk of Mortality

PSI	Presumed Serious Infection
PT	Prothrombin time
PTT	Partial Thromboplastin Time
RDW	Red cell Distribution Width
ROC	Receiver Operating Characteristics
SBI	Serious Bloodstream Infection
SCID	Severe Combined Immunodeficiency
SEG	Segmented Neutrophils
SHAP	SHapley Additive exPlanations
TN	True Negative
TF-IDF	Term Frequency-Inverse Document Frequency
TP	True Positive
TPA	Tissue Plasminogen Activator
TPN	Total Parenteral Nutrition
WAT	Withdrawal Assessment Tool
WBC	White Blood Cell
XGBoost	Extreme Gradient Boosting

Bibliography

[1] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[2] J. A. Alten, A. F. Rahman, H. J. Zaccagni, A. Shin, D. S. Cooper, J. J. Blinder, L. Retzloff, I. B. Aban, E. M. Graham, J. Zampi, et al. The epidemiology of health-care associated infections in pediatric cardiac intensive care units. *The Pediatric infectious disease journal*, 37(8):768, 2018.

[3] F. Amrollahi, S. P. Shashikumar, F. Razmi, and S. Nemati. Contextual embeddings from clinical notes improves prediction of sepsis. In *AMIA Annual Symposium Proceedings*, volume 2020, page 197. American Medical Informatics Association, 2020.

[4] L. B. Amusa, A. V. Bengesai, and H. T. Khan. Predicting the vulnerability of women to intimate partner violence in south africa: evidence from tree-based machine learning techniques. *Journal of interpersonal violence*, page 0886260520960110, 2020.

[5] S. Bagchi, J. Watkins, D. A. Pollock, J. R. Edwards, and K. Allen-Bridson. State health department validations of central line–associated bloodstream infection events reported via the national healthcare safety network. *American journal of infection control*, 46(11):1290–1295, 2018.

[6] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[7] C. Beeler, L. Dbeibo, K. Kelley, L. Thatcher, D. Webb, A. Bah, P. Monahan, N. R. Fowler, S. Nicol, A. Judy-Malcolm, et al. Assessing patient risk of central line-associated bacteremia via machine learning. *American journal of infection control*, 46(9):986–991, 2018.

[8] R. A. Berk and S. B. Sorenson. Algorithmic approach to forecasting rare violent events: An illustration based in intimate partner violence perpetration. *Criminology & Public Policy*, 19(1):213–233, 2020.

[9] R. Biassoni, E. Di Marco, M. Squillario, A. Barla, G. Piccolo, E. Ugolotti, C. Gatti, N. Minuto, G. Patti, M. Maghnie, et al. Gut microbiota in t1dm-onset pediatric patients: machine-learning algorithms to classify microorganisms as disease linked. *The Journal of Clinical Endocrinology & Metabolism*, 105(9):e3114–e3126, 2020.

[10] J. R. Blosnich, J. Cashy, A. J. Gordon, J. C. Shipherd, M. R. Kauth, G. R. Brown, and M. J. Fine. Using clinician text notes in electronic medical record data to validate transgender-related diagnosis codes. *Journal of the American Medical Informatics Association*, 25(7):905–908, 2018.

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

[12] J. C. Campbell. Health consequences of intimate partner violence. *The lancet*, 359(9314):1331–1336, 2002.

[13] J. C. Campbell. Danger Assessment. `https://www.dangerassessment.org/uploads/DA_NewScoring_2019.pdf`, 2003. [Online; accessed 8-October-2021].

[14] J. C. Campbell, D. W. Webster, and N. Glass. The danger assessment: Validation of a lethality risk assessment instrument for intimate partner femicide. *Journal of interpersonal violence*, 24(4):653–674, 2009.

[15] A. Carreño, I. Inza, and J. A. Lozano. Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework. *Artificial Intelligence Review*, 53(5):3575–3594, 2020.

[16] CDC. Intimate Partner Violence. `https://www.cdc.gov/violenceprevention/intimatepartnerviolence/index.html`, 2020. [Online; accessed 29-September-2021].

[17] CDC. Intimate Partner Violence, Sexual Violence, and Stalking Among Men. `https://www.cdc.gov/violenceprevention/intimatepartnerviolence/men-ipvsvandstalking.html`, 2020. [Online; accessed 29-September-2021].

[18] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.

[19] N. S. Chaudhary, J. P. Donnelly, and H. E. Wang. Racial differences in sepsis mortality at united states academic medical center-affiliated hospitals. *Critical care medicine*, 46(6):878, 2018.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[21] I. Y. Chen, E. Alsentzer, H. Park, R. Thomas, B. Gosangi, R. Gujrathi, and B. Khurana. Intimate partner violence and injury prediction from radiology reports. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, pages 55–66. World Scientific, 2020.

[22] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[23] T. Chen and T. He. Higgs boson discovery with boosted trees. In *NIPS 2014 workshop on high-energy physics and machine learning*, pages 69–80. PMLR, 2015.

[24] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *arXiv preprint arXiv:1608.05745*, 2016.

[25] A. L. Coker, K. E. Davis, I. Arias, S. Desai, M. Sanderson, H. M. Brandt, and P. H. Smith. Physical and mental health effects of intimate partner violence for men and women. *American journal of preventive medicine*, 23(4):260–268, 2002.

[26] T. Desautels, J. Calvert, J. Hoffman, M. Jay, Y. Kerem, L. Shieh, D. Shimabukuro, U. Chettipally, M. D. Feldman, C. Barton, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR medical informatics*, 4(3):e5909, 2016.

[27] T. Desautels, J. Hoffman, C. Barton, Q. Mao, M. Jay, J. Calvert, and R. Das. Pediatric severe sepsis prediction using machine learning. *bioRxiv*, page 223289, 2017.

[28] M. Dewan, N. Muthu, E. Shelov, C. P. Bonafide, P. Brady, D. Davis, E. S. Kirkendall, D. Niles, R. M. Sutton, D. Traynor, et al. Performance of a clinical decision support tool to identify picu patients at high-risk for clinical deterioration. *Pediatric critical care medicine: a journal of the Society of Critical Care*

Medicine and the World Federation of Pediatric Intensive and Critical Care Societies, 21(2):129, 2020.

[29] M. Ellsberg, H. A. Jansen, L. Heise, C. H. Watts, C. Garcia-Moreno, et al. Intimate partner violence and women's physical and mental health in the who multi-country study on women's health and domestic violence: an observational study. *The lancet*, 371(9619):1165–1172, 2008.

[30] D. P. Evans, D. Z. Shojaie, K. M. Sahay, N. W. DeSousa, C. D. Hall, and M. A. Vertamatti. Intimate partner violence: barriers to action and opportunities for intervention among health care providers in são paulo, brazil. *Journal of interpersonal violence*, page 0886260519881004, 2019.

[31] L. M. Figueroa-Phillips, C. P. Bonafide, S. E. Coffin, M. E. Ross, and J. P. Guevara. Development of a clinical prediction model for central line-associated bloodstream infection in children presenting to the emergency department. *Pediatric emergency care*, 36(11):e600, 2020.

[32] C. for Disease Control, Prevention, et al. Vital signs: central line–associated blood stream infections—united states, 2001, 2008, and 2009. *Annals of emergency medicine*, 58(5):447–450, 2011.

[33] M. Ford-Gilboe, C. N. Wathen, C. Varcoe, H. L. MacMillan, K. Scott-Storey, T. Mantler, K. Hegarty, and N. Perrin. Development of a brief measure of intimate partner violence experiences: the composite abuse scale (revised)—short form (casr-sf). *BMJ open*, 6(12):e012824, 2016.

[34] S. Fotouhi, S. Asadi, and M. W. Kattan. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of biomedical informatics*, 90:103089, 2019.

[35] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[36] K. H. Goh, L. Wang, A. Y. K. Yeow, H. Poh, K. Li, J. J. L. Yeow, and G. Y. H. Tan. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature communications*, 12(1):1–10, 2021.

[37] H. Han, W.-Y. Wang, and B.-H. Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.

[38] A. K. I. Hassan and A. Abraham. Modeling insurance fraud detection using imbalanced data classification. In *Advances in nature and biologically inspired computing*, pages 117–127. Springer, 2016.

[39] H. He and Y. Ma. *Imbalanced learning: foundations, algorithms, and applications.* Wiley-IEEE Press, 2013.

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

[41] S. Horng, D. A. Sontag, Y. Halpern, Y. Jernite, N. I. Shapiro, and L. A. Nathanson. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PloS one*, 12(4):e0174708, 2017.

[42] H. E. Hsu, F. Abanyie, M. S. Agus, F. Balamuth, P. W. Brady, R. J. Brilli, J. A. Carcillo, R. Dantes, L. Epstein, A. E. Fiore, et al. A national approach to pediatric sepsis surveillance. *Pediatrics*, 144(6), 2019.

[43] K. Huang, J. Altosaar, and R. Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[44] K. Iwasawa, W. Suda, T. Tsunoda, M. Oikawa-Kawamoto, S. Umetsu, L. Takayasu, A. Inui, T. Fujisawa, H. Morita, T. Sogo, et al. Dysbiosis of the salivary microbiota in pediatric-onset primary sclerosing cholangitis and its potential as a biomarker. *Scientific reports*, 8(1):1–10, 2018.

[45] N. Japkowicz. The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence*, volume 56. Citeseer, 2000.

[46] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[47] R. Kamaleswaran, O. Akbilgic, M. A. Hallman, A. N. West, R. L. Davis, and S. H. Shah. Applying artificial intelligence to identify physiomarkers predicting severe sepsis in the picu. *Pediatric Critical Care Medicine*, 19(10):e495–e503, 2018.

[48] A. Khojandi, V. Tansakul, X. Li, R. S. Koszalinski, and W. Paiva. Prediction of sepsis and in-hospital mortality using electronic health records. *Methods of information in medicine*, 57(04):185–193, 2018.

[49] B. Khurana, S. E. Seltzer, I. S. Kohane, and G. W. Boland. Making the 'in-visible'visible: transforming the detection of intimate partner violence. *BMJ quality & safety*, 29(3):241–244, 2020.

[50] M. Kubat, S. Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, volume 97, pages 179–186. Citeseer, 1997.

[51] E. N. Larsen, N. Gavin, N. Marsh, C. M. Rickard, N. Runnegar, and J. Webster. A systematic review of central-line–associated bloodstream infection (clabsi) diagnostic reliability and error. *Infection Control & Hospital Epidemiology*, 40(10):1100–1106, 2019.

[52] S. Le, J. Hoffman, C. Barton, J. C. Fitzgerald, A. Allen, E. Pellegrini, J. Calvert, and R. Das. Pediatric severe sepsis prediction using machine learning. *Frontiers in pediatrics*, 7:413, 2019.

[53] D. E. Leisman, M. O. Harhay, D. J. Lederer, M. Abramson, A. A. Adjei, J. Bakker, Z. K. Ballas, E. Barreiro, S. C. Bell, R. Bellomo, et al. Development and reporting of prediction models: guidance for authors from editors of respiratory, sleep, and critical care journals. *Critical care medicine*, 48(5):623, 2020.

[54] S. Leteurtre, A. Duhamel, J. Salleron, B. Grandbastien, J. Lacroix, F. Leclerc, G. F. de Réanimation et d'Urgences Pédiatriques (GFRUP, et al. Pelod-2: an update of the pediatric logistic organ dysfunction score. *Critical care medicine*, 41(7):1761–1773, 2013.

[55] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier, 1994.

[56] C. Li and S. Liu. A comparative study of the class imbalance problem in twitter spam detection. *Concurrency and Computation: Practice and Experience*, 30(5):e4281, 2018.

[57] R. C. Li, S. M. Asch, and N. H. Shah. Developing a delivery science for artificial intelligence in healthcare. *NPJ digital medicine*, 3(1):1–3, 2020.

[58] H. Liang, B. Y. Tsui, H. Ni, C. C. Valentim, S. L. Baxter, G. Liu, W. Cai, D. S. Kermany, X. Sun, J. Chen, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nature medicine*, 25(3):433–438, 2019.

[59] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object

detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[60] C. X. Ling and C. Li. Data mining for direct marketing: Problems and solutions. In *Kdd*, volume 98, pages 73–79, 1998.

[61] R. J. Little, R. D'Agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, et al. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367(14):1355–1360, 2012.

[62] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

[63] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017.

[64] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1903–1911, 2017.

[65] Q. Mao, M. Jay, J. L. Hoffman, J. Calvert, C. Barton, D. Shimabukuro, L. Shieh, U. Chettipally, G. Fletcher, Y. Kerem, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and icu. *BMJ open*, 8(1):e017833, 2018.

[66] A. J. Masino, M. C. Harris, D. Forsyth, S. Ostapenko, L. Srinivasan, C. P. Bonafide, F. Balamuth, M. Schmatz, and R. W. Grundmeier. Machine learning

models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data. *PloS one*, 14(2):e0212665, 2019.

[67] C. Matava, E. Pankiv, L. Ahumada, B. Weingarten, and A. Simpao. Artificial intelligence, machine learning and the pediatric airway. *Pediatric Anesthesia*, 30(3):264–268, 2020.

[68] P. McDermott, C. Dominguez, N. Kasdaglis, M. Ryan, I. Trhan, and A. Nelson. Human-machine teaming systems engineering guide. Technical report, MITRE CORP BEDFORD MA BEDFORD United States, 2018.

[69] M. R. Miller, M. Griswold, J. M. Harris, G. Yenokyan, W. C. Huskins, M. Moss, T. B. Rice, D. Ridling, D. Campbell, P. Margolis, et al. Decreasing picu catheter-associated bloodstream infections: Nachri's quality transformation efforts. *Pediatrics*, 125(2):206–213, 2010.

[70] U. Nations. Global study on homicide. `https://www.unodc.org/unodc/en/data-and-analysis/global-study-on-homicide.html`, 2019. [Online; accessed 29-September-2021].

[71] S. Nemati, A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, and T. G. Buchman. An interpretable machine learning model for accurate prediction of sepsis in the icu. *Critical care medicine*, 46(4):547, 2018.

[72] H. M. Nguyen, E. W. Cooper, and K. Kamei. Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 3(1):4–21, 2011.

[73] N. P. O'grady, M. Alexander, L. A. Burns, E. P. Dellinger, J. Garland, S. O. Heard, P. A. Lipsett, H. Masur, L. A. Mermel, M. L. Pearson, et al. Guidelines for the prevention of intravascular catheter-related infections. *Clinical infectious diseases*, 52(9):e162–e193, 2011.

[74] W. H. Organization. WHO multi-country study on women's health and domestic violence against women. `https://www.who.int/reproductivehealth/publications/violence/24159358X/en/`, 2005. [Online; accessed 29-September-2021].

[75] J. P. Parreco, A. E. Hidalgo, A. D. Badilla, O. Ilyas, and R. Rattan. Predicting central line-associated bloodstream infections and mortality using supervised machine learning. *Journal of critical care*, 45:156–162, 2018.

[76] R. Petering, M. Y. Um, N. A. Fard, N. Tavabi, R. Kumari, and S. N. Gilani. Artificial intelligence to predict intimate partner violence perpetration. *Artificial intelligence and social work*, page 195, 2018.

[77] E. Pinker. Reporting accuracy of rare event classifiers. *NPJ digital medicine*, 1(1):1–2, 2018.

[78] M. M. Pollack, K. M. Patel, and U. E. Ruttimann. Prism iii: an updated pediatric risk of mortality score. *Critical care medicine*, 24(5):743–752, 1996.

[79] Y. Raita, C. A. Camargo, C. G. Macias, J. M. Mansbach, P. A. Piedra, S. C. Porter, S. J. Teach, and K. Hasegawa. Machine learning-based prediction of acute severity in infants hospitalized for bronchiolitis: a multicenter prospective study. *Scientific reports*, 10(1):1–11, 2020.

[80] B. Renaud and C. Brun-Buisson. Outcomes of primary and catheter-related bacteremia: a cohort and case–control study in critically ill patients. *American journal of respiratory and critical care medicine*, 163(7):1584–1590, 2001.

[81] L. M. Renner, Q. Wang, M. E. Logeais, and C. J. Clark. Health care providers' readiness to identify and respond to intimate partner violence. *Journal of interpersonal violence*, page 0886260519867705, 2019.

[82] M. A. Reyna, C. Josef, S. Seyedi, R. Jeter, S. P. Shashikumar, M. B. Westover, A. Sharma, S. Nemati, and G. D. Clifford. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. In *2019 Computing in Cardiology (CinC)*, pages Page–1. IEEE, 2019.

[83] C. Rhee, R. Dantes, L. Epstein, D. J. Murphy, C. W. Seymour, T. J. Iwashyna, S. S. Kadri, D. C. Angus, R. L. Danner, A. E. Fiore, et al. Incidence and trends of sepsis in us hospitals using clinical vs claims data, 2009-2014. *Jama*, 318(13):1241–1249, 2017.

[84] C. Rhee, R. B. Dantes, L. Epstein, and M. Klompas. Using objective clinical data to track progress on preventing and treating sepsis: Cdc's new 'adult sepsis event'surveillance strategy. *BMJ quality & safety*, 28(4):305–309, 2019.

[85] C. Rhee, S. Kadri, S. S. Huang, M. V. Murphy, L. Li, R. Platt, and M. Klompas. Objective sepsis surveillance using electronic clinical data. *infection control & hospital epidemiology*, 37(2):163–171, 2016.

[86] M. E. Rupp and D. Majorant. Prevention of vascular catheter-related bloodstream infections. *Infectious Disease Clinics*, 30(4):853–868, 2016.

[87] M. Saqib, Y. Sha, and M. D. Wang. Early prediction of sepsis in emr records using traditional ml techniques and deep learning lstm networks. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4038–4041. IEEE, 2018.

[88] J. Schaefer, M. Lehne, J. Schepers, F. Prasser, and S. Thun. The use of machine learning in rare diseases: a scoping review. *Orphanet Journal of Rare Diseases*, 15(1):1–10, 2020.

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

[90] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

[91] S. P. Shashikumar, C. Josef, A. Sharma, and S. Nemati. Deepaise–an end-to-end development and deployment of a recurrent neural survival model for early prediction of sepsis. *arXiv preprint arXiv:1908.04759*, 2019.

[92] E. K. Shin, R. Mahajan, O. Akbilgic, and A. Shaban-Nejad. Sociomarkers and biomarkers: predictive modeling in identifying pediatric asthma patients at risk of hospital revisits. *NPJ digital medicine*, 1(1):1–5, 2018.

[93] Y. Si, J. Wang, H. Xu, and K. Roberts. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304, 2019.

[94] M. A. Straus. Conflict Tactics Scale. `http://rzukausk.home.mruni.eu/wp-content/uploads/Conflict-Tactics-Scale.pdf`, 2010. [Online; accessed 8-October-2021].

[95] L. Sung, C. Corbin, E. Steinberg, E. Vettese, A. Campigotto, L. Lecce, G. A. Tomlinson, and N. Shah. Development and utility assessment of a machine learning bloodstream infection classifier in pediatric patients receiving cancer treatments. *BMC cancer*, 20(1):1–9, 2020.

[96] A. Tabaie, S. Nemati, J. W. Allen, C. Chung, F. Queiroga, W.-J. Kuk, and A. B. Prater. Assessing contribution of higher order clinical risk factors to prediction of outcome in aneurysmal subarachnoid hemorrhage patients. In *AMIA Annual Symposium Proceedings*, volume 2019, page 848. American Medical Informatics Association, 2019.

[97] A. Tabaie, E. W. Orenstein, S. Nemati, R. K. Basu, G. D. Clifford, and R. Kamaleswaran. Deep learning model to predict serious infection among children with central venous lines. *Frontiers in Pediatrics*, page 910.

[98] A. Tabaie, E. W. Orenstein, S. Nemati, R. K. Basu, S. Kandaswamy, G. D. Clifford, and R. Kamaleswaran. Predicting presumed serious infection among hospitalized children on central venous lines with machine learning. *Computers in Biology and Medicine*, 132:104289, 2021.

[99] N. Thai-Nghe, Z. Gantner, and L. Schmidt-Thieme. Cost-sensitive learning methods for imbalanced data. In *The 2010 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2010.

[100] P. J. Thoral, M. Fornasa, D. P. de Bruin, H. Hovenkamp, R. H. Driessen, A. R. Girbes, M. Hoogendoorn, and P. W. Elbers. Developing a machine learning prediction model for bedside decision support by predicting readmission or death following discharge from the intensive care unit. 2020.

[101] J. Todahl and E. Walters. Universal screening for intimate partner violence: A systematic review. *Journal of marital and family therapy*, 37(3):355–369, 2011.

[102] K. Tollestrup, D. Sklar, F. J. Frost, L. Olson, J. Weybright, J. Sandvig, and M. Larson. Health indicators and intimate partner violence among women who are members of a managed care organization. *Preventive medicine*, 29(5):431–440, 1999.

[103] E. E. Tress, P. M. Kochanek, R. A. Saladino, and M. D. Manole. Cardiac arrest in children. *Journal of Emergencies, Trauma and Shock*, 3(3):267, 2010.

[104] F. Van Wyk, A. Khojandi, and R. Kamaleswaran. Improving prediction performance using hierarchical analysis of real-time data: a sepsis case study. *IEEE journal of biomedical and health informatics*, 23(3):978–986, 2019.

[105] N. C. A. D. Violence. Domestic violence. `https://assets.speakcdn.com/assets/2497/domestic_violence-2020080709350855.pdf?1596811079991`, 2020. [Online; accessed 29-September-2021].

[106] L. W. Walker, A. J. Nowalk, and S. Visweswaran. Predicting outcomes in central venous catheter salvage in pediatric central line–associated bloodstream infection. *Journal of the American Medical Informatics Association*, 28(4):862–867, 2021.

[107] S. L. Weiss, F. Balamuth, M. Chilutti, M. J. Ramos, P. McBride, N.-A. Kelly, K. J. Payton, J. C. Fitzgerald, and J. W. Pennington. Identification of pediatric sepsis for epidemiologic surveillance using electronic clinical data. *Pediatric critical care medicine: a journal of the Society of Critical Care Medicine and the World Federation of Pediatric Intensive and Critical Care Societies*, 21(2):113, 2020.

[108] S. L. Weiss, J. C. Fitzgerald, J. Pappachan, D. Wheeler, J. C. Jaramillo-Bustamante, A. Salloo, S. C. Singhi, S. Erickson, J. A. Roy, J. L. Bush, et al. Global epidemiology of pediatric severe sepsis: the sepsis prevalence, outcomes, and therapies study. *American journal of respiratory and critical care medicine*, 191(10):1147–1157, 2015.

[109] B. Yang, J.-T. Sun, T. Wang, and Z. Chen. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 917–926, 2009.

[110] D. Zhang, J. Thadajarassiri, C. Sen, and E. Rundensteiner. Time-aware transformer-based network for clinical notes series prediction. In *Machine Learning for Healthcare Conference*, pages 566–588. PMLR, 2020.

[111] G. Zhang, J. Xu, M. Yu, J. Yuan, and F. Chen. A machine learning approach for

mortality prediction only using non-invasive parameters. *Medical & Biological Engineering & Computing*, pages 1–44, 2020.

[112] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.