# DeepSOFA: Clinical Deep Learning for Real-Time Acuity Assessments of Critically Ill ICU Patients

Benjamin Shickel[1], Tyler Loftus[2], Lasith Adhikari[3,5], Tezcan Ozrazgat-Baslanti[3,5], Azra Bihorac[3,5], Parisa Rashidi[1,4,5]

[1]Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, USA
[2]Department of Surgery, University of Florida, Gainesville, FL, USA
[3]Department of Medicine, University of Florida, Gainesville, FL, USA
[4]Department of Biomedical Engineering University of Florida, Gainesville, FL, USA
[5]Precision and Intelligent Systems in Medicine (PRISMAP), University of Florida, Gainesville, FL, USA

## Background

Critically ill patients in the intensive care unit (ICU) have a life-threatening condition or the propensity to develop one at any moment, and early recognition of evolving illness severity in the ICU is invaluable. Timely and accurate illness severity assessments may identify patients in need of life-saving interventions prior to the occurrence of an adverse event and may inform shared decision-making processes among patients, providers, and families regarding goals of care and optimal resource utilization.

One of the most commonly used tools for assessing ICU patient acuity is the Sequential Organ Failure Assessment (SOFA) score[1]. SOFA considers 13 physiological variables representing six different organ systems (cardiovascular, respiratory, nervous, liver, coagulation, and renal) and uses their worst measurements over a 24-hour period in conjunction with static value thresholds to assign numerical scores for each component. The sum of these component scores yields a patient's overall SOFA score, which can be used to assess illness severity and predict mortality[2–4]. Although SOFA provides a reasonably accurate assessment of overall condition and mortality risk, in practice it is often time-consuming, error-prone, and has accuracy hindered by the use of fixed cutoff points for each component score that do not capitalize on individual physiological patterns.

## Objective

The emerging availability of high-fidelity physiologic measurements in the ICU from streaming electronic health record (EHR) systems offers the opportunity to apply computational approaches beyond existing conventional models[5–7]. We propose DeepSOFA[8], a mortality risk prediction framework that utilizes the full scope of patients' temporal measurements in conjunction with deep learning, a collection of machine learning techniques characterized by their ability to automatically learn optimal and often complex variable patterns directly from raw data without requiring manual feature extraction based on a priori domain knowledge[9,10].

## Methods

DeepSOFA is composed of a recurrent neural network (RNN) with gated recurrent units (GRU) and a state-of-the-art self-attention mechanism, and operates by continuously updating its parameters based on multivariate inputs from both the current and previous hours in the ICU. Individual mortality predictions incorporate patterns detected across the entirety of an ICU admission for generating and recalculating dynamic, real-time, and patient-centered acuity assessments.

One of the weaknesses of deep learning techniques is the inherent difficulty in understanding the relative importance of model inputs in generating the output. In the case of mortality prediction, clinicians are interested not only in the likelihood of death, but also in knowing which factors are primarily responsible for the risk of death. To improve clinical interpretability and inspired by state-of-the-art results in other deep learning domains[11], we modified the traditional GRU-RNN network to include a self-attention mechanism to allow clinicians to understand why the deep network is making its predictions. At each hour during a real-time ICU stay, the model's attention mechanism focuses on salient deep representations of data from all previous time points, assigning interpretable relevance scores to every preceding hour that

determine the magnitude of each hour's contribution to the model's overall mortality prediction. By aligning attention scores with the original input time series, the model is able to justify its mortality predictions in a clinician-interpretable manner for increased trust in patient acuity assessments.

DeepSOFA was compared with two baseline models using traditional SOFA scores calculated at each hour using the previous 24 hours of a patient's EHR data. The baseline mortality predictions associated with calculated SOFA scores were derived from both published mortality rate correlations with any given score[2] (Bedside SOFA), and to overall AUC derived from raw SOFA scores[12] (Traditional SOFA). At any hour during an ICU admission, the Bedside SOFA baseline model associated the current SOFA score with a predicted probability of mortality, as would be performed using an online calculator. Traditional SOFA is based on retrospective analysis that derives AUC from raw SOFA scores and mortality outcomes using a given cohort, and while not suitable for real-time prediction in practice, is a reasonable and contemporary baseline and an appropriate challenger to compare with DeepSOFA.

Using both a private hospital cohort of 36,216 ICU admissions (referred to as P1) and a cohort of 48,948 ICU admissions from the publicly available *MIMIC-III* database[13] (referred to as P2), we compared DeepSOFA to baseline approaches using both internal and external validation methods. For internal validation, 5-fold cross validation was performed with each of the two cohorts independently. For external validation, models were trained on the entirety of one cohort and tested on the other. Performance was evaluated using AUC (area under the receiving operating characteristic curve) at each ICU hour.

**Results**

DeepSOFA significantly outperformed traditional SOFA models in external validation cohorts regardless of which cohort was used for model development. For brevity, we list only results from training on the P2 cohort and evaluating on the P1 cohort but note that performance was approximately equivalent between cohorts. Additionally, we omit internal validation results, which resulted in higher performance but carry fewer cross-institutional implications.

Across all hours of each ICU encounter, DeepSOFA had a mean AUC of 0.90, 95% CI 0.90-0.91 ($p<0.05$ compared with Bedside SOFA AUC of 0.79, 95% CI 0.79-0.80 and $p<0.05$ compared with Traditional SOFA AUC of 0.85, 95% CI 0.85-0.86). At the final hour of each ICU stay, DeepSOFA yielded an AUC of 0.93, 95% CI 0.93-0.94, $p<0.05$ compared to Bedside SOFA AUC of 0.82, 95% CI 0.81-0.83 and $p<0.05$ compared to Traditional SOFA AUC of 0.88, 95% CI 0.88-0.89). Although model performance decreased slightly when prediction occurred earlier in the ICU encounter, DeepSOFA retained excellent AUC above 0.87, 95% CI 0.87-0.88 up to 100 hours away from hospital discharge or death, regardless of the mortality time point of interest. These findings were consistent across both cohorts.

**Discussion**

In large, heterogeneous populations of ICU patients, we have developed and externally validated a dynamic and interpretable deep learning framework (DeepSOFA) that uses a time-honored illness severity score framework to predict in-hospital mortality with significantly greater accuracy than traditional methods. Previous work has often employed multivariable regression models in predicting mortality using methods such as SOFA[1], SAPS[14–16], MPM[16,17], and APACHE[16,18]. Although these methods have produced reasonably accurate predictions, their accuracy is inferior to that of deep models, and their clinical application is cumbersome compared with automated models that are well suited to capitalize on the emerging availability of streaming EHR data. In this regard, deep models may augment clinical decision-making by serving as an early warning system to identify patients in need of therapeutic interventions and by informing the shared decision-making processes among patients, providers, and families regarding goals of care and resource utilization by instantaneously assessing large volumes of data over time, a task which is difficult and time-consuming for clinicians.

**Conclusion**

To our knowledge, DeepSOFA is the first application of deep learning toward generating real-time patient acuity scores[8]. Our interpretability mechanism is also a novel application of recent advances in deep

learning self-attention, where we visualize the severity of fundamental time series patterns and their overall effect on the resulting acuity scores and mortality predictions. DeepSOFA may be applied to individual patients, exhibiting consistent and proportionate responses to clinical events, with visual representation of the probability of death and time periods during which model inputs disproportionately contributed to predictions. These findings suggest that the SOFA score can be augmented with more nuanced and intelligent mechanisms for assessing patient acuity. Deep learning technology may be used to augment clinician decision-making by generating accurate real-time prognostic data to identify patients in need of therapeutic interventions and inform shared decision-making processes among patients, providers, and families.

## References


1. Vincent, J. L. *et al.* The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med* **22,** 707–710 (1996).
2. Ferreira, F., Bota, D., Bross, A., Mélot, C. & Vincent, J. Serial evaluation of the sofa score to predict outcome in critically ill patients. *J. Am. Med. Assoc.* **286,** 1754–1758 (2001).
3. Vincent, J.-L. *et al.* Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: Results of a multicenter, prospective study. *Crit. Care Med.* **26,** (1998).
4. Minne, L., Abu-Hanna, A. & de Jonge, E. Evaluation of SOFA-based models for predicting mortality in the ICU: A systematic review. *Crit. Care* **12,** R161 (2008).
5. Sujin Kim, Woojae Kim & Rae Woong Park. A Comparison of Intensive Care Unit Mortality Prediction Models through the Use of Data Mining Techniques. *Healthc. Inform. Res.* **17,** 232–243 (2011).
6. Meyfroidt, G., Güiza, F., Ramon, J. & Bruynooghe, M. Machine learning techniques to examine large patient databases. *Best Pract. Res. Clin. Anaesthesiol.* **23,** 127–143 (2009).
7. Clermont, G., Angus, D. C., DiRusso, S. M., Griffin, M. & Linde-Zwirble, W. T. Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models. *Crit. Care Med.* **29,** (2001).
8. Shickel, B. *et al.* DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning. *Sci. Rep.* **9,** 1879 (2019).
9. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015).
10. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning*. (MIT Press, 2016).
11. Vaswani, A. *et al.* Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* 5998–6008 (2017). doi:10.1017/S0952523813000308
12. Badawi, O., Liu, X., Hassan, E., Amelung, P. J. & Swami, S. Evaluation of ICU Risk Models Adapted for Use as Continuous Markers of Severity of Illness Throughout the ICU Stay. *Crit. Care Med.* **46,** 361–367 (2018).
13. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. Data* **3,** 1–9 (2016).
14. Metnitz, P. G. H. *et al.* SAPS 3-From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description. *Intensive Care Med.* **31,** 1336–1344 (2005).
15. Moreno, R. P. *et al.* SAPS 3 - From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med.* **31,** 1345–1355 (2005).
16. Afessa, B., Gajic, O. & Keegan, M. T. Severity of Illness and Organ Failure Assessment in Adult Intensive Care Units. *Crit. Care Clin.* **23,** 639–658 (2007).
17. Higgins, T. L. *et al.* Assessing contemporary intensive care unit outcome: an updated Mortality Probability Admission Model (MPM0-III). *Crit. Care Med.* **35,** 827–35 (2007).
18. Zimmerman, J. E., Kramer, A. a, McNair, D. S. & Malila, F. M. Acute Physiology and Chronic

Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit. Care Med.* **34,** 1297–1310 (2006).