Self-explaining Neural Network with Concept-based Explanations for ICU Mortality Prediction

(I²) Institute for Informatics

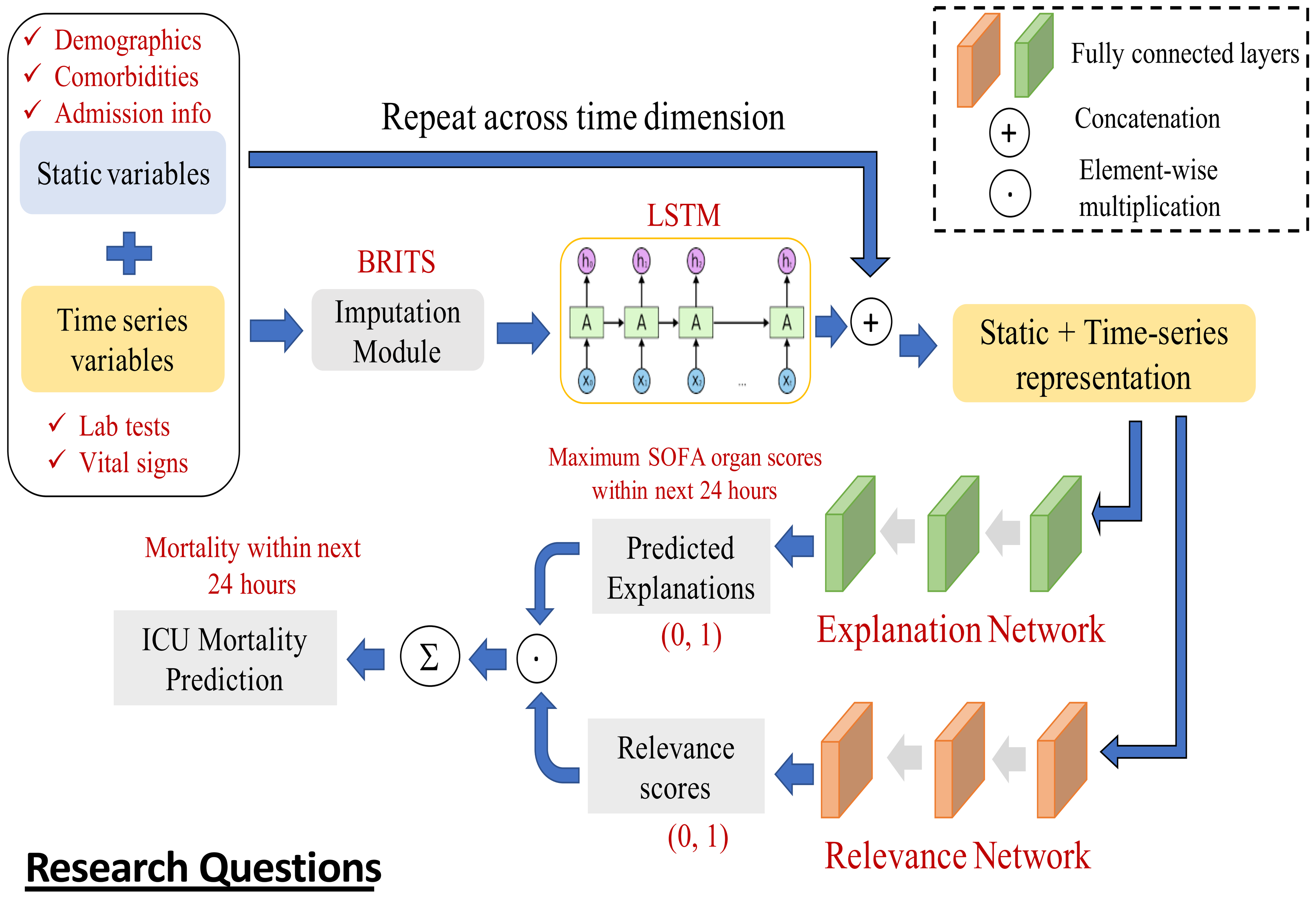
Sayantan Kumar, MS; Sean C. Yu, MS; Thomas Kannampallil, PhD; Zachary B. Abrams, PhD; Andrew Michelson, MD; Philip R.O. Payne, PhD

Washington University in St. Louis School of Medicine

Introduction

- **Challenges**:
  - ➢ Machine learning models deployed in a healthcare setting need to be interpretable, they cannot be a black box.
  - ➢ Post-hoc explanations, key issue of ownership, not reliable for clinical understanding.[1]
  - ➢ Raw clinical variables (e.g. pixels in medical imaging) as units of explanation posey major challenge for clinical interpretation.[2,3]
- **Contributions**:
  - ➢ Expert-knowledge driven **intermediate high-level concepts** derived from raw clinical features as units of explanation.
  - ➢ Deep learning framework to **jointly predict and explain** in end-to-end setting.
- **Use case:** Intensive care unit (ICU) mortality prediction
  - ➢ **Sequential Organ Failure Assessment (SOFA) organ scores** as high-level concepts (explanation units).
  - ➢ **Relevance scores** – Which organ system failures correlate with mortality?

Methods



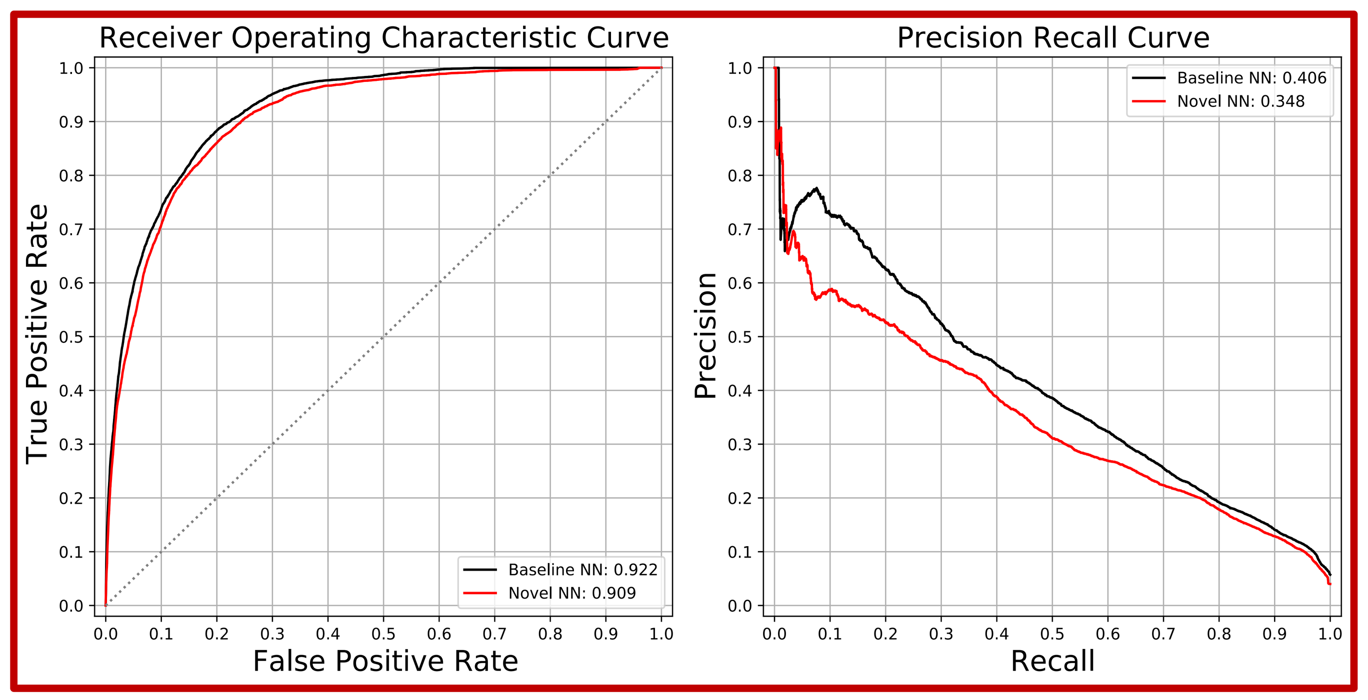
Research Questions

- ✓ Does adding explainability components to a deep learning framework affect it's prediction performance (interpretability-performance trade-off)?
- ✓ Are the predicted explanations grounded in terms of expert domain knowledge?
- ✓ Are the explanations generated by the model understandable for clinicians?

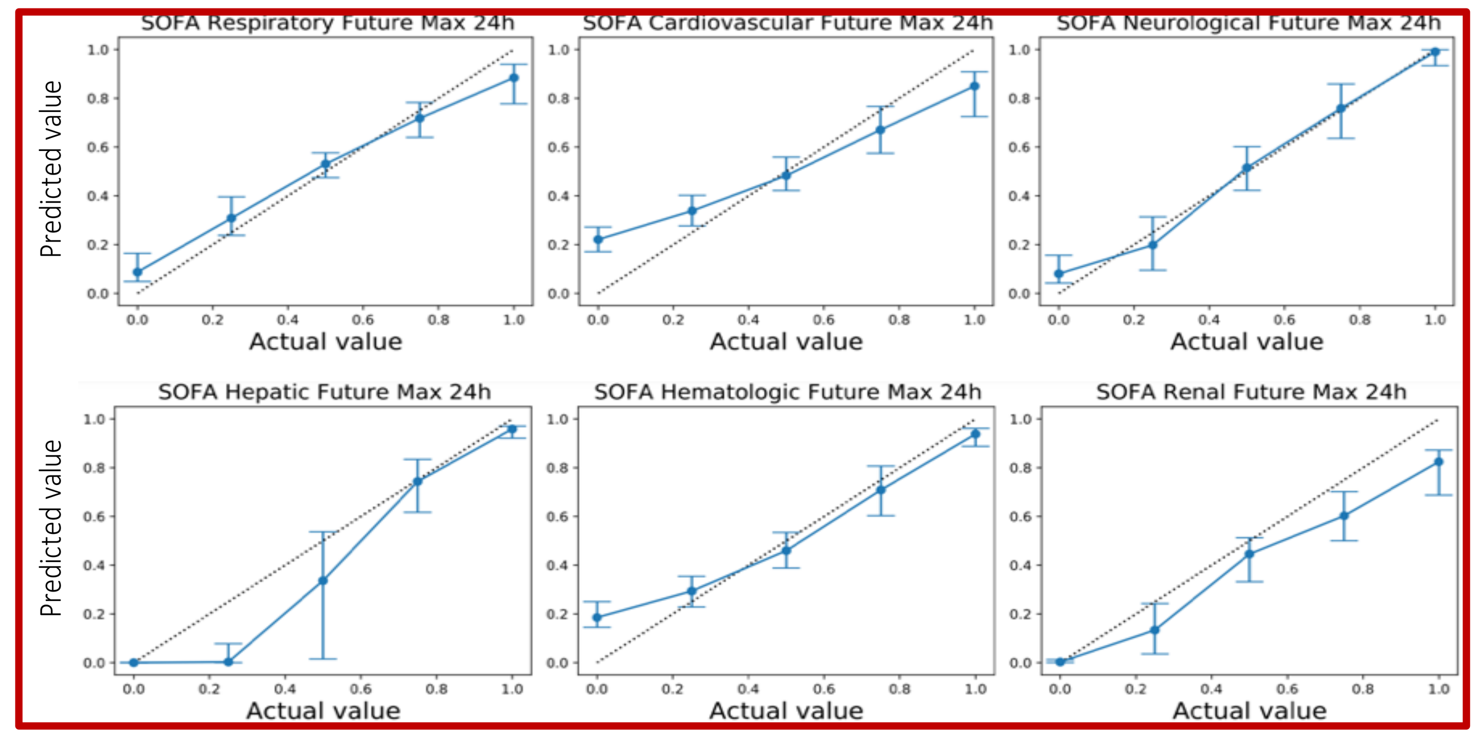
Experimental Results

- **Data Sources**:
  - ➢ Dataset: MIMIC-IV
  - ➢ 2,043 (8.9%) experienced in-hospital mortality out of 22,944 admissions.
- **Features**:
  - ➢ Time-invariant (n = 24) – demographics, comorbidities, admission information.
  - ➢ Time-series (n = 87) – lab tests and vital signs calculated at hourly time interval.
  - ➢ Quantile-based outlier removal.
- **Ablation Study:**
  - ➢ **Baseline** framework without the explanations and relevance scores.
  - ➢ Performance-interpretability trade-off

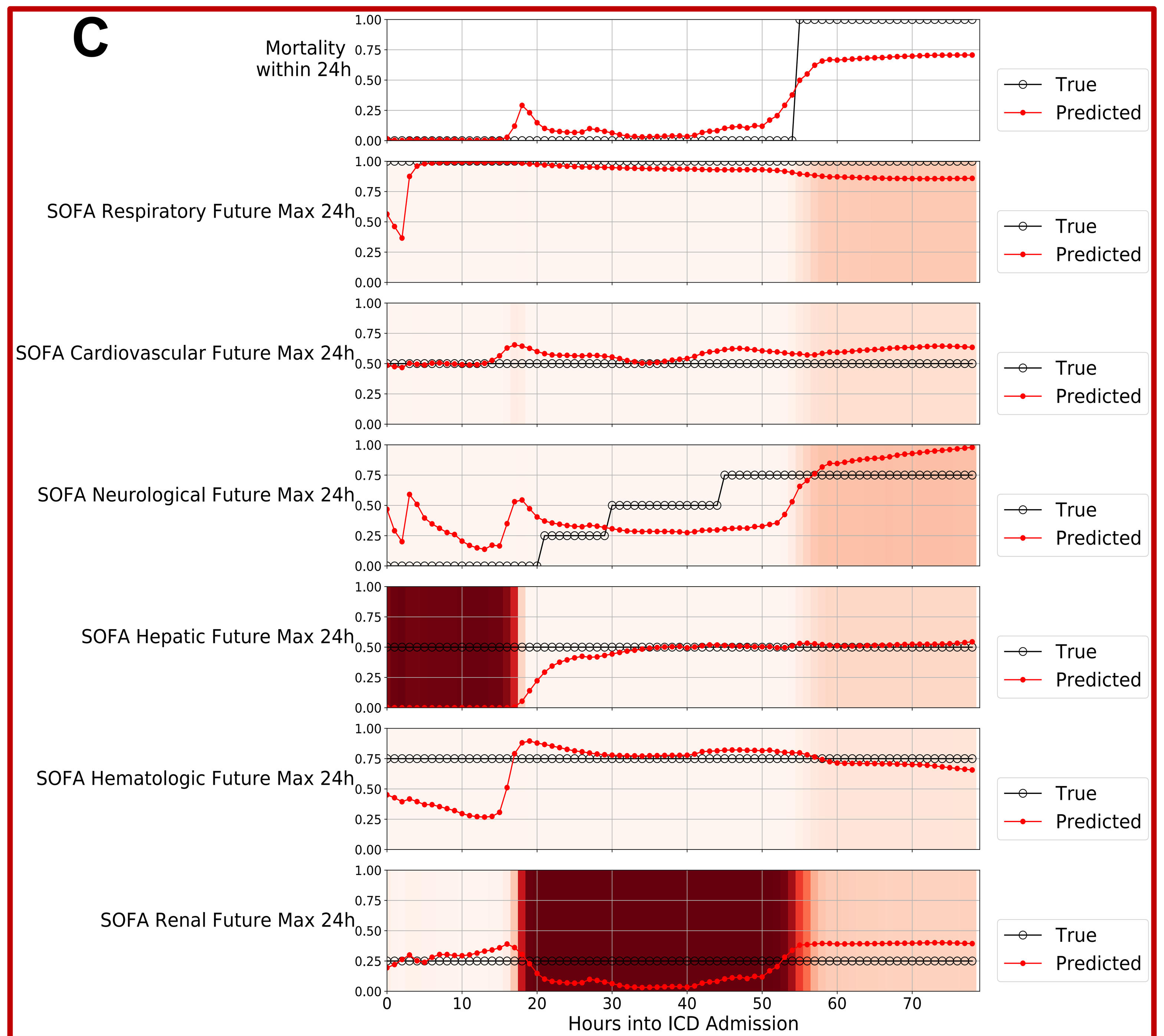
A



B



C



- **A**: AUROC and AUPRC of proposed and baseline.
- **B**: Performance of proposed model on auxiliary tasks.
- **C**: Explanations and relevance scores at each timepoint within a patient's ICU trajectory.

Discussion and Conclusion

- **Performance-interpertability tradeoff**:
  - ➢ Adding explainability components does not impact prediction performance of model **(Figure A).**
  - ➢ Relevance scores are learned without additional training effort.
- **Pre-defining SOFA as units of explanation**:
  - ➢ Predicted explanations are grounded in terms of expert knowledge (**Figure B**).
  - ➢ SOFA – derived from raw variables and used by clinicians as intermediate knowledge to analyse ICU mortality.
- **Quality of explanations:**
  - ➢ As the predicted probability of mortality rises, the model is shown to pay more importance to anticipated respiratory, neurological, hepatic and renal organ failure, highlighting their contribution towards mortality **(Figure C).**
  - ➢ Help clinicians understand the health status of patient throughout length of ICU stay.

References

(1) Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence* 1.5 (2019): 206-215.
(2) Alvarez Melis, David, and Tommi Jaakkola. "Towards robust interpretability with self-explaining neural networks." Advances in neural information processing systems 31 (2018).
(3) Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." International conference on machine learning. PMLR, 2018.