Explaining Longitudinal Clinical Outcomes using Domain-Knowledge driven Intermediate Concepts

Sayantan Kumar
Department of Computer Science and Engineering
Washington University in St. Louis
St. Louis, MO, USA
sayantan.kumar@wustl.edu

Thomas Kannampallil
Department of Anesthesiology, School of Medicine
Washington University in St. Louis
thomas.k@wustl.edu

Aristeidis Sotiras
Department of Radiology, School of Medicine
Washington University in St. Louis
aristeidis.sotiras@wustl.edu

Philip Payne
Institute for Informatics, Data Science, and Biostatistics
Washington University in St. Louis
prpayne@wustl.edu

Abstract

The black-box nature of complex deep learning models makes it challenging to explain the rationale behind model predictions to clinicians and healthcare providers. Most of the current explanation methods in healthcare provide explanations through feature importance scores, which identify clinical features that are important for prediction. For high-dimensional clinical data, using individual input features as units of explanations often leads to noisy explanations that are sensitive to input perturbations and less informative for clinical interpretation. In this work, we design a novel deep learning framework that predicts domain-knowledge driven intermediate high-level clinical concepts from input features and uses them as units of explanation. Our framework is self-explaining; relevance scores are generated for each concept to predict and explain in an end-to-end joint training scheme. We perform systematic experiments on a real-world electronic health records dataset to evaluate both the performance and explainability of the predicted clinical concepts.

1 Introduction

Wider availability of Electronic Health Records (EHR) has led to an increase in deep learning applications for clinical diagnosis and prognosis [18, 13, 20, 5]. While these deep neural networks allow for accurate and dynamic performance, the black-box nature of these models makes it challenging to understand the rationale behind model predictions. Particularly, in the healthcare

XAI in Action: Past, Present, and Future Applications @ NeurIPS 2023.

setting, explainability is critical in engendering trust amongst clinicians in the usage of deep learning based clinical decision support systems [17, 24].

Multiple approaches have been proposed to provide explanations for deep learning models applied to EHR data, with a focus on using input clinical features as the units of explanation [21]. Feature-based explanations in healthcare involve assigning weights to individual clinical features highlighting their contribution towards model prediction, e.g. saliency maps [14] and Shapley explanations [16]. For high-dimensional inputs, feature-based explanations are sensitive to input perturbations leading to noisy explanations and are difficult for coherent interpretation [8, 22]. To address this challenge, we can operate on higher-level concepts or feature intermediates, derived from a combination of raw features. These high-level concepts can be understood as aggregated knowledge which clinical experts often rely on to make decisions. For example, in medical imaging, high-level concepts driven by expert knowledge such as tissue ruggedness or elongation are strong predictors of cancerous tumours and can be the natural "units" of explanation for doctors to make their diagnosis [15]. Recent work on concept-based explanations focus on learning concepts from images of simpler toy datasets like MNIST, CIFAR10 and UCI datasets [1, 11, 25]. Learning unsupervised clinical concepts from EHR data without any kind of domain-knowledge supervision makes it challenging to learn clinically meaningful concepts.

In this work, we propose a novel deep learning framework that learns high-level intermediate clinical concepts, supervised by domain knowledge as the units of model explanation. The clinical concepts should satisfy the following desiderata: (i) **expert-knowledge driven:** well-validated metric used by clinicians for analyzing the particular clinical outcome. (ii) **Intermediate knowledge:** intermediate features derived from an aggregated assessment of individual clinical variables (iii) **high-level**: easier for clinical interpretation. Our proposed framework consists of a recurrent module with missing value imputation for time-series EHR variables, concept network for predicting clinical concepts as both auxiliary tasks and units of interpretation, relevance network for generating relevance scores (contribiutions/importance) for each concept, and a regression module to predict the clinical outcome using the concepts (features) and relevance scores (weights). Our framework is self-explaining in nature; predictions and explanations are generated in an end-to-end joint training scheme. Our work is alligned to Mincu et al.[19] where clinical concepts based on EHR data were used for post-hoc explanations, without learning them in a supervised setting. To the best of our knowledge, ours is the first end-to-end approach which learns both supervised high-level clinical concepts from EHR data and intrinsically developed model explanations in the context of predicting a clinical outcome.

We tested our framework on a publicly available EHR dataset and evaluated explainability based on the following criteria: (i) **explainability-performance tradeoff**: does the self-explaining nature of our model sacrifice prediction performance? (ii) **clinically meaningful:** are the explanations understandable to clinicians, (iii) **faithfulness:** are the relevance scores indicative of "true" importance? and (iv) **grounding:** are the concepts grounded or close to domain knowledge?

2 Proposed Methodology

2.1 Problem formulation

We denote a dataset of multivariate longitudinal EHR data of a single subject as $X = \{x_1, x_2, ...X_T\}$ as a sequence of T observations. The t^{th} observation $x_t \in \mathbb{R}^D$ consists of D features $\{x_t^1, x_t^2, ...x_t^D\}$ and was observed at timestamp s_t. Let $y_t \in \{0, 1\}$ represent the outcome label at timestep t. Due to missing values in EHR data, m_t^d represents a masking vector such that $m_t^d = 0$ if x_t^d is not observed and 1 otherwise. Since EHR features can be missing for consecutive timestamps, we denote δ_t^d to be the time gap from the last observation to the current timestamp s_t such that $\delta_t^d = s_t - s_{t-1} + \delta_{t-1}^d$ for $m_{t-1}^d = 0$,$t > 1$; $\delta_t^d = s_t - s_{t-1}$ for $m_{t-1}^d = 1$,$t > 1$ and $\delta_t = 0$ for $t = 1$. For clarity, all notations in the following sections represent a single subject.

2.2 Time-series embedding with missing value imputation

Following the BRITS algorithm [4], we use a recurrent neural network for embedding time-series features where the missing values are imputed based on recurrent dynamics. We represent a

standard recurrent neural network as $h_t = \sigma(W_h h_{t-1} + U_h x_t + b_h)$ where σ is the sigmoid function, W_h, U_h and b_h are parameters, and h_t is the hidden state of previous time steps. The missing values are imputed by a regression component which transfers the hidden state h_{t-1} to the estimated vector $\hat{x}_t$ (Equation (1)). The hidden state h_t is updated by (Equation (4)) where x_t^c represents the complement input when x_t is missing (Equation (2)) and γ_t represents the temporal decay factor to decay the hidden vector h_{t-1}. The temporal decay factor (Equation (3)) represents the missing patterns in the time-series with smaller γ to decay the hidden state h_{t-1} if δ_t is large (values missing for a long time). The imputation loss (Equation (5)) can be calculated by the mean squared error between input x_t and imputed vector $\hat{x}_t$.

$$\hat{x}_t = W_x h_{t-1} + b_x \tag{1}$$
$$x_t^c = m_t \cdot x_t + (1 - m_t) \cdot \hat{x}_t \tag{2}$$
$$\gamma_t = \exp(-max(0, W_\gamma \delta_t + b_\gamma)) \tag{3}$$
$$h_t = \sigma(W_h[h_{t-1} \cdot \gamma_t] + U_h[x_t^c \cdot m_t] + b_h) \tag{4}$$
$$L_{imp} = MSE(x_t, \hat{x}_t) \tag{5}$$

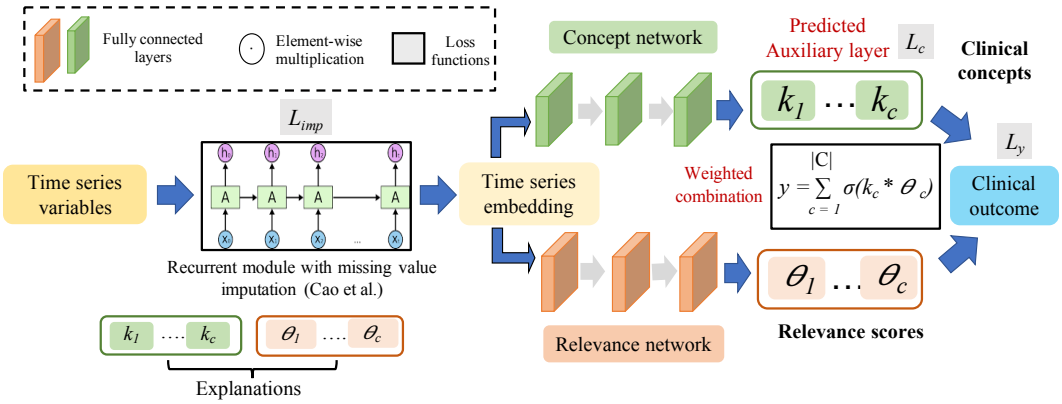

Figure 1: Our proposed framework has a recurrent module with missing value imputation for time-series EHR variables, concept network for predicting clinical concepts as both auxiliary tasks and units of interpretation, relevance network for generating relevance scores (contribiutions/importance) for each concept, and a regression module to predict the clinical outcome using the concepts (features) and relevance scores (weights).

2.3 Concepts and relevance scores

The output of the recurrent module represents the latent vector (imputed vector $\hat{x}_t \in \mathbb{R}^{\mathbb{T} \times \mathbb{H}}$) where H is the hidden layer dimension. $\hat{x}_t$ can be passed through the concept network $C(\hat{x}_t)$ to generate |C| concepts $K = \{k_1, k_2, ...k_C)\}$ such that the c-th concept vector $k_c = \{k_c^1, k_c^2, ..k_c^T\}$ represent the concept values at each timepoint. Similarly, $\hat{x}_t$ is passed through the relevance network $\theta(\hat{x}_t)$ to generate the relevance scores for |C| concepts $\theta = \{\theta_1, \theta_2, ...\theta_C)\}$ such that the c-th relevance vector $\theta_c = \{\theta_c^1, \theta_c^2, ..\theta_c^T\}$ represent the relevance scores at each timepoint corresponding to the c-th concept $k_c = \{k_c^1, k_c^2, ..k_c^T\}$. Both $C(\hat{x}_t)$ and $\theta(\hat{x}_t)$ can be represented by a fully-connected neural network.

2.4 Explanations and final outcome

The predicted probability of the final clinical outcome Y outcome can be estimated by a regression module with the concepts (features) and the relevance scores (weights) as $Y = \{y_1, y_2, ..y_T\}$ where $y_t = \sigma(\sum_{j=1}^{|C|} \theta_j^t \times K_j^t)$ represents the predicted outcome probability at the t-th timepoint. The use of a regression module using the concepts and relevance scores is motivated by the interpretability of a linear model $y_t = \theta_i^t \times x_i^t$ where the constant coefficient θ_i^t represents the explicit contribution of input feature x_i^t. Building on a linear regression model and making it more complex, both $\theta(x)$ and $K(x)$ can be learnt from input features and $\theta_j^t(x)$ represents the contribution (relevance) of concept $K_j^t(x)$ towards the final outcome at each timestep.

2.5 Loss function

The final loss function L can be represented by $L = \lambda_1 L_y + \lambda_2 \sum_{c=1}^{|C|} L_c + \lambda_3 L_{imp}$ where L_y is the cross-entropy classification loss for the final predicted outcome, L_c is the auxiliary concept loss for the cth concept, represented by the mean squared error and L_{imp} is the imputation loss calculated in Equation (5). λ_1, λ_2 and λ_3 represent the coefficients of the different losses respectively. The supervised auxiliary concept loss encourages the predicted concepts to be similar to the domain-knowledge values. The coefficients of the loss terms are set as $\lambda_1 = 1$, $\lambda_2 = 0.8$ and $\lambda_3 = 0.05$ after hyper-parameter tuning using grid search.

3 Experimental Design

3.1 Dataset and feature preprocessing

We conducted our experiments on the Medical Information Mart for Intensive Care IV (MIMIC-IV v0.4) database [12]. Our cohort consists of 22,944 ICU admitted patients between 2008-2019, of which 2043 (8.9%) experienced in-hospital mortality. Only the first admission was considered in case of multiple ICU admissions. For each patient, we extracted 87 time-series features which includes laboratory test results and vital signs. Feature pre-processing of time-series variables include clipping the outlier values to the 1st and 99th percentile values and standardization using the RobustScalar package from sklearn [2]. Time-varying variables were aggregated into hourly time buckets using the median for repeated values.

3.2 Clinical outcome

Our aim is to predict the risk(probability) of a patient's death at each timestep within his/her stay in the ICU. At each timepoint within the patient's ICU trajectory, the model predicts if the patient will die within the next 24 hours period (Figure 2 left). For example, if a patient stays within the ICU from t=0 to t=56 hours(death time) then the clinical outcome labels (ground truth) from t=0 to t=32 will be 0 (survival) and 1 (death) from t=32 onwards.

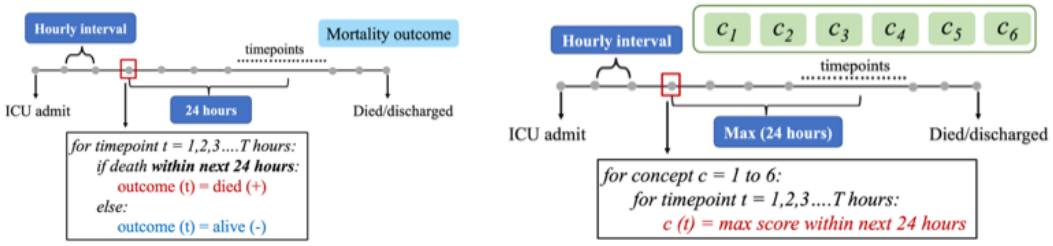

Figure 2: Calculating longitudinal ground-truth label for mortality (**left**) and auxiliary concepts (**right**).At each timepoint, mortality label is positive if patient dies within the next 24 hours window. The concept ground truth value at each timepoint is the maximum SPOFA organ-failure scores within a 24-hour window following the current timepoint.

3.3 Domain-knowledge concepts

Our proposed framework uses Sequential Organ Failure Assessment (SOFA) organ-failure risk scores as high-level clinical concepts to explain patient mortality in the ICU [23]. The six concepts in our model correspond to the organ-failure risk scores for each of the six organ systems: respiratory, cardiovascular, neurological hepatic, hematologic (coagulation) and renal systems. SOFA organ scores vary between 0-4 with high scores indicating severe organ system conditions [10]. The SOFA organ scores satisfy the properties of clinical concepts since they are (i) **expert-knowledge driven:** well-validated metric used by clinicians to understand ICU mortality, (ii) **intermediate knowledge:** each organ-failure score is an aggregated assessment of clinical features and (iii) **high-level:** easier for human cognition since we are moving one level up from feature-level to organ level. Similar to the final clinical outcome, the predicted explanations are also anticipated longitudinal predictions. At each timestep, the model predicts anticipated organ-failure risk scores (maximum score within the next 24 hour window) (Figure 2 right).

3.4 Baselines and ablation studies

Our goal is to analyze if the self-explaining nature of our proposed model sacrifices prediction performance (explainability-performance trade-off). We compared the prediction performance of our proposed model with following state-of-the-art baseline methods on ICU mortality prediction from MIMIC: (i)Support Vector Machines (SVM) [9](ii) Random Forests (RF) [9], (iii) XgBoost (XGB) [7], (iv) Fully Connected Network (FCN) [3] (v) Gated Recurrent Unit (GRU) + MLP (multi-layer perceptron) [6]. We also performed a set of ablation studies to understand the utility of learning concepts and the imputation module. Our ablation baselines are as follows: (vi) proposed framework without the concepts and relevance scores (RNN+MLP), (vii) Proposed framework without the missing value imputation and (viii) using only the SOFA organ scores as input features (same network as in (vi)) to predict mortality.

3.5 Implementation details

Both the proposed and baseline models were trained with the same set of longitudinal EHR variables. The dataset was split into training, validation and test set (70:15:15), with the validation set used for early stopping. All deep learning models were trained using the same set of parameter configurations as follows: Adam optimizer with learning rate $= 0.001, \beta_1 = 0.9, \beta_2 = 0.999$ and L2 regularization factor $= 0.0001$ and the number of recurrent layers was set to 3, each of dimension 128. Both the concept network and the relevance network were implemented using fully connected layers of dimensions 256, 12, 64 respectively. All the models were trained for 500 epochs with batch size 128 and dropout rate $= 0.5$. The non-neural models (SVM, RF and XGB) were implemented using Scikit-learn 1.0.1 (default parameters) with Python. The deep learning models were implemented using PyTorch 1.10.2.

3.6 Evaluation metric

We evaluated the explainability and performance of our framework based on the following criteria: (i) **explainability-performance tradeoff**: does the self-explaining nature of our model sacrifice prediction performance? (ii) **added cost of explainability**: extra computational cost (training time) of our proposed model compared to baselines (iii) **clinically meaningful:** are the explanations clinically informative for clinicians and (iv) **faithfulness:** are the relevance scores indicative of "true" importance? Prediction performance of all models were compared using the Area under the Receiver Operating Characteristics (AUROC) and Area under the Precision Recall (AUPRC) curves.

4 Experimental Results

4.1 Explainability-performance trade-off

Our proposed model has better prediction performance (AUROC/AUPRC) compared to the both state-of-the-art machine learning and deep learning models on ICU mortality prediction using MIMIC (Table 1). We believe that the improved performance of our proposed model can be attributed to missing value imputation using recurrent dynamics and using the anticipated organ-failure scores to predicted mortality. Performance metrics of the non-deep learning models are comparatively lower, suggesting that deep neural networks are necessary for predictive modelling with high-dimensional complex datasets. The ablation model using only the six SOFA organ-specific scores as features performs poorly, indicating that it's necessary to learn the concepts (anticipated organ-failure risk scores) compared to directly using them as features. Our proposed model without the missing value imputation shows significant drop in performance. Our proposed model without the concepts and relevance scores does not lead to any performance improvement. This suggests that the self-explaining nature of our proposed model does not sacrifice prediction performance.

4.2 Added complexity of explainability layer

Our proposed model more training time than the baselines, but it's still within a reasonable limit (6 minutes). We believe that the added computational cost is due to the missing value

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
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
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
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

imputation and the additional concept and relevance network. We believe that the trade-off between performance and computational cost of our model is reasonable.

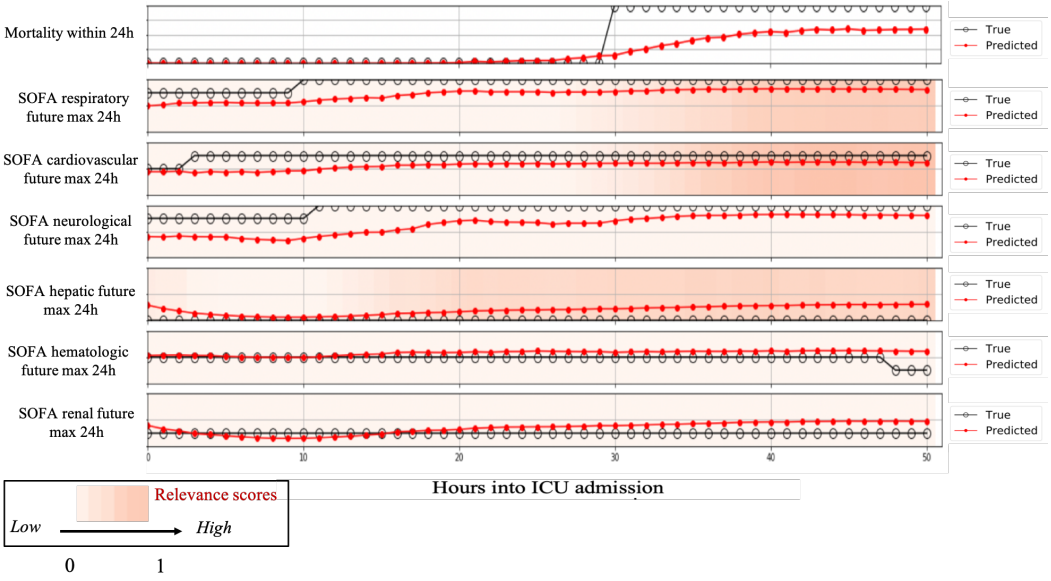

Figure 3: Explaination-relevance visualization of a single patient who died 56 hours after ICU admission. The topmost plot represents mortality probability while the following points represent the explanations. The x-axis represents time (hours into ICU admission) and the plotted values represent either the ground truth label (black) or the predicted value (red).The colormap within each row indicates the weight/importance (relevance score) given to the explanations where dark hues corresponds to higher relevance given to a particular organ system at a specific time point. Both concepts and relevances were scaled between 0 and 1 for easier interpretation.

4.3 Clinical interpretation : Are the explanations clinically informative for clinicians?

We investigated the explanation relevance visualization of the longitudinal trajectory (death) of a single patient who died 56 hours after ICU admission (Figure 3). At each timestep, the clinical outcome (mortality) labels from t=0 to t= 32 will be 0 (survival) and the the labels will be 1 from t=32 onwards. As the predicted probability of mortality rises, the model is shown to pay more importance to anticipated failures in the respiratory, cardiovascular and hepatic organ systems, highlighting their contribution towards mortality. The predicted auxiliary concepts along with the relevance scores are clinically informative in the sense that they can inform clinicians 24 hours in advance the potential organ-system failures that can lead to patient mortality.

4.4 Are relevance scores indicative of true concept importance?

In MIMIC IV, the exact reason behind patient mortality is not available.We measured the faithfulness of relevance scores with respect to the model using a proxy notion of importance: observing the effect of removing features on the model's prediction. In this case, we dropped each of the six concepts (6 SOFA organ-failure scores) during inference time, measured the drop in predicted probability of mortality, and compared them with the corresponding concept's relevance scores. The relevance scores (aggregated across all patients and timepoints) are correlated with the probability drop with for each concept. (Figure 4 left).SOFA respiratory, SOFA cardiovascular and SOFA hepatic with higher relevance scores have a greater drop in mortality probability indicating the importance of these organ-failure risk scores for estimating mortality.

4.5 Grounding of concepts

Our proposed model performs relatively well in predicting risk scores for all the six organ systems (Figure 4 right). This indicates that the predicted organ scores are grounded in terms of expert knowledge and are clinically meaningful choices to select as concepts for our framework.

5 Discussion and Conclusion

In this work, we designed a novel deep learning framework that predicts domain-knowledge driven intermediate high-level clinical concepts from input features and uses them as units of explanation. Our framework is self-explaining; relevance scores are generated for each concept to predict and explain in an end-to-end joint training scheme. Experiments on a real-world electronic health records dataset suggest that (i) the self-explaining nature of our model does not sacrifice prediction performance, (ii) the generated explanations for clinically informative and (iii) the relevance scores reflects true importance (contribution) of the concepts towards the final outcome.

Due to the ever-increasing volume of EHR data, clinicians often rely on existing intermediate knowledge derived from clinical variables for their diagnosis. With the availability of intermediate knowledge as concepts specific to each clinical prediction problem, our proposed framework is generalizable and can be applied on other datasets too. Future work includes learning the plausible clinical concepts in an unsupervised manner without relying on expert knowledge.

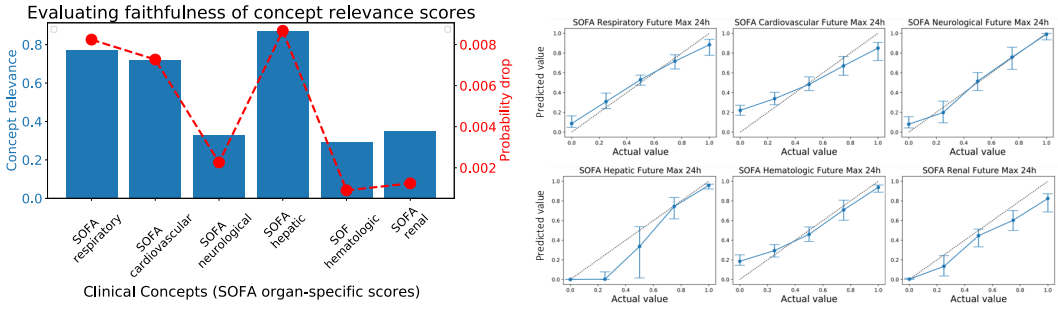

Figure 4: **Left:** Relation between the relevance score and corresponding probability drop for each concept. **Right:** Performance of our proposed model on the auxiliary concept prediction (aggregated across all patients and all timepoints). SOFA organ scores ranging between (0,1,2,3,4) were scaled between (0,1) to (0,0.25,0.5,0.75.1). The dotted line in each figure represents the ideal scenario where the predicted and actual values are same.

References

[1] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

[2] Ekaba Bisong. Introduction to scikit-learn. In *Building machine learning and deep learning models on Google cloud platform*, pages 215–229. Springer, 2019.

[3] William Caicedo-Torres and Jairo Gutierrez. Iseeu: Visually interpretable deep learning for mortality prediction inside the icu. *Journal of biomedical informatics*, 98:103269, 2019.

[4] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems*, 31, 2018.

[5] Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 507–516, 2015.

[6] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Interpretable deep models for icu outcome prediction. In *AMIA annual symposium proceedings*, volume 2016, page 371. American Medical Informatics Association, 2016.

[7] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[8] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.

[9] MIT Critical Data, Joon Lee, Joel A Dubin, and David M Maslove. Mortality prediction in the icu. *Secondary analysis of electronic health records*, pages 315–324, 2016.

[10] Flavio Lopes Ferreira, Daliana Peres Bota, Annette Bross, Christian Mélot, and Jean-Louis Vincent. Serial evaluation of the sofa score to predict outcome in critically ill patients. *Jama*, 286(14):1754–1758, 2001.

[11] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

[12] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[13] David C Kale, Zhengping Che, Mohammad Taha Bahadori, Wenzhe Li, Yan Liu, and Randall Wetzel. Causal phenotype discovery via deep networks. In *AMIA Annual Symposium Proceedings*, volume 2015, page 677. American Medical Informatics Association, 2015.

[14] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280. Springer, 2019.

[15] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020.

[16] I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pages 5491–5500. PMLR, 2020.

[17] Owen Lahav, Nicholas Mastronarde, and Mihaela van der Schaar. What is interpretable? using machine learning to design interpretable decision-support systems. *arXiv preprint arXiv:1811.10799*, 2018.

[18] Thomas A Lasko, Joshua C Denny, and Mia A Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6):e66341, 2013.

[19] Diana Mincu, Eric Loreaux, Shaobo Hou, Sebastien Baur, Ivan Protsyuk, Martin Seneviratne, Anne Mottram, Nenad Tomasev, Alan Karthikesalingam, and Jessica Schrouff. Concept-based model explanations for electronic health records. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 36–46, 2021.

[20] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10, 2016.

[21] Seyedeh Neelufar Payrovnaziri, Zhaoyi Chen, Pablo Rengifo-Moreno, Tim Miller, Jiang Bian, Jonathan H Chen, Xiuwen Liu, and Zhe He. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *Journal of the American Medical Informatics Association*, 27(7):1173–1185, 2020.

[22] Benjamin Shickel, Tyler J Loftus, Lasith Adhikari, Tezcan Ozrazgat-Baslanti, Azra Bihorac, and Parisa Rashidi. Deepsofa: a continuous acuity score for critically ill patients using clinically interpretable deep learning. *Scientific reports*, 9(1):1–12, 2019.

[23] J-L Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, CK Reinhart, PeterM Suter, and Lambertius G Thijs. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure, 1996.

[24] Akbar K Waljee and Peter DR Higgins. Machine learning in medicine: a primer for physicians. *Official journal of the American College of Gastroenterology| ACG*, 105(6):1224–1226, 2010.

[25] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Pradeep Ravikumar, and Tomas Pfister. On concept-based explanations in deep neural networks. 2019.