Machine Learning Approaches for Patient State Prediction in Pediatric ICUs

Muhammad Aurangzeb Ahmad
KenSci Inc
Department of Computer Science
University of Washington Bothell
Seattle, WA, USA
maahmad@uw.edu

Eduardo Antonio Trujillo Rivera
Children's National Medical Center
School of Medicine and Health Sciences
George Washington University
Washington D.C., USA
eduardotrujillo@email.gwu.edu

Murray Pollack M.D.
Children's National Medical Center
Washington D.C., USA
mpollack@childrensnational.org

Carly Eckert M.D.
KenSci Inc
Seattle, WA, USA
carly@kensci.com

Anita Patel M.D.
Children's National Medical Center
Washington D.C., USA
apatel4@childrensnational.org

Ankur Teredesai
KenSci Inc
Department of Computer Science
University of Washington Tacoma
Seattle, WA, USA
ankurt@uw.edu

***Abstract*—We consider the problem of characterizing and predicting the condition of pediatric patients in intensive care units (ICUs). This population is often typified by rapid changes in patient conditions which necessitate predictions that can capture transition in patient states. While the assessment of patient's condition is currently usually done using domain based scoring systems, we employ machine learning models for predicting the state of the pediatric patient. Additionally, we explore how model explainability could affect the usage of predictive models in a real world settings.**

***Index Terms*—Pediatric ICU, Pediatric Patient Deterioration, Patient Deterioration**

Clinical deterioration is the often rapid change in the condition of a patient. A delay in recognizing the change in patient condition can lead to worse clinical outcomes, decline in quality of life and in extreme cases, even death. Consequently, better prediction of clinical deterioration is a priority as many patients today get harmed when precursors go unrecognized, leading to potentially preventable morbidity, mortality, and cost. Providing care to very seriously sick or injured children may require intensive care. While less than 2% of pediatric patients require intensive care services, those that do typically experience significant morbidity. Reducing this morbidity is highly contingent on the timely recognition of changing clinical needs. In the hospital setting, errors and delays in beginning effective therapies contribute to patient morbidity. Better, more accurate predictions of a patient's criticality of illness, supplied via machine learning algorithms that take time into account, could save valuable lives.

Patients in ICUs require high level of care as compared to patients in other units in the hospital. The condition of certain patients can deteriorate quickly which require sudden attention from the medical staff. This can often times prove challenging especially if risk stratification is needed for resource allocation for the medical staff, a problem that has especially become prominent in the post COVID-19 world. To support this, we propose the use of machine learning models that can be used to characterize patient state (i.e., the relative condition of pediatric patients), predict future states of patients (being in ICU, not in ICU or in transition) and algorithmically generate explanations for why the patient is in the ICU. The insights generated from these models could be used for patient prioritization and care optimization.

Several factors make such a risk prediction among critically ill pediatric patients even more challenging. Normal vital signs ranges change with age, and thus must be compared to age appropriate referent values. Pediatric patients generally have highly effective physiologic compensatory mechanisms, which can obfuscate early markers of hemodynamic instability. Finally, while children have good physiologic compensation, they have smaller reserves, which means that decomposition can progress from subtle to overt quickly. In this paper, we consider the problem of criticality of illness as a key concept. **Criticality** is the combination of physiologic variables and therapeutic intensity. We present preliminary results from a project where the long term objective is to establish criticality as a conceptual framework for severity of illness by predicting that patient's current care location using physiologic variables and therapies.

Interpretable machine learning models have become an integral part of creating responsible and accountable machine learning systems [1]. Since the long term goal of the current work is to create an adaptable machine learning based scoring system to predict patient risk of mortality in pediatric ICUs, it is important to know how the machine learning models work as it would be necessary to determine what factors are contributing to the prediction. This would enable the healthcare personnel to make quick informed decisions and also improve upon current risk scoring systems. The main contrition of the current work is to address the problem of criticality in the context of patient risk prediction in pediatric

Category	Feature Example	Total Features
Demographics	Race, Gender, Age	16
Temporal	Sliced Hour, Stay Length	3
Labs	WBC, Bun, Calcium	434
Medications	Antidepressants, Vitamins etc.	580
Vitals	Heart Rate, Temperature, Systolic	98
Total		1,131

TABLE I
SUMMARY OF MAXIMAL FEATURE SET CATEGORIES

ICUs and laying the groundwork for models that can be used for risk stratification in a real world setting.

I. RISK SCORING IN HEALTHCARE

In the healthcare domain, risk scoring is used to assess the risk associated with respect to a particular outcome. Risk scoring can be defined with respect to a population, the feature/predictor space and the outcome space. Another dimension of interest is whether the risk scoring setting is static or dynamic. Static risk scoring is a well studied problem and a relative scale for risk scoring can be defined and assessed on historical data. Dynamic risk scoring is needed in scenarios where patient prioritization is needed in relatively short spans of time e.g., in the matter of hours vs. days. Prioritization implies addressing questions like can the clinician use the risk scores to prioritize and characterize the relative degrees of risk for their patients. In a hospital setting physicians making the rounds have a list of patients on their units and having a list of high risk patients can help them to prioritize patients who are likely to need transition into ICU at some point in their encounter. Reducing this morbidity is highly contingent on the timely recognition of changing clinical needs.

Modifiable factors and/or preemptive measures that may improve care can affect/decrease risk of the patient needing care in the ICU setting. Physicians or advanced practice provider often have a list of patients to cover and round on. Knowing which patients have recently shown higher risks of needing ICU care and/or transition can help to prioritize different treatment plans to lower the risk of needing ICU or transitioning to higher acuity care. Additionally House Supervisors or similar staffing personnel may use such models for resource optimization i.e., right place, right time, right level of care (nurse ratios), and right level of clinical expertise. Similarly, House Supervisors have a limited number of resources in staffing and beds and thus having a list of patients who may be at higher risk of transitioning to ICU can help to prioritize where the high risk patients may go and also facilitate discharge of patients who are improving and no longer need ICU levels of care. To summarize, the problem of risk scoring of patients is an important problem in healthcare which affects multiple stakeholders in the care continuum spectrum.

What the examples above illustrate is that dynamic risk scoring is needed in settings like ICU where the risk of a patient can change rapidly. It should however be noted that dynamic risk scoring can be defined in a number of ways depending upon the requirements for a use case. The main possible scenarios are as follows:

1) The underlying population may be changing
2) The availability of features/predictors may change over time
3) The nature of the features/predictors may change over time
4) The target variable may change over time
5) Combination of all of the above factors

In the current study we address the problem of dynamic risk scoring when the predictors or the targets may change over time. We formally define the prediction problem as follows:

Problem: Given a set of patients $p_i \in P$ with historical data comprising of the predictor variable set $v_{pi} = \{v_0, v_1, v_2, ..., v_n\}$ and the target variable $O_i = \{o_1 i, o_2 i, o_3 i, ..., o_n i\}$ in the output space $o_i \in \{ICU, notinICU, transition\}$ at time $t = \{t_0, t_1, t_3, ...t_k\}$ predict the target variable at times $t_{k+1}, t_{k+2}, t_{k+3}, ..., t_n$ where $n > k$.

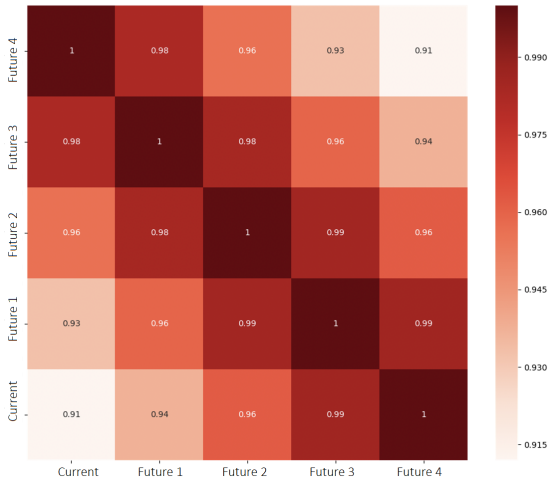

Fig. 1. Correlation between the current and the future patient states

II. RELATED WORK

Risk scores for determining the condition of a patient are widely used in the healthcare domain especially for determining the risk of mortality. One of the oldest scoring system in healthcare which is still used today is the AGPAR Scoring System which is a risk score for newborns [2]. Glasgow Coma Scale is a commonly used scale that is used to assess an individual's neurologic status by evaluating three subscores: eye, verbal, and motor responses [3]. The APACHE score is another commonly used scale that link the risk of mortality to the number of organs failed [4] [5]. The most widely used pediatric mortality risk score is the PRISM score developed by Pollack et al [6] [7]. PRISM is a physiologically based score used to quantify the physiologic derangement of a pediatric patient. When combined with other variables, it can be used to compute expected mortality risk [6]. The current version of PRISM is PRISM III score which has 17 physiologic variables

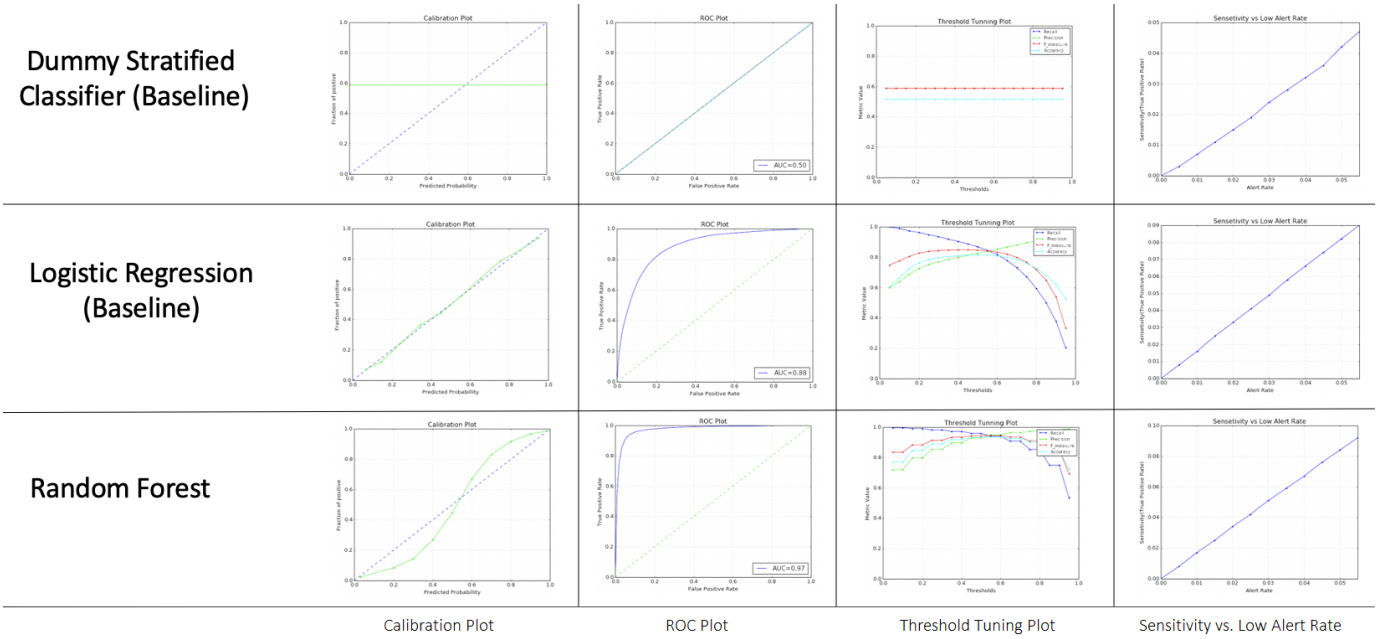

Fig. 2. Summary of results for predicting Future State 1

subdivided into 26 ranges [7]. The PRISM III variables include vital signs, blood gas measures, and lab results from metabolic tests, blood counts, and clotting studies. The Pediatric Index of Mortality (PIM) is another metric that is used for mortality prediction for pediatric ICU populations [8] [9].

There are a number of studies on using machine learning models for predictive models in pediatric ICUs. Kamaleswaran et al [10] and Le et al [11] use machine learning for predicting sepsis in PICU. Che et al [12] considered the problem of predicting 60 day after admit mortality and also predict ventilation Free Days using data from Children's Hospital Los Angeles. Ghassemi et al [13] studied predicting 30 day post-discharge and 1 year post-discharge mortality using information extracted from clinical notes. Nguyen et al [14] used deep learning to asses risk in ICUs for 3-hour mortality prediction. The work of Rubin et al [15] is also relevant to this work as they focus on predicting transfer in PICU using machine learning models. Given the limitation of space it is not possible to cover all of relevant literature on machine learning in PICU, we refer the reader to literature survey papers in this field [16] [17].

III. Dataset

The data that we use comes from the HealthFacts dataset created by Cerner [18]. The timespan for the data is from January 2009 to June 2016. The dataset consists of 1,901,437 records and 42,581 encounters. We also exclude a small number of the patients (768) who passed away during the timespan that we considered. Data which had incorrect dates associated associated with it was also removed. There were 69 tables after taking the union of tables which had the same type of data but for different years. The data can be divided into five broad categories" Pharmacy data, lab data, billing data (codified diagnosis, procedure codes), clinical events (vital signs, pain, sedation, coma scales, respiratory, therapy etc) and microbiology (culture details, organisms isolated, antimicrobial medications, susceptibles). From a union of these sources the final feature set consisted of 1131 features. The breakdown of various categories of features is given in Table I.

The majority of the encounters have more than 5 or more records associated with them. Data cleansing included: a) eliminating duplicate and null values; and b) eliminating data inconsistent with valid entries (e.g. admission times of zero, data inconsistent with life, or negative values). The pediatric population that was considered is defined as the set of ICU pediatric patients and pediatric inpatients. The definition of ICU pediatric patients is as follows:

- Age in years is not missing
- Age in years < 22
- Age in hours > 0 OR Age in hours is not missing
- At least one care setting in (55, 56, 57 ,59, or 60) from source in (lab order, medication dispensed medication request, discharge care setting)

The pediatric inpatient cohort is defined as follows:

- Age in years is not missing
- Age in years < 22
- Age in hours > 0 OR Age in hours is not missing
- Patient type id = 87 (which corresponds to inpatient).

The final cohort that was used was a union of these two cohorts. While the definition of the cohort is relatively straightforward, it is not possible to determine from the data if a patient is in ICU in a straightforward manner. Thus proxies and a complex set of rules need to be used to determine if a patient is in ICU. These rules are based on what is

	Current	Future 1	Future 2	Future 3	Future 4
Accuracy	0.93	0.94	0.93	0.92	0.91
Precision	0.94	0.94	0.94	0.93	0.92
Recall	0.95	0.95	0.95	0.94	0.93
F-Score	0.94	0.94	0.94	0.93	0.92
AUC	0.97	0.97	0.97	0.97	0.96
MCC	0.86	0.87	0.86	0.85	0.82

TABLE II
SUMMARY OF RESULTS FOR PREDICTING CURRENT AND FUTURE STATES

Metric	State	Baseline	LR	RF
Precision	ICU	0.54	0.88	0.88
	not-ICU	0.45	0.83	0.88
	Transition	0.01	0.01	0.60
Recall	ICU	0.53	0.85	0.91
	not-ICU	0.45	0.89	0.87
	Transition	0.01	0.01	0.10
F-Score	ICU	0.53	0.86	0.90
	not-ICU	0.45	0.85	0.87
	Transition	0.01	0.01	0.17

TABLE III
SUMMARY OF RESULTS FOR TERTIARY PREDICTION FOR FUTURE STATE 1

commonly known in the domain and the also the experience of the clinicians involved in the project. The ICU entry and exit times were determined as follows: The ICU entry time was recorded as the earliest hour based on the laboratory order, medication request or medication dispensed associated with an ICU location. Following ICU entry, if there were no laboratory orders, medication requests or medication dispersal, the location was assumed to be the ICU.

We identified the ICU exit time using the following two conditions: First, the medication request, medication dispensed, or laboratory order came from a non-ICU location following ICU care. If the care setting was null or not mapped, the previous care setting was assumed. If multiple care settings including an ICU were observed during the same hour, the care setting was assumed to be an ICU care setting. Second, if the patient was discharged from the hospital from the ICU, the hospital discharge time was used as the ICU discharge time.

Since we recognized that there is imprecision in the assignment of entry and exit times, we assigned a minimum of 10 hours to each ICU stay for survivors (deaths could have a time period $<$ 10 hours). For patients readmitted to the ICU during a hospitalization, a minimum of 40 hours was assigned to each non-ICU time period between ICU admissions. This resulted in 97.0% of PICU (Pediatric ICU) patients having only 1 ICU admission, and 2.3% have 2 ICU admissions, and 0.7% have 3 or more. For the rare patient that had more than 5 ICU admissions (0.13%), the ICU admissions after the 5th admission were not included in the ICU admission.

The target variable is the state of these patient i.e., whether they are in the ICU or not. There are multiple ways to define this variable by varying the time window for definition. We use a time window of 4 hours to define the state of the patient as it made sense from a domain perspective and the condition of the patients did not change as much. We considered multiple targets for prediction by varying the timeslots: the current state of the patient and the next four states of the patient. The correlation between these states is given in Figure 1. As expected the correlation between the states decreases as the time between the states increases.

IV. EXPERIMENTS

We pose the problem of predicting patient states as two classification problems where in the first problem the two classes that need to be predicted are: 'patient in ICU' and 'patient not in ICU' and in the second problem the additional class to be predicted is 'in-transition. This is the state where the patient is being moved from ICU to outside of ICU or vice versa. The data for this state is limited and corresponds to around one percent of all the states. We employ standard 10-fold cross validation with additional constraints to ensure that there is no data leakage as follows: Since a patient may have multiple records which indicate that they are in the hospital, we exclude records in the training set for the patient for whom prediction needs to be made. We employed a set of standard set of classifiers (Decision Trees, AdaBoost with Decision Stump, Naive Bayes, SVM, Random Forest and XGBoost) to choose the algorithm with the best performance. For each of these algorithms, whenever applicable, we applied grid search to find the optimal set of hyperparameters. In addition to these algorithms we also employ two baseline models: A dummy stratified classifier which randomly predicts based on the relative distribution of the classes and Logistic Regression as the baseline. We note that it was not possible to use any of the rule based scoring models currently being used in clinical settings since all the variables data that are used to create these scoring models are not available in HealthFacts.

A summary of prediction results for predicting the next future state are given in Figure 2 which shows the Calibration Plot, the ROC plot, the threshold tuning plot and Sensitivity vs. Low Alert Rate plot. Due to limitations in space we only report the results for the best model which corresponds to Random Forest. We also note that the performance from Random Forest is slightly better than what we get from the RNN model. This implies that having a deep learning model may not be advantageous in this case. The plots in Figure 2 reveal that the Random Forest model is sufficiently calibrated, although model calibration can be improved further. The AUC of the model is 0.97 and the threshold tuning plots also reveal that the model has sufficiently good predictive power as measured by precision and recall. Lastly, the SLA (Sensitivity vs. Low Alert Rate) plots reveal that the model does not degrade when the model confidence is low.

In Table III a summary of results for predicting the current and future states is given. The tables demonstrate that the sufficiently high results are obtained across the board for predicting future states. We also computed the Matthews Correlation Coefficient for the results in Table III since this metric is generally regarded as a good way to capture different aspects of a Confusion Matrix in a single number [19]. The overall performance gives us confidence that the models perform sufficiently well. For the tertiary classification problem

we used the same exact prediction setup i.e., the same set of classifiers were used with 10-fold cross validation. The summary of results for tertiary classification is given in Table III. Due to limitations of space we only show the results from the best model and the baselines. Here the baseline refers to the stratified baseline where the prediction is based on random prediction with respect to the relative distribution of the classes. LR corresponds to Logistical Regression which is being used as another baseline. Lastly, RF corresponds to Random Forest which is the model with the best results. The main thing to note here is that while the results are much better than the baseline and the model is doing well for the two main classes, the results for the minority 'transition' class can be improved greatly.

As described in the previous sections a pivotal aspect of assessment of PICU models described here is the transparency of the models. Towards this end we computed the global factors that drive the overall prediction of the models as well as local predictions which are responsible for individual predictions. The factors are computed using the SHAP framework [20], a summary of the top factors is given in Figure 3. The global factors that were identified by the model was a mix of expected and unexpected features e.g., pCO2 (partial pressure of carbon dioxide is the measure of carbon dioxide within arterial or venous blood) is an important factor with the population bifurcated into two subsets, one with high importance and another with low importance. Vital signs (heart rate, max respiratory rate etc) are highly represented as top features. Receiving medications from the following classes were present: narcotic analgesics, CNS stimulants, and anti-infectives are associated with future ICU care. Higher levels of Albumin are negatively associated with the positive class. Additionally, very young age (age measured in hours) is associated with increased risk for ICU care.

From this set of features we tried to create a set of minimal features that could be used to create a minimal model that could be used in the real world. From a usability perspective features like "stay length" cannot be used in a model if it is predicting at the admit time or if length of stay is not available at the time of prediction, thus it was excluded. The variables related to the stimulants were also excluded since they could show up because of over-representation of certain sub-populations like newborns. We had three physicians who were involved in this study review the top factors to determine if the factors made sense and if they could be used in a minimal model for prediction. We then used the features from the global model as well as input from the domain experts to create a minimal model consisting of 20 features. A summary of the results of the minimal model are given in Table IV which shows only a small decrease in performance except for predicting Future State 1. This indicates that a large number of features may not be needed for prediction to create a sufficiently good model that can be used in practice.

	Current	Future 1	Future 2	Future 3	Future 4
Accuracy	0.87	0.91	0.89	0.89	0.85
Precision	0.84	0.89	0.88	0.87	0.81
Recall	0.86	0.90	0.87	0.86	0.82
F-Score	0.85	0.90	0.87	0.87	0.82
AUC	0.91	0.94	0.88	0.89	0.83

TABLE IV
SUMMARY OF RESULTS FOR MINIMAL MODEL

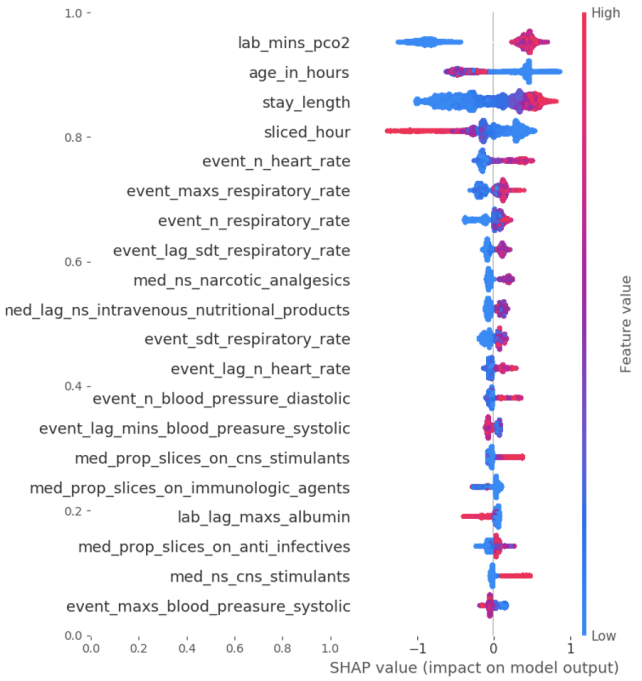

Fig. 3. Top factors for prediction model for Future State 1 as determined by the SHAP model

V. DISCUSSION

The prediction results obtained from the binary classification version of the problem are quite good, however for the tertiary prediction the results can still be greatly improved as both the precision and recall of for the transition class was relatively low. One surprising result from the experiments is that a traditional machine learning model slightly better than a deep learning model. It has been noted in other studies [21] as well that while there are a number of applications in healthcare and medicine where deep learning has proven to be greatly effective, it is not a panacea for machine learning problems. One limitation of this work is that the data used in the study came from a general hospital setting (HealthFacts) and not from a pediatric hospital.

Data on ICU patients is much more comprehensive and complete than other patient populations thus there is lots of high quality data to build and validate predictive models. There is a great deal of room for extending the current models since massive resources of data - real-time telemetry, real-time ventilator data, frequent laboratory studies etc. are frequently under-utilized within models, and indeed most of this data never enters the EHR where it can be used. On the flip-

side, it is also important to acknowledge the bias inherent in curated datasets like HealthFacts since they may not be true representations of ICU patients. This would limit the generalizability of the results. To address this issue we plan to use pediatric ICU data from other source in future follow up work.

VI. Conclusion and Future Work

The problem of predicting the future state of a pediatric patient in ICU is an important problem with far reaching consequences for patient quality of life and even patient survival. Methods that are currently being used for gauging pediatric mortality risk and thus for risk stratification are mostly scoring based systems. One limitation of the current methods is that they cannot be customized to particular settings and cannot make use of rich data that is now being captured and analyzed in the healthcare domain which was not available when these methods were first proposed. To address these issues, we proposed a machine learning approach the problem of predicting a patient's state in pediatric ICUs in this paper.

The predictive models showed sufficiently high predictive performance performed (AUC = 0.97) which was significantly higher than the baseline models. The descriptive explanation models characterized what are the main factors for predicting the next state at the population at the patient level. In addition, the current work has demonstrated that a minimal set of features can produce a highly accurate model as compared to other models that use the entire feature set. This is both encouraging and exciting and gives us confidence that access to additional pediatric ICU data will help develop a dynamic risk scoring solution using machine learning that is able to identify critical attributes that contribute to predicting deterioration in the pediatric population.

For the purpose of assigning patients to ICU, the current models may be sufficient. We hope to these these models in a follow up study in the future. However ICUs often require risk stratification if they are resource constrained. In such use cases relative ordering of severity or criticality of patients would be needed for risk stratification. For this purpose we plan to modify and extend the current set of models so that the probability scores could be used as risk scores for the patients. This would of course require the presence of highly calibrated prediction models. Our long term goal is to create a risk scoring system that can be used in lieu of current risk scoring systems used in PICUs. One advantage of using machine learning approaches as compared to traditional risk scoring systems is that the later can be adapted for particular hospital settings and populations. Additionally, for future work, we propose to integrate the models described in this manuscript in a platform that may be integrated into the workflow of physicians and other healthcare personnel.

References

[1] M. A. Ahmad, C. Eckert, and A. Teredesai, "Interpretable machine learning in healthcare," in *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, 2018, pp. 559–560.

[2] V. Apgar, "A proposal for a new method of evaluation of the newborn," *Classic Papers in Critical Care*, vol. 32, no. 449, p. 97, 1952.

[3] G. Teasdale and B. Jennett, "Assessment of coma and impaired consciousness: a practical scale," *The Lancet*, vol. 304, no. 7872, pp. 81–84, 1974.

[4] W. A. Knaus, J. E. Zimmerman, D. P. Wagner, E. A. Draper, and D. E. Lawrence, "Apache-acute physiology and chronic health evaluation: a physiologically based classification system." *Critical care medicine*, vol. 9, no. 8, pp. 591–597, 1981.

[5] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman, "Apache ii: a severity of disease classification system." *Critical care medicine*, vol. 13, no. 10, pp. 818–829, 1985.

[6] M. M. Pollack, U. E. Ruttimann, and P. R. Getson, "Pediatric risk of mortality (prism) score." *Critical care medicine*, vol. 16, no. 11, pp. 1110–1116, 1988.

[7] M. M. Pollack, K. M. Patel, and U. E. Ruttimann, "Prism iii: an updated pediatric risk of mortality score," *Critical care medicine*, vol. 24, no. 5, pp. 743–752, 1996.

[8] F. Shann, G. Pearson, A. Slater, and K. Wilkinson, "Paediatric index of mortality (pim): a mortality prediction model for children in intensive care," *Intensive care medicine*, vol. 23, no. 2, pp. 201–207, 1997.

[9] A. Slater, F. Shann, and G. Pearson, "Pim2: a revised version of the paediatric index of mortality," *Intensive care medicine*, vol. 29, no. 2, pp. 278–285, 2003.

[10] 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— Society of Critical Care Medicine*, vol. 19, no. 10, pp. e495–e503, 2018.

[11] 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*, vol. 7, p. 413, 2019.

[12] Z. Che, S. Purushotham, R. Khemani, and Y. Liu, "Interpretable deep models for icu outcome prediction," in *AMIA annual symposium proceedings*, vol. 2016. American Medical Informatics Association, 2016, p. 371.

[13] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits, "Unfolding physiological state: Mortality modelling in intensive care units," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 75–84.

[14] P. Nguyen, T. Tran, and S. Venkatesh, "Deep learning to attend to risk in icu," *arXiv preprint arXiv:1707.05010*, 2017.

[15] J. Rubin, C. Potes, M. Xu-Wilson, J. Dong, A. Rahman, H. Nguyen, and D. Moromisato, "An ensemble boosting model for predicting transfer to the pediatric intensive care unit," *International journal of medical informatics*, vol. 112, pp. 15–20, 2018.

[16] J. B. Williams, D. Ghosh, and R. C. Wetzel, "Applying machine learning to pediatric critical care data," *Pediatric Critical Care Medicine*, vol. 19, no. 7, pp. 599–608, 2018.

[17] H. Lonsdale, A. Jalali, L. Ahumada, and C. Matava, "Machine learning and artificial intelligence in pediatric research: current state, future prospects, and examples in perioperative and critical care," *The Journal of Pediatrics*, vol. 221, pp. S3–S10, 2020.

[18] J. Heneghan, A. Patel, E. Trujillo-Rivera, Q. Zeng, F. Faruqe, D. Kim, H. Morizono, J. Bost, and M. Pollack, "370: Medication profiles of children in the intensive care unit: A national assessment," *Critical Care Medicine*, vol. 47, no. 1, p. 166, 2019.

[19] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

[20] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017.

[21] M. Medvedeva, M. Kroon, and B. Plank, "When sparse traditional models outperform dense neural networks: the curious case of discriminating between similar languages," in *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 2017, pp. 156–163.