Surgery Open Science 16 (2023) 77–81

Contents lists available at ScienceDirect

Surgery Open Science

journal homepage: www.journals.elsevier.com/surgery-open-science

Research Paper

Children are small adults (when properly normalized): Transferrable/generalizable sepsis prediction

Caitlin Marassi, BA, Damien Socia, MS, Dale Larie, BS, Gary An, MD, R. Chase Cockrell, PhD [*]

Department of Surgery, University of Vermont, 89 Beaumont Ave, Given D319, Burlington, VT 05405, United States of America

ARTICLE INFO

Keywords:
Machine learning
Sepsis
Pediatric sepsis

ABSTRACT

Background: Though governed by the same underlying biology, the differential physiology of children causes the temporal evolution from health to a septic/diseased state to follow trajectories that are distinct from adult cases. As pediatric sepsis data sets are less readily available than for adult sepsis, we aim to leverage this shared underlying biology by normalizing pediatric physiological data such that it would be directly comparable to adult data, and then develop machine-learning (ML) based classifiers to predict the onset of sepsis in the pediatric population. We then externally validated the classifiers in an independent adult dataset.

Methods: Vital signs and laboratory observables were obtained from the Pediatric Intensive Care (PIC) database. These data elements were normalized for age and placed on a continuous scale, termed the Continuous Age-Normalized SOFA (CAN-SOFA) score. The XGBoost algorithm was used to classify pediatric patients that are septic. We tested the trained model using adult data from the MIMIC-IV database.

Results: On the pediatric population, the sepsis classifier has an accuracy of 0.84 and an F1-Score of 0.867. On the adult population, the sepsis classifier has an accuracy of 0.80 and an F1-score of 0.88; when tested on the adult population, the model showed similar performance degradation ("data drift") as in the pediatric population.

Conclusions: In this work, we demonstrate that, using a straightforward age-normalization method, EHR's can be generalizable compared (at least in the context of sepsis) between the pediatric and adult populations.

Introduction

Sepsis is a pathological manifestation of the body's acute inflammatory response to infection and injury, and, despite decades of research, continues to have a significant mortality [1–3]. Specifically, pediatric sepsis has a significant health impact world-wide [4], with well-defined differences in the clinical trajectories seen in these patients compared to adults [5,6]. In recent years, machine learning (ML) has been increasingly employed to improve risk prediction for sepsis, with variable success [4,5]. However, a persistent issue with ML-sepsis prediction is the problem of data drift, where the eventual application population has different statistical distributions than the training population/data, and the inevitable performance of these systems over time. While retraining has been proposed as a maintenance strategy for these ML systems after deployment [6], the need to do retraining intrinsically limits the utility of such systems in a mission-critical intensive care setting. With respect to pediatric sepsis, the fact that there are fewer extensive data sets of these patients, compared to adult sepsis, accentuates the limitations of ML predictive algorithms, being more subject to brittleness, overfitting, and a failure to generalize. Our goal is to augment the performance of ML prediction of pediatric sepsis by increasing the amount of available ML training data by utilizing the totality of available sepsis datasets (e.g., pediatric and adult cases). We recognize that accomplishing this will require developing a normalization process to allow direct comparisons between the disease trajectories between these groups. While it may seem counter-intuitive to increase the variability and heterogeneity of a training set, we believe that a normalization process can uncover a shared fundamental biology between adults and children, and in so doing actually improve the generalizability of a ML algorithm trained on such data. Therefore we aim to demonstrate this enhanced generalizability by cross-testing the trained ML algorithm in what are generally considered very distinct populations (i.e. pediatric versus adult sepsis).

As we are interested in evaluating the clinical courses of disease, we look to the Sequential Organ Failure Assessment (SOFA) score as a means of categorizing the progressive organ dysfunction that is seen in sepsis. The SOFA score is a well-established and accepted metric that has been used to quantify the degree of organ dysfunction in sepsis and

[*] Corresponding author at: 89 Beaumont Ave, Given D319, Burlington, VT 05405, United States of America.
E-mail address: Robert.cockrell@med.uvm.edu (R.C. Cockrell).

https://doi.org/10.1016/j.sopen.2023.09.013
Received 16 June 2023; Received in revised form 29 August 2023; Accepted 17 September 2023
Available online 22 September 2023

provide prognostic information in these patients [7–9]. Though initially developed as a means of establishing in-hospital mortality, there has been a natural evolution towards examining trajectories of SOFA scores as a potential means of clinically phenotyping patients into different risk categories [10]. However, by its very nature and intent of being relatively simple to calculate, the SOFA score necessarily aggregates wide ranges of input variables that can represent different and progressive degrees of physiological derangement into a single value. The impact of the step-function nature of the SOFA score manifests in two ways: 1) small differences at the boundary between two SOFA categories would imply a greater physiological derangement than is actually present, and 2) large differences that are encompassed within the range of a SOFA category are not detected at all.

We adopt a strategy that utilizes fundamental properties of the SOFA score to address the limitations of ML prediction in sepsis. Our basic rationale is that the SOFA score quantifies the degree of physiological derangement in sepsis as a deviance from baseline health; however, the step-function/discretized nature of the SOFA score obscures differences in physiological derangement. Nevertheless, despite its aggregating limitations (or possibly because of it), the SOFA score remains a reliable metric for quantifying the degree of physiological derangement across very divergent patient populations. For instance, a the numerical value of a SOFA score in the adult population is considered to be equivalently deranged as the same value in the Pediatric SOFA (pSOFA) score in the pediatric population [11–13] (in fact, this underlying assumption is present in the development of the pSOFA score). Therefore, the SOFA score (within its varied forms), represents a robust and generalizable quantification of the degree of physiological derangement seen in sepsis. Additionally, there is a continuum of measured values and associated levels of physiological derangement that is present within each category of the SOFA score, with the reasonable assumption that worsening values reflect worsening derangement, that is not reflected by the actual SOFA score. A fundamental limitation at present of ML prediction in sepsis is that the specific values of the incorporated data points will have varying statistical distributions within a specific data set, and these differences lead to data drift when applying external or longitudinal validation. In essence, the ML algorithms learn a limited physiological "truth" present in the training set that may not apply to the specifics of a different data set. Thus, while the SOFA score does apply a generalizable metric for quantifying physiological derangement, in of itself it is too coarse to allow for sufficient discrimination between data points necessary for modern ML approaches. Therefore, we make the logical extrapolation and assume that there is, at least as a first approximation, a linear progression of the component measurements used to calculate the SOFA score where the greater deviance from normal values reflect greater physiological derangement: we term this transform the Continuous Age-Normalized SOFA (CAN-SOFA) score.

The continuous aspect of the CAN-SOFA can address more nuanced alterations in the patients' condition. The normalization aspect of the CAN-SOFA is predicated on the recognition that baseline "normal" measurements of physiologic parameters can vary based on age, with a consequent impact on how "deviations" might present in the pediatric population. Therefore, we incorporate an age-based normalization process to allow comparison between pediatric and adult physiological parameters. These steps allow us to use the CAN-SOFA score as a means of normalizing and quantifying the degree of physiological derangements across populations such that ML can be applied.

Herein we present an example of using the CAN-SOFA score to evaluate: 1) the efficacy of the normalization process by determining whether training on a pediatric sepsis cohort could generalize to an adult sepsis population, and then 2) determining if the generalizing effect of the CAN-SOFA could effectively use training on an adult sepsis cohort to effectively classify pediatric sepsis patients. Note that this represents two distinct training and testing tasks.

Methods

Data sources

Generation of a Sepsis classifier and testing its generalizability required the use of two datasets. The data used for the training and evaluation of the Sepsis classifier came from the Pediatric Intensive Care Database (PIC) [14]. This database is a pediatric specific dataset collected from Children's Hospital of Zhejiang University School of Medicine in China. From the dataset all the patient time-points with at least one lab and vital sign data element were used.

This resulted in 12,749 patients with a total of 111,532 data points. Each data point contains data that was collected at the same time. This resulted in a sparse dataset due to vitals, and labs being recorded at different frequencies. To reduce the amount of missing data, each patient's data was aggregated by hour based on admission time. For columns with multiple entries per hour the mean was taken. This resulted in 37,558 data points. For each data point a label was given, Septic or Non-septic, this was the classification label that our model predicted. To assign these labels the ICD_10 discharge code and the ICU admission time were used. The ICD_10 code represents the diagnoses and procedures, using these codes patients were classified as having sepsis while hospitalized, of these 12,749 patients, 296 had sepsis. The issue with just using the ICD_10 code is that there is no time of diagnosis. This caused all the data of patients discharged with an ICD_10 sepsis code to be marked as septic. To address this problem ICU transfer time was used. Using this time point, all the data points before ICU were classified as not septic and the data after were classified as Septic. This resulted in 19,781 data points being classified as Septic. We recognize that using the combination of ICD_10 code to determine the presence of sepsis and the ICU transfer time to define the onset of sepsis is controversial and does not rigorously meet the Sepsis-3 definition [15]; however, we choose to utilize this approximation as publicly available databases do not contain sufficiently granular information to meet this definition without the use of data imputation techniques. Further, as the focus of this paper is on the pre-processing of EHR data such that models informed by the preprocessed data are maximally generalizable, we posit that the use of our consistent definition among the different populations (children and adults) is an excellent exemplar as we do not have to rely on synthetic or imputed data.

To test the generalizability of the classifier The Medical Information Mart for Intensive Care (MIMIC) [16], an adult dataset, was used. The same data collection and cleaning process was used except for one key difference. Due to only having ICU data in MIMIC, data from 10,000 separate non-septic patients was collected to be used as the non-septic data points. This resulted in 970,627 data points after aggregating on the hour with 70,272 being non-septic and 900,355 being septic.

The main difference between the two datasets is the age range they contain with PIC containing patients under 18 and MIMIC containing patients over 18. This results in Creatinine, Total Bilirubin, C Reactive

Table 1
Data features used to inform that machine-learning classifier.

Category	Measurement
Vital	pO2
Vital	Oxygen saturation
Vital	Temperature
Vital	Respiratory rate
Vital	Pulse
Vital	Mean arterial pressure
Lab	Fshunt
Lab	C reactive protein
Lab	Total bilirubin
Lab	Indirect bilirubin
Lab	Creatinine

Protein, Pulse, Respiratory Rate and Mean Arterial Pressure having different ranges between datasets.

Continuous/age-normalizing transformation

The numerical value ranges for the SOFA and pSOFA scores were normalized using an assumption of a linear progression of those values and the reflected physiological derangement. The complete set of variables that were used in this work are shown in Table 1. A subset of the formulas for the generation of the Continuous SOFA score can be seen in Table 2. We present an example for a specific patient in Table 3, in which the first column shows the variable of interest, the second column shows the raw value, and the third column determines the continuous SOFA score for metrics that are associated with SOFA scores or a deviance score for other metrics. Variables marked with an asterisk are not converted to deviance scores as they are equally informative for children and adults and are simply normalized at the time of training. Variables that have been bolded were transformed using the continuous SOFA transformations. Details regarding the full transformation can be found in the supplementary material. We note that we include additional variables not used in the SOFA score: Pulse, Respiratory Rate, and C Reactive Protein.

Machine learning algorithm

In general, the purpose of an ML algorithm is to discover a complex pattern in data that would be otherwise occluded due to the nature of the data (i.e., dimensionality, missingness, etc.). To address the problem of classifying patients as septic based on their vitals and labs, an ensemble method was chosen, specifically gradient-boosted trees. Ensemble methods are a machine learning technique that uses many distinct classifiers or regressors and combines their results by either averaging or taking the majority. This results in a model with better performance than its individual components. The gradient-boosted trees algorithm XGBoost [17] was chosen due to its ability to deal with sparse data, as described below. To illustrate how the XGBoost algorithm works, we provide an explanatory diagram in Fig. 1. In this diagram, each node of the decision tree contains a mathematical operation, or 'decision,' labeled $\mathbf{DX_y}$, in which X represents the sequence in which the decision is operated upon and y represents an individual tree model in the ensemble. Results are indicated by the **A's**. In this work, the result is a prediction of whether or not that patient will evolve into a state of sepsis in the course of their hospitalization. The mathematical operations represented by the decisions could be something like, 'Is the normalized renal sofa score>2?' with the answer then informing subsequent decisions. One explicit strength of the XGBoost algorithm is its ability to deal with missing (i.e., NaN) data in the training datasets by using the presence of missing data for a specific variable as a node in the decision tree, for example, if there is no data on serum creatinine, then other variables may be more informative towards the final prediction. Ultimately, the results from all of the individual tree model are aggregated together to give the final result (Fig. 2).

For training of the model, the data was randomly split with 80 % of the data going to the training set and 20 % going to validation. To accommodate for the imbalance in the data, a class weighting was applied. This was done for each class using the equation below:

$$weightX = numberOfSamples/(numberOfClasses*numberOfSamplesInClassX)$$

The models were evaluated based on their accuracy and F1 score. Multiple models were trained to tune XGBoost's hyperparameters.

Results

The model was initially trained on the PIC data achieving an accuracy of 0.84. Then all the MIMIC data was fed to the algorithm to assess

Table 2
Continuous age-normalization for serum creatinine: the transformation for serum creatinine concentrations is presented here as a function of age and raw value, e.g., for a 15-month old patient with a serum creatinine level of 1.2, the value for the renal component of the CAN-SOFA score would be 3.33, as opposed to standard pediatric SOFA score of 3.

Creatinine:										
Age	Value	Formula	Value	Formula	Value	Formula	Value	Formula	Value	Formula
<1 mo.	<0.8 mg/dL	0	0.8–0.9	1 + (C-0.8) * 10	1.0–1.1	2 + (C-1.0) * 10	1.2–1.5	3 + (C-1.2) * (10/3)	>1.6	MIN(5,4 + C-1.6)
1–11 mo.	<0.3	0	0.3–0.4	1 + (C-0.3) * 10	0.5–0.7	2 + (C-0.5) * 5	0.8–1.1	3 + (C-0.8) * (10/3)	>1.2	MIN(5,4 + C-1.2)
12–23 mo.	<0.4	0	0.4–0.5	1 + (C-0.4) * 10	0.6–1.0	2 + (C-0.6) * 2.5	1.1–1.4	3 + (C-1.1) * (10/3)	>1.5	MIN(5,4 + C-1.5)
24–59 mo.	<0.6	0	0.6–0.8	1 + (C-0.6) * 5	0.9–1.5	2 + (C-0.9) * (5/3)	1.6–2.2	3 + (C-1.6) * (5/3)	>2.3	MIN(5,4 + C-2.3)
60–143 mo.	<0.7	0	0.7–1.0	1 + (C-0.7) * (10/3)	1.1–1.7	2 + (C-1.1) * 2	1.8–2.5	3 + (C-1.8) * (10/7)	>2.6	MIN(5,4 + C-2.6)
144–216 mo.	<1.0	0	1.0–1.6	1 + (C-1) * (5/3)	1.7–2.8	2 + (C-1.7) * (10/11)	2.9–4.1	3 + (C-2.9) * (5/6)	>4.2	MIN(5,4 + C-4.5)
>18 yr.	<1.2	0	1.2–1.9	1 + (C-1.2) * (10/7)	2.0–3.4	2 + (C-2.0) * (5/7)	3.5–4.9	3 + (C-3.5) * (5/7)	>5	MIN(5,4 + (C-5)/10)

Table 3
Example patient data: in the first column, we present a list of variables used in the ML model; in the second column, we present the raw value for that variable, with NaN indicating missing data for that patient; in the third column, we present the continuous age-normalized score. Variables marked with an asterisk are not converted to deviance scores as they are equally informative for children and adults and are simply normalized at the time of training. Variables that have been bolded were transformed using the continuous SOFA transformations.

	Raw value	Age-normalized score
Age	1.1 yrs	1.1 yrs
C Reactive protein_mean	**1.86**	**2.32**
Creatinine_mean	**1.58484163**	**4.28280543**
Fshunt_mean*	5.7	5.7
MAP_max	**53**	**1.12**
MAP_mean	**53**	**1.12**
MAP_median	**53**	**1.12**
MAP_min	**53**	**1.12**
Oxygen Saturation_max*	100	100
Oxygen Saturation_mean*	98.4	98.4
Oxygen Saturation_median*	98.4	98.4
Oxygen Saturation_min*	96.8	96.8
Pulse_max	**NaN**	**NaN**
Pulse_mean	**NaN**	**NaN**
Pulse_median	**NaN**	**NaN**
Pulse_min	**NaN**	**NaN**
Pulse_std	**NaN**	**NaN**
Respiratory Rate_max	**70**	**0.0008329**
Respiratory Rate_mean	**69**	**0.00074961**
Respiratory Rate_median	**69**	**0.00074961**
Respiratory Rate_min	**68**	**0.00066632**
Temperature_max*	37	37
Temperature_mean*	36.8333333	36.8333333
Temperature_median*	37	37
Temperature_min*	36.5	36.5
Total Bilirubin_mean*	14.55	5
White Blood Cells_mean*	NaN	NaN
p02_max*	173	173
p02_mean*	124	124
p02_median*	124	124
p02_min*	75	75
paCO2_mean*	42.45	42.45

the model's ability to generalize to another dataset. The classifier achieved an accuracy of 0.80 (Industry Standard for Good Performance is between 70 % and 90 %, with >90 % being Very Good [18]). The reverse was then done, training the model on the MIMIC data and testing on the PIC data, to see if it was possible to leverage more readily available adult sepsis data to inform classification in a pediatric context. To test this the model was trained on a subset of the MIMIC dataset to match the size of the PIC dataset with the data points being selected at random. After the classifier was retrained it had an accuracy of 0.95 on the MIMIC test set and an accuracy of 0.77 on the PIC data.

Discussion

We demonstrate a novel interpretation of the SOFA score, the CAN-SOFA score, as means of normalizing and quantifying physiological derangement in sepsis that allows the training of a robust ML sepsis prediction algorithm that generalizes from pediatric patients to adult patients. This approach is novel as it incorporates the knowledge present in a well-established sepsis scoring system that has already demonstrated efficacy in generalizing across patient populations considered to be widely divergent. This pre-processing of data provides a degree of interpretability and generalizability often missing in standard ML approaches.

There have been other attempts to reconcile the step function nature of the SOFA score with the fact that the component values are more continuous, most notably, DeepSOFA [19]. A significant difference between CAN-SOFA and DeepSOFA is that the underlying methodology of DeepSOFA, seeking a "hidden" complex function that associates the time series measurements, and is subject to exactly the same limitations of brittleness, overfitting and need for retraining present in standard ML sepsis prediction. Further, we note that, while CAN-SOFA was only evaluated using data from two institutions (PIC and MIMIC), as was the case with DeepSOFA, CAN-SOFA was trained and tested on two distinct populations (pediatric vs adult), demonstrating that the pre-processing/normalization of the data increases the overall generalizability of the method.

Additionally, in contrast with other studies using machine learning to predict or detect sepsis [20], we did not use the Sepsis-3 [15] definition for sepsis, rather we used ICD diagnosis codes present in the electronic health record (EHR). The reasoning for this is that, in order to meet this definition, the patient must maintain vital signs and laboratory observables outside of the normal regime for at least 5 h; in our databases, no patient EHRs contain sufficient data to determine this. Other studies [19,20] have used data imputation techniques to fill out missing data, though choice of imputation technique can significantly affect the reproducibility, generalizability, and accuracy of results [21]; as such, we chose to explore a technique that did not rely on explicit data imputation. Instead, we utilized the diagnosis code that reflects a combination of the diagnosing clinician's expertise (including awareness of the Sepsis-3 guidelines) and known sepsis-identifier present in EHR systems.

A key point in the development and deployment of CAN-SOFA is that some basic interpretation of the properties of the underlying data can

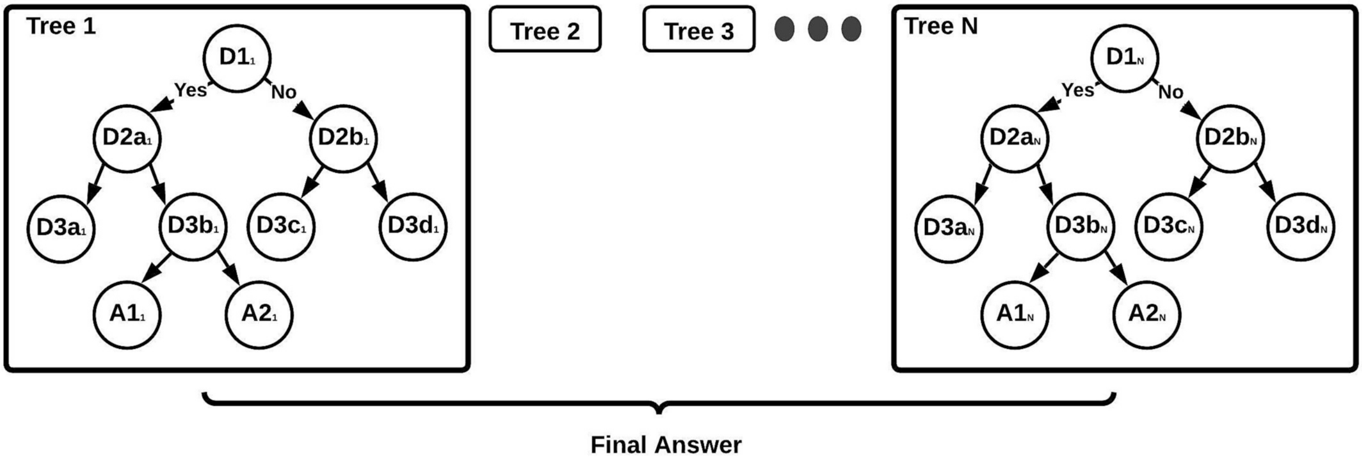

Fig. 1. Explanatory diagram of the XGBoost algorithm. Each node of the decision tree contains a mathematical operation, or 'decision,' labeled **DX$_y$**, in which X represents the sequence in which the decision is operated upon and y represents an individual tree model in the ensemble. Results are indicated by the **A's**.

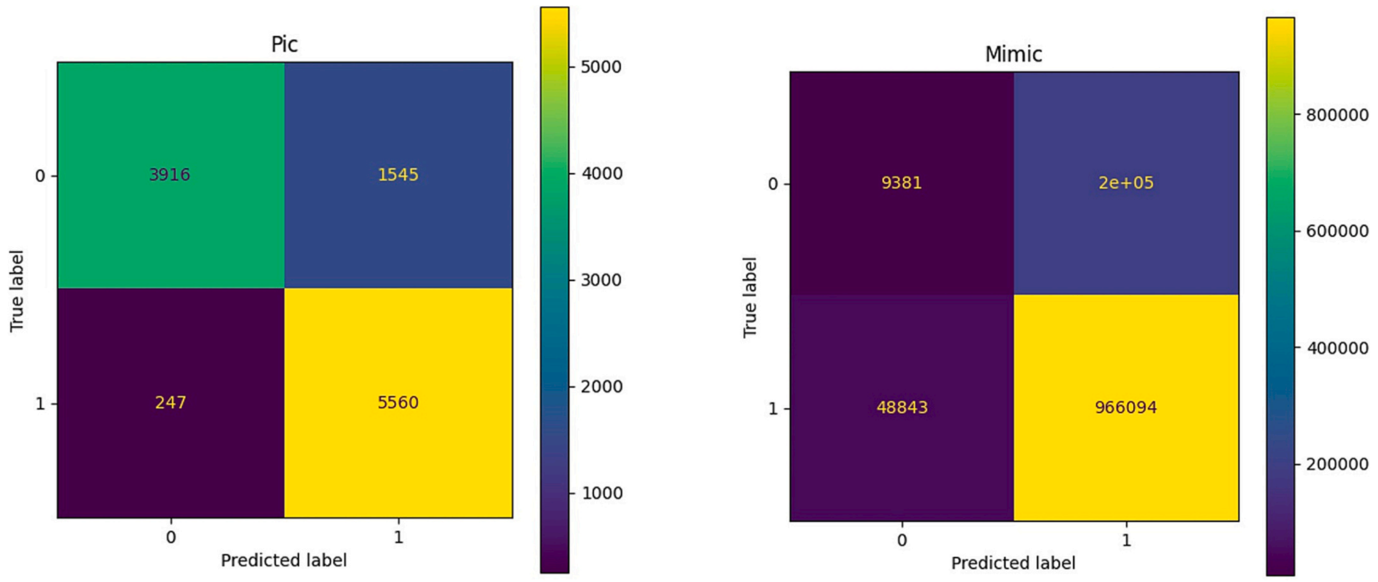

Fig. 2. Panel on Left: Confusion matrix showing performance of model on internal validation on PIC (Accuracy 0.84). Panel on Right: Confusion matrix showing performance of external validation of model trained on pediatric data on adult data/MIMIC (Accuracy 0.80).

provide considerable assistance in overcoming the intrinsic limitations of ML methods. In fact, this approach is increasingly being recognized in the general artificial intelligence community, where the limitations of deep learning methods are becoming more evident even as those systems reach higher levels of performance.

CRediT authorship contribution statement

CM, DS, and DL annotated data, implemented machine learning algorithms, and contributed to the manuscript, GA contributed to the manuscript and provided clinical guidance to the development and interpretation of the machine-learning techniques, RCC conceived of the project, directed the machine learning tasks, and contributed to the manuscript.

Funding sources

This work was funded by internal funds from the Larner College of Medicine at the University of Vermont.

Ethical approval

This work required no ethics approval as all data was deidentified.

Declaration of competing interest

The authors report no proprietary or commercial interest in any product mentioned or concept discussed in this article.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.sopen.2023.09.013.

References

[1] Barriere SL, Lowry SF. An overview of mortality risk prediction in sepsis. Crit Care Med 1995;23(2):376–93.

[2] Carcillo JA, et al. A systemic inflammation mortality risk assessment contingency table for severe sepsis. Pediatr Crit Care Med 2017;18(2):143.

[3] Zhang K, et al. Development and validation of a sepsis mortality risk score for sepsis-3 patients in intensive care unit. Front Med 2021;7:609769.

[4] Nemati S, et al. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med 2018;46(4):547.

[5] Fleuren LM, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 2020;46: 383–400.

[6] Rahmani K, et al. Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. Int J Med Inform 2022:104930.

[7] Lambden S, et al. The SOFA score—development, utility and challenges of accurate assessment in clinical trials. Crit Care 2019;23(1):1–9.

[8] Ferreira FL, et al. Serial evaluation of the SOFA score to predict outcome in critically ill patients. Jama 2001;286(14):1754–8.

[9] Innocenti F, et al. SOFA score in septic patients: incremental prognostic value over age, comorbidities, and parameters of sepsis severity. Intern Emerg Med 2018;13: 405–12.

[10] Lie KC, et al. Utility of SOFA score, management and outcomes of sepsis in Southeast Asia: a multinational multicenter prospective observational study. J Intensive Care 2018;6:1–8.

[11] Aulia M, et al. Pediatric SOFA score for detecting sepsis in children. Paediatr Indones 2021;61(1):1–7.

[12] Matics TJ, Sanchez-Pinto LN. Adaptation and validation of a pediatric sequential organ failure assessment score and evaluation of the sepsis-3 definitions in critically ill children. JAMA Pediatr 2017;171(10):e172352.

[13] Lalitha A, et al. Sequential organ failure assessment score as a predictor of outcome in sepsis in pediatric intensive care unit. J Pediatr Intensive Care 2021;10(02): 110–7.

[14] Zeng X, et al. PIC, a paediatric-specific intensive care database. Sci Data 2020;7(1): 14.

[15] Singer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). Jama 2016;315(8):801–10.

[16] Johnson AE, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3(1):160035.

[17] Chen T, et al. Xgboost: extreme gradient boosting. R package version 0.4-21(4); 2015. p. 1–4.

[18] Bertolini M, et al. Machine learning for industrial applications: a comprehensive literature review. Expert Syst Appl 2021;175:114820.

[19] Shickel B, et al. DeepSOFA: a continuous acuity score for critically ill patients using clinically interpretable deep learning. Sci Rep 2019;9(1):1879.

[20] Scherpf M, et al. Predicting sepsis with a recurrent neural network using the MIMIC III database. Comput Biol Med 2019;113:103395.

[21] Kamble V, Deshmukh S. Comparision between accuracy and MSE, RMSE by using proposed method with imputation technique. Orient J Comput Sci Technol 2017; 10(4):773–9.