Explaining Longitudinal Clinical Outcomes using Domain-Knowledge driven Intermediate Concepts

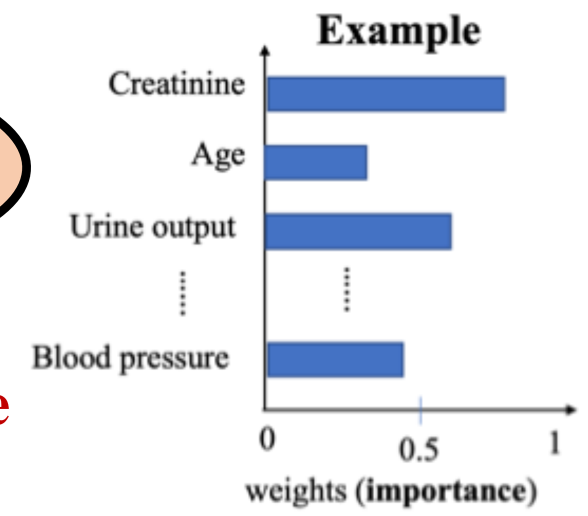
Sayantan Kumar [1], Thomas Kannampallil [1,2,4], Philip R.O. Payne [1,4], Aristeidis Sotiras [3,4]

[1]Department of Computer Science and Engineering, Washington University in St. Louis, [2]Division of Anesthesiology, Washington University School of Medicine, [3]Department of Radiology, Washington University School of Medicine, [4]Institute for Informatics, Data Science and Biostatistics, Washington University School of Medicine

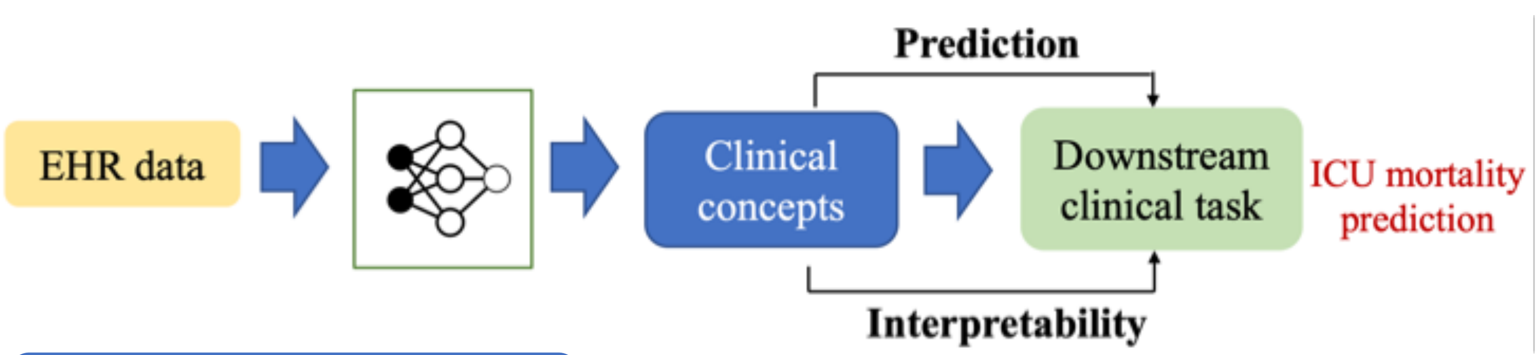
Limitations : Existing XAI methods

Input features as units of interpretation

- ✓ Interpretation changes with change in features.
- ✓ Interpretation is less informative for high-dimensional data.



Contributions: Clinical concepts



Latent bottleneck	Should preserve **relevant information**
Intermediate	**Aggregated knowledge**, derived from input features
Interpretable	**Associated** with clinical outcome

Use case: ICU mortality using organ-failure scores

Clinical outcome: ICU mortality

- At each point, predict mortality risk within next 24 hours
- Tracks patient's health status in ICU.
- 0 – 4 (**higher** = more risk of organ failure).

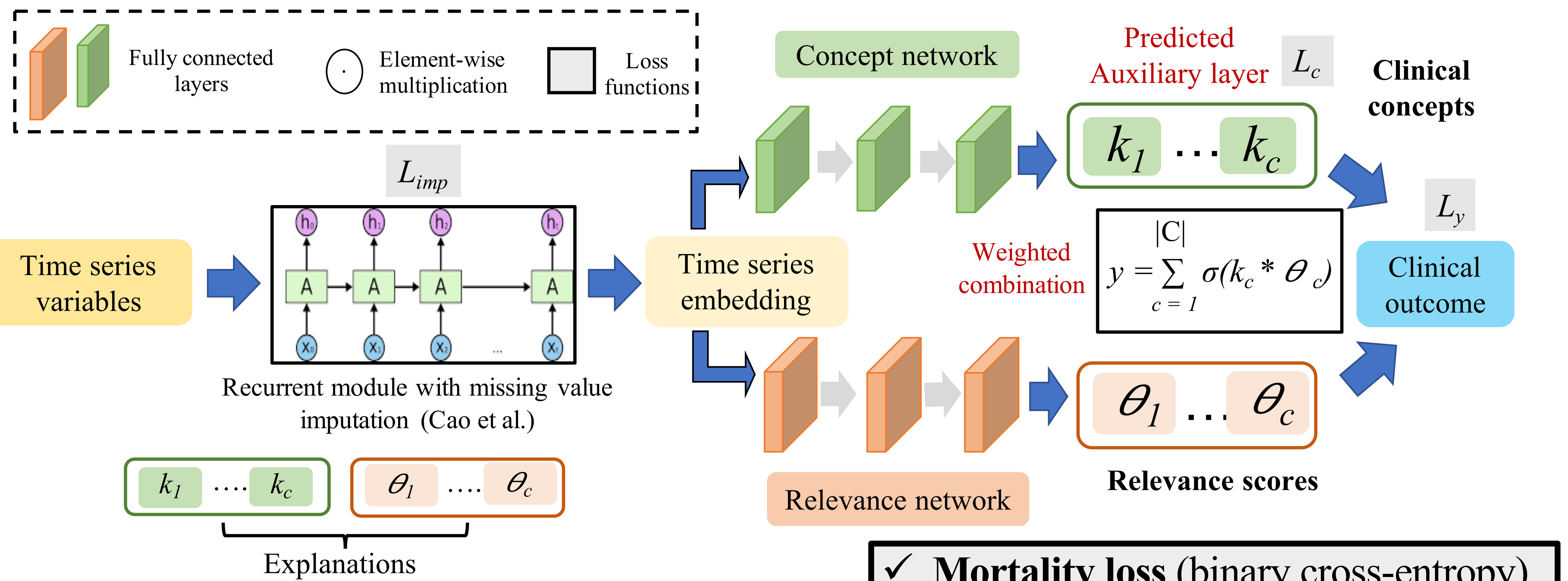
Concepts: Sequential Organ Failure Assessment (SOFA)

- Extent of failure (**risk scores**) for organ systems.
- Tracks patient's health status in ICU.
- 0 – 4 (**higher** = more risk of organ failure).

Dataset: MIMIC IV

- 2043/22904 (8.9%) experienced in-hospital mortality.
- Time-series variable (n = 87) – lab tests and vital signs calculated at hourly time interval.

Proposed Method

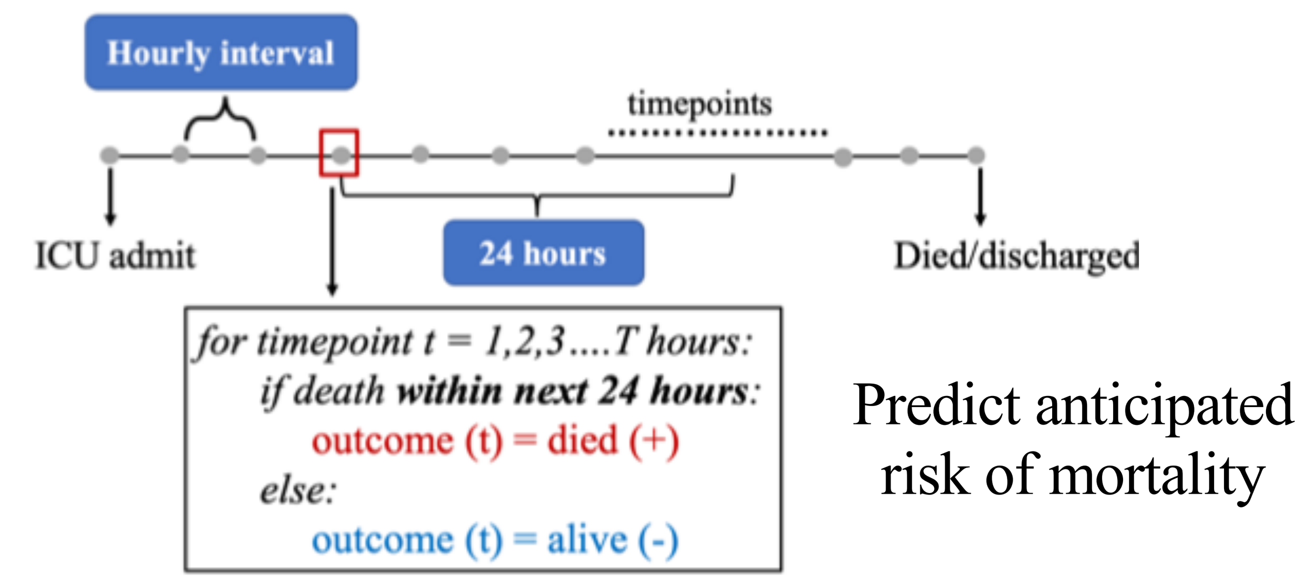


$$\text{Loss} = \lambda_1 L_{mort} + \lambda_2 L_{aux} + \lambda_3 L_{imp}$$

- ✓ **Mortality loss** (binary cross-entropy)
- ✓ **Auxiliary loss** (mean-squared error)
- ✓ **Imputation loss** (mean-squared error)

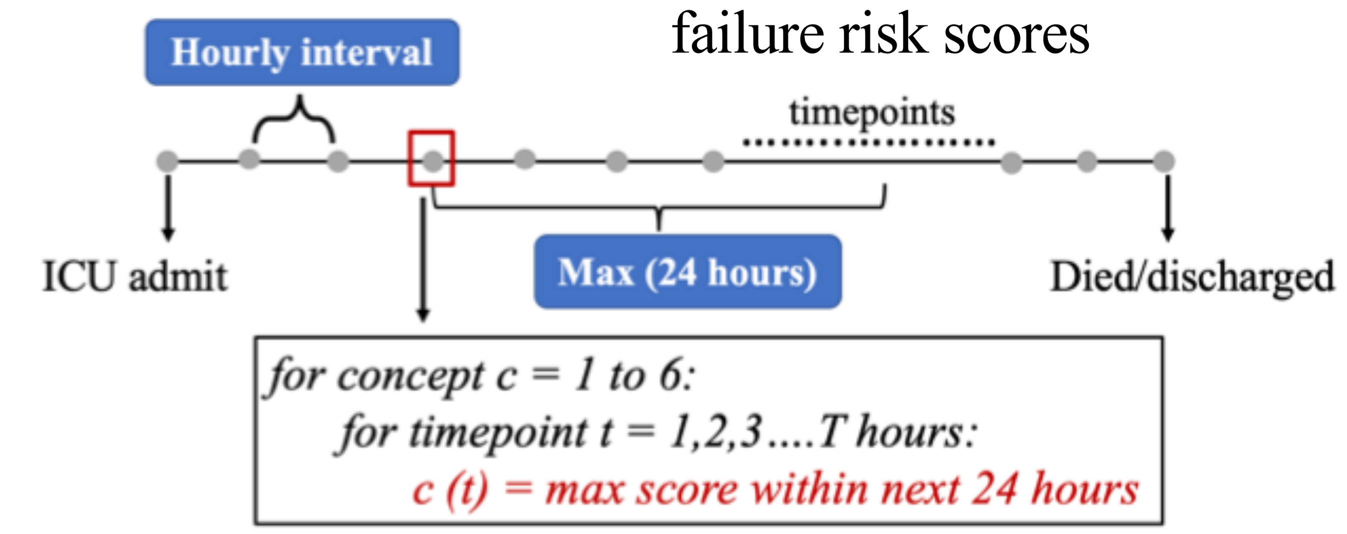
- ✓ Expert-knowledge driven **intermediate high-level concepts** as units of explanation.
- ✓ Deep learning framework to **jointly predict and explain** in end-to-end setting.

Longitudinal prediction



Predict anticipated risk of mortality

Predict anticipated organ-failure risk scores



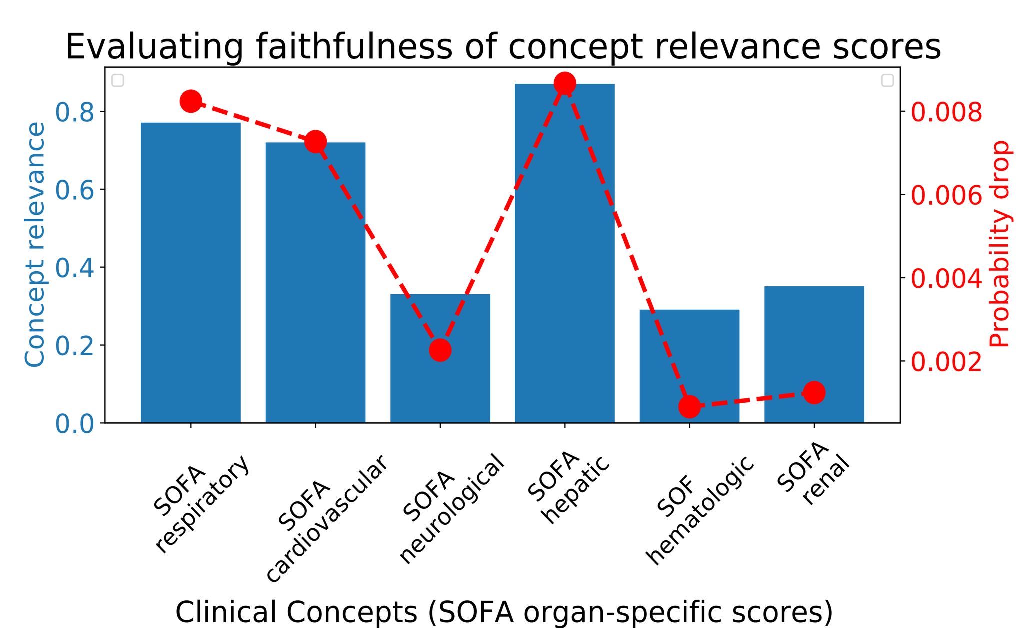
Experimental Results

Self-explaining nature of model does not sacrifice prediction performance

Computational cost is reasonable

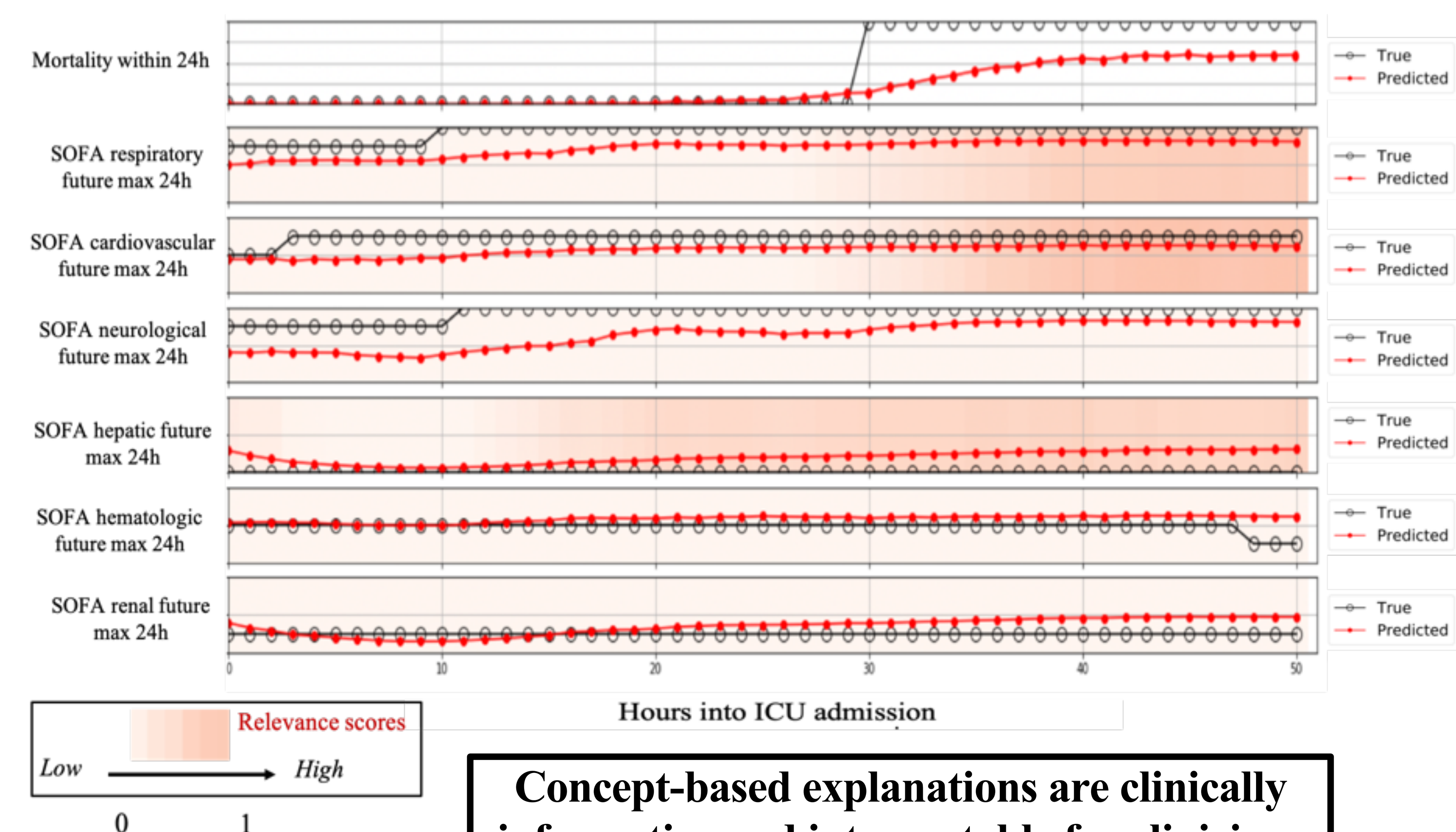
Table 1: Mortality prediction performance (AUROC, AUPRC and training time in seconds). The values indicate mean and 95% values confidence interval after repeated experiments.

	Model	AUROC	AUPRC	Training time
Conventional ML models	SVM [9]	0.737 [0.725-0.747]	0.432 [0.415-0.452]	56s
	Random Forest [9]	0.723 [0.708-0.734]	0.443 [0.422-0.461]	74s
	XgBoost [7]	0.784 [0.775-0.791]	0.526 [0.497-0.556]	36s
SOTA baselines	FCN [3]	0.812 [0.785-0.847]	0.495 [0.482-0.514]	112s
	GRU + MLP [6]	0.894 [0.865-0.918]	**0.532 [0.517-0.558]**	140s
Ablation study	SOFA_only	0.715 [0.695-0.727]	0.378 [0.362-0.385]	78s
	No missing imputation	0.825 [0.805-0.847]	0.472 [0.452-0.481]	282s
	RNN + MLP (no concept)	0.902 [0.875-0.933]	0.524 [0.517-0.533]	128s
	Proposed	**0.923 [0.915-0.947]**	0.529 [0.505-0.551]	342s



Relevance scores indicative of true importance of concept

- Drop concepts one at a time.
- Measure drop in prediction probability.
- Correlation between relevance scores and probability drop.



Concept-based explanations are clinically informative and interpretable for clinicians

As predicted probability of mortality rises, **model is shown to pay more attention to *anticipated* respiratory, cardiovascular and hepatic failure**