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

Sayantan Kumar [1] **Sean Yu** [2] **Thomas Kannampallil** [2 3] **Zachary Abrams** [2] **Andrew Michelson** [2 4]
Philip R.O. Payne [2]

Abstract

Complex deep learning models show high prediction tasks in various clinical prediction tasks but their inherent complexity makes it more challenging to explain model predictions for clinicians and healthcare providers. Existing research on explainability of deep learning models in healthcare have two major limitations: using post-hoc explanations and using raw clinical variables as units of explanation, both of which are often difficult for human interpretation. In this work, we designed a self-explaining deep learning framework using the expert-knowledge driven clinical concepts or intermediate features as units of explanation. The self-explaining nature of our proposed model comes from generating both explanations and predictions within the same architectural framework via joint training. We tested our proposed approach on a publicly available Electronic Health Records (EHR) dataset for predicting patient mortality in the ICU. In order to analyze the performance-interpretability trade-off, we compared our proposed model with a baseline having the same set-up but without the explanation components. Experimental results suggest that adding explainability components to a deep learning framework does not impact prediction performance and the explanations generated by the model can provide insights to the clinicians to understand the possible reasons behind patient mortality.

[1]Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, MO, USA [2]Institute for Informatics, Washington University School of Medicine, St. Louis, MO, USA [3]Department of Anaesthesiology, Washington University School of Medicine, St. Louis, MO, USA [4]Department of Pulmonary Critical Care and Medicine, Washington University School of Medicine, St. Louis, MO, USA. Correspondence to: Sayantan Kumar <sayantan.kumar@wustl.edu>.

Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML).

1. Introduction

Linear machine models such as logistic regression and shallow decision trees have been successfully employed in the healthcare domain due to their inherent interpretative nature and are widely used by clinicians for clinical predictive tasks such as disease diagnosis (Bonner, 2001; Yao et al., 2005). However, these models can often easily perform poorly on large heterogeneous clinical datasets. On the other hand, complex deep learning models (particularly neural networks) have been shown to achieve high levels of performance in various downstream healthcare tasks because of their ability to detect complex patterns in the data (Lasko et al., 2013; Kale et al., 2015; Miotto et al., 2016; Che et al., 2015). However, the inherent complexity of black-box deep learning models makes it more challenging to explain model predictions especially for those unfamiliar, specially clinicians and healthcare providers (Waljee & Higgins, 2010; Lahav et al., 2018). This trade-off between model performance and interpretability can lead to an important research question: how can we develop deep learning models having high predictive accuracy and at the same time can be easily interpreted and understood by health care professionals? (Che et al., 2016) In spite of considerable research in recent years, there exists no single widely-accepted definition of explainability or interpretability of deep learning models (Karim et al., 2018). Model developers might be more interested in the working mechanism of the model for the purpose of debugging while clinicians might focus on understanding the rationale behind the clinical predictions obtained as model output (Ras et al., 2022). In order to avoid any ambiguity in the paper, we define both explainability and interpretability as the extent to which the model produces explanations about its predictions that are generally accepted as being understandable, useful, and plausible by subject matter experts (clinicians and healthcare providers) (Tonekaboni et al., 2019). We will use the terms explainability and interpretability interchangeably throughout the manuscript, both referring to the same definition as above.

In the past, multiple approaches have been proposed to provide explanations for deep learning models applied on clinical data [see (Payrovnaziri et al., 2020) for review].

However, these approaches suffer from two major drawbacks: post-hoc explanations and using raw clinical variables as units of explanation. Post-hoc explanation methods are usually motivated by the trade-off between predictive performance and interpretability. Hence, instead of modifying the existing model architecture with the risk of a lower predictive performance, post-hoc explanations accompany the original model as a separate model to provide insights about model predictions (Li et al., 2018). However, post-hoc explainability methods have a key issue of ownership. If an explanation for a given prediction is incorrect, it may be difficult to isolate whether the original model was responsible for this behaviour, or the explanation method generated the error. Due to these challenges, post-hoc explanations in healthcare are often not fully accepted by clinicians and hence, are not reliable enough to integrate into clinical workflow (Rudin, 2019). A potential solution to this problem is developing self-explaining models which can learn self-explainable representation (i.e. no need for training separate models for explanation) and have a tight association with the prediction task via joint training (mutual benefits for both accurate prediction and accurate explanation) (Alvarez Melis & Jaakkola, 2018).

Another challenge of existing explainability approaches in healthcare is to explain the model predictions in terms of the raw clinical features used as input to deep learning models. Weights are assigned to individual features highlighting their contribution (importance) towards model prediction, e.g. logistic regression (Thomas et al., 2008), saliency maps (Kindermans et al., 2019) and Shapley explanations (Kumar et al., 2020). However, using raw features (e.g. pixels in medical images) as units of explanation is often difficult for human interpretation and can lead to unstable explanations that are sensitive to noise or perturbations in the input. To mitigate some of these challenges, we can operate on higher-level features or feature intermediates, which can be referred to as high-level concepts 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 (Koh et al., 2020).

In this work, we aim to address both the above-mentioned challenges of generating deep learning explanations when applied in the healthcare setting. We designed a self-explaining deep learning framework using the expert-knowledge driven clinical concepts or intermediate features derived from raw clinical variables as units of explanation. Our proposed framework uses Sequential Organ Failure Assessment (SOFA) scores as high-level concepts to explain patient mortality in the ICU. SOFA score is a composite score derived from the organ-based sub scores that measure an individual patient's degree of organ dysfunction using raw clinical data and can be tracked over time. The SOFA score is a well-validated metric associated with ICU mortality outcome and functions as an aggregated assessment of clinical status thus, these scores can become ideal candidates to be units of explanation and provide insights into the reasoning behind mortality prediction.

In order to address the performance-interpretability trade-off, we designed our framework to be intrinsically interpretable by jointly predicting and explaining the predictions. We leverage the benefits of multi-task learning to predict the explanations or concepts as supervised auxiliary tasks and utilize the predicted explanations to predict the final outcome, both in a joint end-to-end training setting. The self-explaining nature of our proposed model comes from the relevance scores we generate for each concept within the framework itself without any additional training effort. The relevance scores can provide insights into the question of "why did the patient die?" and quantifies the importance of each concept (SOFA organ system score) in deciding the final mortality outcome. Our contributions can be summarized as follows:

- To the best of our knowledge, our proposed approach is the first study which uses both supervised high-level clinical concepts and intrinsically developed model explanations in the context of predicting a clinical outcome.
- We test our proposed approach on a publicly available longitudinal Electronic Health Records (EHR) dataset for predicting anticipated mortality within the next 24 hours at each point in the ICU trajectory.
- We answer the following research questions through our experiments: (a) Does adding explainability components to a deep learning framework affect its prediction performance (interpretability-performance trade-off), (b) Are the predicted explanations grounded in terms of expert domain knowledge? and (c) Do the explanations generated by our method provide insights into patient mortality?

2. Related Work

In the following section, we briefly describe the studies related to our work. We divide the related research works into 2 broad categories: (i) interpretability methods for neural networks, describing both existing post-hoc and intrinsically explainable neural network based frameworks and (ii) methods which provide explanations through high-level concepts or prototypes.

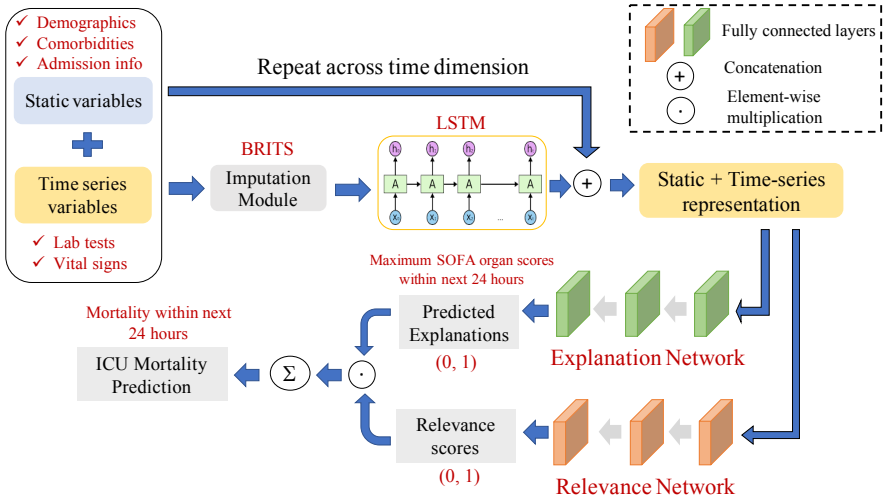

Figure 1. The time-series variables (lab tests and vital signs) are passed through an imputation module BRITS and the imputed variables are propagated through recurrent (LSTM) layers to yield a latent (hidden) representation. The final latent representation, formed by concatenating the static variables (repeated across time dimension) is passed through a series of fully connected neural network layers to generate predicted explanations or auxiliary scores (predicted maximum SOFA organ scores within next 24 hours). The relevance scores (weights/significance) of each concept are generated through a separate set of fully connected layers. The weighted sum of the explanations and relevance scores can be used to yield the predicted probabilities of mortality within next 24 hours.

2.1. Interpretability methods for neural networks

There exist several interpretability methods for neural networks which mostly focus on gradient and perturbation-based methods mentioned here (Bach et al., 2015; Shrikumar et al., 2017; Simonyan et al., 2014; Sundararajan et al., 2017). All these methods do not modify the existing architecture, but explain model predictions in terms of the importance values (weights) of the input features or sensitivities of the inputs with change in the outcome. Our proposed approach differs from the above studies in both the units of explanation - high level concepts instead of raw features and how they are used, relying on the relevance scores produced intrinsically by the model, eliminating the need for additional computation. Studies focusing on intrinsic explanations associated with medical prediction tasks have mostly focused on attention mechanism to find a set of input variables with the most relevant information to the prediction task. Attention has been used to (i) highlight when input features have influenced predictions of clinical events of ICU patients (Kaji et al., 2019), (ii) design an interpretable acuity score based on DeepSOFA that can evaluate a ICU patient's severity of illness (Shickel et al., 2019) and (iii) learn a representation of EHR data that captures the relationships between clinical events for each patient (Patient2Vec) (Zhang et al., 2018). Choi et al. (Choi et al., 2016) designed a reverse time attention model (RETAIN) which mimic the behaviour of a medical professional and incorporate sequential information. Kwon et al. (Kwon et al., 2018) developed a visually interpretable deep learning framework for heart failure based on RETAIN. Unlike our proposed approach, all the above attention based methods use individual clinical features as units of explanation and do not use high level feature concepts or any kind of aggregated knowledge.

2.2. Explanations through concepts and prototypes

A recent study (Alvarez Melis & Jaakkola, 2018) used a self-explainable neural network, learning concepts in an unsupervised way to explain model predictions. We adopt a similar approach in the context of clinical predictive modelling and design high-level clinical concepts which can be learned in a supervised way driven by medical expert-knowledge. Li et al. (Li et al., 2018) proposes an interpretable deep learning framework whose predictions are based on the similarity of the input to a small set of prototypes which are learned during training. Kim et al. (Kim et al., 2018) follows a post-hoc approach of learning interpretable concepts to explain model predictions. Their proposed model learns concept activation vectors representing human friendly concepts of interest and the directional derivatives along these vectors can be used to estimate the sensitivity of the predictors with respect to semantic changes in the direction of the concept. Mincu et al. (Mincu et al., 2021) extended the TCAV concept to longitudinal EHR data. Their approach differs from ours is that the contribution of the concept to the primary task can only be determined globally (through aggregation over multiple samples) and not locally (for each patient). Our proposed approach allows local explanations through

the relevance scores that can explain the mortality of each patient.

3. Proposed Framework

In this section, we describe our proposed explainability framework (Figure 1). Our method has a multi-task setting where the explanations or high-level concepts are predicted as auxiliary tasks and the predicted auxiliary scores are used to generate the final prediction output. Figure 1 shows all the components of our proposed framework in detail.

Let X be a multivariate time series of a particular patient which can be represented by a sequence of T observations $\{x_t\}$, where $t = 1, 2, ...T$. We denote the set of time invariant or static features of that patient as $\{x_s\}$. The time-series features are passed through an imputation algorithm named BRITS which imputes the missing values according to recurrent dynamics (Cao et al., 2018). In BRITS, the imputed variables are regarded as variables in a bidirectional RNN graph. Hence the missing values get delayed gradients in both forward and backward direction, which makes the estimation of missing values more accurate. The imputation is supervised by the imputation loss function L_{impute} which can be calculated by the Mean Absolute Error (MAE) between the actual and imputed variables, as shown in Equation 4. The imputed variables $\{\tilde{x_t}\}$ (Equation 2) are then passed through a series of recurrent Long Short Term Memory (LSTM) layers to yield a latent representation $\{h_t\}$. We define a mask m_t (Equation 1) which keeps track of the time points where a particular feature variable is missing. In case of missing time points, we replace the missing values in x_t with the corresponding values in x_{impute} (Equation 3). The latent representation obtained as LSTM output is then concatenated with the time-invariant variables $\{x_s\}$, repeated across time dimension to form the final latent representation f (Equation 5)

$$m_t = \begin{cases} 0 & x_t \text{ is not observed at timestep t} \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

$$x^t_{impute} = BRITS(\{x_t\}) \tag{2}$$

$$\tilde{x_t} = m_t * x_t + (1 - m_t) * x^t_{impute} \tag{3}$$

$$L_{impute} = \sum_{i=1}^{T} |x_t - x^t_{impute}| \tag{4}$$

$$f = \tilde{x_t} \oplus x_s \tag{5}$$

The concatenated representation f is then passed through a series of fully-connected layers, followed by a sigmoid activation function to generate the auxiliary layer output or predicted explanations. The auxiliary task performance is supervised by the auxiliary loss L_{aux}, which is calculated by the Mean Squared Error (MSE) between the predicted (y^{pred}_{aux}) and ground truth (y_{aux}) labels (Equation 6). The number of auxiliary tasks or explanation units is a predefined value based on the number of expert-knowledge driven medical concepts. We also generate relevance scores for each concept, which is estimated by passing the concatenated representation through a separate set of fully connected layers, followed by the sigmoid activation function to make the relevance scores range between 0 and 1. The relevance score of a concept at each time point signify weight or contribution of that concept in deciding the final predicted probability. Higher weights (close to 1) indicate greater contribution.

The predicted auxiliary values exp_j can be combined with the corresponding relevance scores α_j to generate the final probability y_{pred}, where j denotes the j-th explanation and N denotes the number of explanations (Equation 7). The final loss function is a weighted sum of 3 supervised losses: L_{mort} (Equation 8), binary cross-entropy (CE) loss between ground truth y mortality label y_{pred} and predicted mortality $\hat{y_t}$, the auxiliary loss L_{aux} and imputation loss L_{impute} (Equation 9). Since the imputation loss is dominated by more frequent variables, there exists a high degree of loss imbalance between the imputation loss and the primary and auxiliary losses, which can lead to negative loss transfer and low prediction performance (Li et al., 2020). To address this challenge, we selected higher weights for the primary and auxiliary loss ($\lambda_1 = 1, \lambda_2 = 10$) compared to the imputation loss ($\lambda_3 = 0.001$).

$$L_{aux} = \sum_{i=1}^{T} (y_{aux} - y^{pred}_{aux})^2 \tag{6}$$

$$y_{pred} = \sigma(\sum_{j=1}^{N} (\alpha_j * exp_j)) \tag{7}$$

$$L_{mort} = -(y \log(y_{pred}) + (1 - y) \log(1 - y_{pred})) \tag{8}$$

$$L = \lambda_1 L_{mort} + \lambda_2 L_{aux} + \lambda_3 L_{impute} \tag{9}$$

4. Experiments

We aim to answer the following research questions through our experiments: Does adding explainability components to a deep learning framework affect its prediction performance (interpretability-performance trade-off), (b) Are the predicted explanations grounded in terms of expert domain knowledge? and (c) Do the explanations generated by our method provide insights into patient mortality? In the remainder of this section, we will describe the dataset, details about feature engineering and pre-processing, the baseline methods for comparison and the implementation details.

4.1. Dataset and Experimental design

STUDY PARTICIPANTS

We conduct our experiments on the publicly available Medical Information Mart for Intensive Care IV (MIMIC-IV v0.4) database. MIMIC-IV comprises of more than a decade worth of de-identified ICU patient records, including vital signs, laboratory and radiology reports, and therapeutic data, from patients admitted to the Intensive Care Units of the Beth Israel Deaconess Medical Centre in Boston, Massachusetts, and is freely available for research (Johnson et al., 2016). We established our cohort based on the following eligibility criteria: (i) ICU stays longer than 48 hours and (ii) patients older than 18 patients at the time of admission. Only the first admission was considered in case of multiple admissions to the ICU. Our cohort comprises of clinical data of 22,944 ICU admissions between 2008-2019, of which 2043 (8.9%) experienced in-hospital mortality. The median age of adult patients (age ¿= 18 years old) is 67 years (IQR: 56-78 years) and the median length of stay (LOS) in the ICU is 84.3 hours (IQR: 62.3 - 132.7 hours).

FEATURE VARIABLES AND PREPROCESSING

For each patient, we extracted 24 static or time-invariant features such as demographics (age, gender, race, ethnicity), admission diagnoses and comorbidity information and 87 time-series features which includes laboratory test results and vital signs. Feature pre-processing of time-series variables steps include clipping the outlier values to the 1st and 99th percentile values and standardization using the RobustScalar package from sklearn (Bisong, 2019). Time-varying variables were aggregated into hourly time buckets using the median for repeated values. All categorical features were one-hot encoded. Missing values in time-series variables were imputed using the BRITS algorithm, a state-of-the-art imputation based on recurrent dynamics (Cao et al., 2018).

HIGH-LEVEL CONCEPTS: UNITS OF EXPLANATION

In our work, we used the different components of the Sequential Organ Failure Assessment (SOFA) scores as our units of explanation. SOFA is a widely-used score validated by clinicians and used for assessing severity of illness measuring the extent of a the person's organ function or failure. The score is based on six subscores, one each for the six organ systems: respiratory, cardiovascular, hepatic, coagulation, renal and neurological systems, allocating a score of 0-4 with high scores indicating severe organ conditions (Ferreira et al., 2001). Most in-hospital ICU deaths are preceded by signs of organ failure; thus, these scores can provide insights into the reasoning behind mortality prediction (Yu et al., 2021). At each timepoint within a patient's ICU trajectory, the maximum SOFA sub-score (for each of the six organ systems) within the next 24 hours were used as ground truth labels to supervise the predicted auxiliary score or explanations. In our model, the predicted SOFA organ scores form the six units of explanation, which when combined with the relevance scores can be used to predict the final mortality. The model predicts mortality at each timepoint as the mortality risk or the probability of patient mortality within next 24 hours. The SOFA organ scores for the six organ systems were calculated based on a set of standard rules proposed in (Ferreira et al., 2001; Lambden et al., 2019). All the SOFA organ scores were scaled between 0 and 1 by the MinMax scaling package from sklearn (Bisong, 2019).

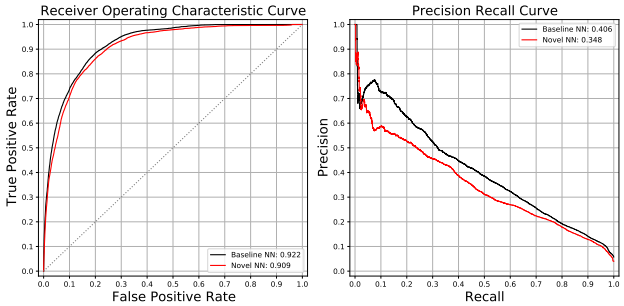

Figure 2. AUROC (AUROC (Area Under the Receiving Operating Characteristics Curve) and AUPRC (Area Under the Precision Recall curve for both our proposed (Novel NN) and baseline method (baseline NN). Both the proposed and baseline model have similar AUROC and AUPRC values indicating that adding explainability components does not impact prediction performance.

4.2. Baseline method and Implementation details

In order to validate our hypothesis that adding explainability components to the network does not affect prediction performance, we compared our proposed approach with a baseline model to analyze the performance-interpretability trade-off. The baseline model has the same set up as our proposed model but without the components generating the auxiliary and relevance scores. In the baseline model framework, the time-series variables are passed through the imputation module BRITS and the imputed variables are propagated through recurrent (LSTM) layers to yield a latent (hidden) representation. The final latent representation, formed by concatenating the static variables (repeated across the time dimension) is passed through a series of fully connected neural network layers, followed by a sigmoid activation function to generate the final mortality probability (mortality within next 24 hours).

Both the proposed and baseline models are trained with the same set of static and time-series input variables. The

dataset was split into training, validation and test set (70:15:15), with the validation set used for early stopping. For a fair comparison, both the proposed and baseline model 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 LSTM layers was set to 3, each of dimension 128. For the imputation module BRITS, we used the same parameter configurations used in the original paper (2 recurrent layers with size 256 each). In the proposed model, the concatenated representation of static and time series variables are fed through 2 separate sets of 3 fully connected layers of size 256, 128, 64 for generating the auxiliary and relevance scores. Both the models were trained for 500 epochs with batch size 128 and dropout rate = 0.5.

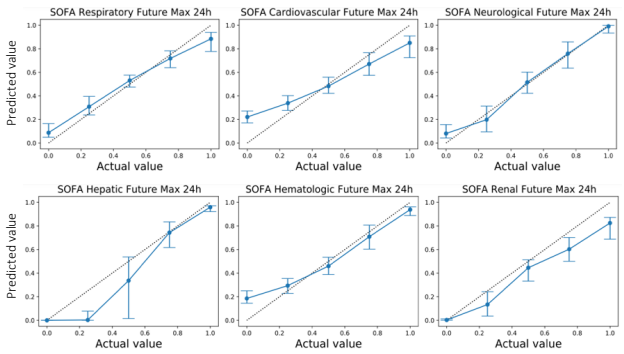

Figure 3. Performance of our proposed model on the auxiliary tasks, showing actual and predicted values of the auxiliary tasks corresponding to the maximum SOFA organ system scores within 24 hours. The dotted line in each figure represents the ideal scenario where the predicted and actual values are same.

5. Results

Figure 2 shows the AUROC (Area Under the Receiving Operating Characteristics Curve) and AUPRC (Area Under the Precision Recall Curve) for both our proposed approach and baseline method without the explainability components (auxiliary scores and relevance scores). Both models have low AUPRC, caused by the low prevalence of mortality, or class imbalance; in our cohort, only 2043 (8.9%) out of 22,944 ICU admissions experienced in-hospital mortality. Both the proposed and baseline model have similar AUROC ($0.923[0.915 - 0.947]$ vs $0.909[0.895 - 0.928], p = 0.676$) and AUPRC ($0.224[0.205 - 0.247]$ vs $0.227[0.198 - 0.237], p = 0.823$) values without any statistically significant differences at 95% confidence interval. This observation addresses our 1st research question regarding performance-interpretability trade-off and indicates that adding intrinsic explainability components does not impact prediction performance of our proposed method.

Figure 3 shows the performance of our proposed model on the auxiliary tasks corresponding to the six SOFA organ systems. For each of the organ systems, the SOFA scores ranging from 0-4 in increasing order of severity are scaled to (0,1). The results suggest that the model performs relatively well in predicting scores for all the organ systems. The decent performance metrics on the auxiliary tasks demonstrate that the predicted auxiliary scores are similar to the actual SOFA organ system scores. In other words, the predicted explanations are grounded in terms of expert knowledge and can act as reliable indicators to provide insights into potential reasons behind the final mortality outcome.

Figure 4 shows the explanation relevance visualization of the longitudinal trajectory of a single patient who died 80 hours after ICU admission. Here we can observe how the predicted explanations (maximum SOFA organ scores within next 24 hours) and their corresponding relevance scores vary at each timepoint throughout the longitudinal trajectory of the patient in the ICU. We can observe that the model initially pays the maximum importance (weight) to the anticipated SOFA hepatic dysfunction till t = 20 hours, followed by anticipated SOFA kidney (renal) dysfunction. 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. The explanations along with the relevance scores in Figure 4 can help clinicians understand the health status of a patient throughout the duration of the ICU stay and analyze possible reasons for mortality (if applicable).

6. Discussion

Our aim was to design a self-explaining deep learning framework using the expert-knowledge derived clinical concepts such as SOFA organ system scores as units of explanation for predicting longitudinal mortality in the Intensive Care Unit (ICU) setting. The self-explaining nature of our proposed model comes from the relevance scores we generate for each concept which can provide insights into the reasoning behind patient mortality. We leverage the benefits of multi-task learning to predict the explanations or concepts as supervised auxiliary tasks and utilize the same explanations to predict the final outcome, both in a joint end-to-end training setting, with the aim of producing both accurate predictions and accurate explanations.

6.1. Interpretability - performance trade-off

Interpretability and performance currently stand in apparent conflict in deep learning. Most relevant literature have focused on using a separate post-hoc model for explaining the

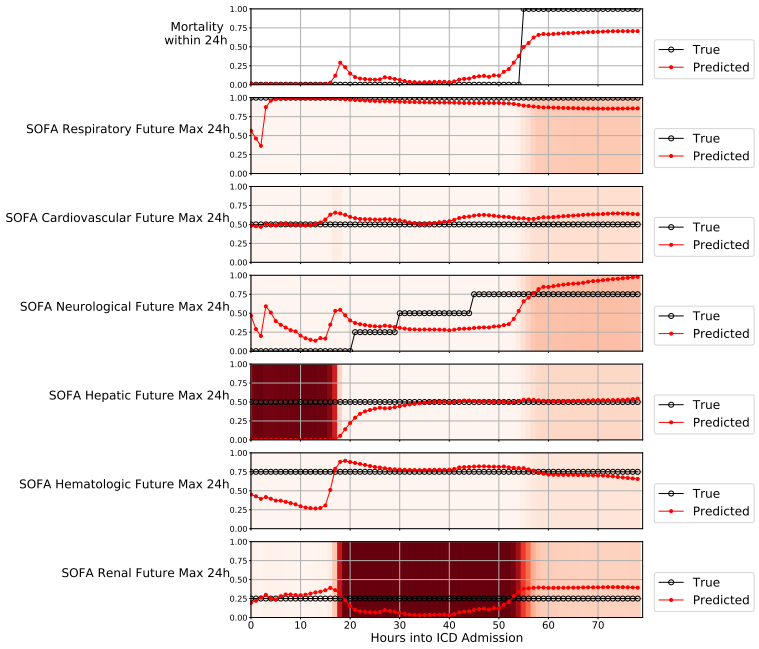

Figure 4. Figure showing how the predicted explanations (maximum SOFA organ scores within next 24 hours) and their corresponding relevance scores at each timepoint throughout the longitudinal trajectory of a particular patient in the ICU. Both the explanations and relevance scores range between 0 and 1. The topmost plot represents mortality 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 highlighting 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.

predictions of a deep learning framework. These methods focus on improving the prediction performance and hence do not modify the existing architecture. In our work, we analyze the interpretability-performance trade-off by designing a framework that is jointly trained to both predict and explain in an end-to-end manner, without any additional training effort. We leveraged the benefits of a multi-task learning by predicting the explanations as auxiliary scores and using the predicted auxiliary scores to generate the final outcome. Both the auxiliary tasks and the final task are learned via joint training in a supervised manner, thus improving the performance on both the final and auxiliary tasks. Our AUROC and AUPRC results demonstrate that model performance does not change significantly even if we add an intrinsic explainability component within the main architecture. We believe that this can motivate future research on designing deep learning frameworks which can predict and explain at the same time without any cost in prediction performance.

6.2. Pre-defining SOFA organ scores as units of explanation

A common approach for generating explanations for a deep learning model when used in the clinical context is to use the raw clinical variables, understanding which of them are correlated with the final prediction and in what capacity. However, for a large number of input clinical features, it often becomes difficult for the clinicians to comprehend the exact reasoning behind the final clinical outcome if we just provide them with a list of features with high weights or importance towards the final prediction. Subject matter experts tend to rely on some intermediate knowledge schemas which make them help decisions. These can be understood as some kind of aggregated knowledge or high-level interpretable concepts, which are derived from the raw clinical variables and more stable with respect to noise and input perturbations. SOFA organ system scores can be a suitable candidate for high-level interpretable concepts that are not only derived from a combination of feature variables but also widely used by clinicians to analyze the risk of patient mortality in the ICU. In our work, we train our model to learn these interpretable concepts from raw clinical variables in a supervised manner through a framework of auxiliary tasks. Our results demonstrate that the predicted explanations are grounded in terms of expert knowledge and can act as reliable indicators to explain the reasoning behind the final mortality outcome. Hence a more natural unit of explanation would be ”the patient died due to organ failures

in the cardiovascular system".

6.3. Generalizability: Application of proposed framework to other clinical problem set-up

Our goal was to design a generalizable interpretable network that can generate both predictions and explanations in any kind of clinical prediction problem. As a starting point, we used ICU mortality prediction as a single use case in our experiments, where the SOFA organ system scores were representatives of the high-level interpretable concepts used as units of explanation. However, our framework can be applied in other clinical domains too. Since we use expert-knowledge driven aggregated knowledge or high-level intermediate concepts in a supervised setting, our only assumption is that the intermediate concepts need to be pre-defined. Clinical experts have designed intermediate constructs or aggregated knowledge for different clinical problems to make decisions. An example of intermediate construct is Clinical Dementia Rating which is derived from individual cognitive performance tasks for predicting Alzheimer's Disease progression. Due to the ever-increasing volume of EHR data, clinicians often rely on similar existing intermediate knowledge derived from clinical variables for their diagnosis. Hence it is possible to obtain the intermediate knowledge as concepts specific to each clinical prediction problem. Thus our proposed framework is generalizable to work on other informatics problems too.

6.4. Limitations and scope for future work

In this work, we applied our proposed explainability framework on a single dataset and a specific clinical prediction problem. Our immediate next step is to test the generalizability of our proposed interpretability framework on multiple EHR datasets. One of the fundamental assumptions of our approach is that the intermediate concepts need to be predefined in the context of the clinical problem. Hence, our explainability framework is not applicable in cases where there exists no available expert knowledge for that particular problem. As a next step, we plan to have our model learn the interpretable concepts in an unsupervised manner without relying on expert knowledge and apply them to more complex domains such as medical imaging, speech recognition and clinical natural language processing.

A potential limitation of SOFA organ system scores is that, it only evaluates a few organs and sometimes markers of damage to one organ actually represent markers of damage to another. For example, bilirubin can be a sign of haemolysis or gallbladder issues and also of liver failure (Yu et al., 2021). Also, chronic/underlying causes that may impact SOFA scores are not accounted for in many of the studies used for clinical validation. However, we re-emphasize that our goal is to analyze if it's possible to generate plausible explanations using an example of expert-knowledge driven intermediate concepts. In our work, SOFA is used as an example test case only which satisfied the characteristics of intermediate aggregated knowledge that can be used by clinicians to make decisions.

7. Conclusion

In this work, we designed a self-explaining deep learning framework with expert-knowledge driven clinical concepts as the units of explanation. The self-explaining nature of our proposed model comes from generating both explanations and predictions via joint training. We tested our proposed approach on the MIMIC IV EHR dataset for predicting patient mortality in the ICU using SOFA organ system scores as the intermediate high-level concepts. The explanations generated by our model along with the corresponding relevance scores help clinicians monitor the health status of a patient and provide insights into possible reasons behind patient mortality. Our experiments analyzed the interpretability-performance trade-off and demonstrate that model performance does not change significantly even if we add an intrinsic explainability component within the main architecture itself. Future work include testing the generalizability of our proposed interpretability framework on multiple EHR datasets and learn the interpretable concepts in an unsupervised manner without relying on expert knowledge.

References

Alvarez Melis, D. and Jaakkola, T. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Bisong, E. Introduction to scikit-learn. In *Building machine learning and deep learning models on Google cloud platform*, pp. 215–229. Springer, 2019.

Bonner, G. Decision making for health care professionals: use of decision trees within the community mental health setting. *Journal of Advanced Nursing*, 35(3):349–356, 2001.

Cao, W., Wang, D., Li, J., Zhou, H., Li, L., and Li, Y. Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems*, 31, 2018.

Che, Z., Kale, D., Li, W., Bahadori, M. T., and Liu, Y. Deep

computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 507–516, 2015.

Che, Z., Purushotham, S., Khemani, R., and Liu, Y. Interpretable deep models for icu outcome prediction. In *AMIA annual symposium proceedings*, volume 2016, pp. 371. American Medical Informatics Association, 2016.

Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., and Stewart, W. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.

Ferreira, F. L., Bota, D. P., Bross, A., Mélot, C., and Vincent, J.-L. Serial evaluation of the sofa score to predict outcome in critically ill patients. *Jama*, 286(14):1754–1758, 2001.

Johnson, A. E., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Kaji, D. A., Zech, J. R., Kim, J. S., Cho, S. K., Dangayach, N. S., Costa, A. B., and Oermann, E. K. An attention based deep learning model of clinical events in the intensive care unit. *PloS one*, 14(2):e0211057, 2019.

Kale, D. C., Che, Z., Bahadori, M. T., Li, W., Liu, Y., and Wetzel, R. Causal phenotype discovery via deep networks. In *AMIA Annual Symposium Proceedings*, volume 2015, pp. 677. American Medical Informatics Association, 2015.

Karim, A., Mishra, A., Newton, M., and Sattar, A. Machine learning interpretability: A science rather than a tool. *arXiv preprint arXiv:1807.06722*, 2018.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280. Springer, 2019.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *International Conference on Machine Learning*, pp. 5338–5348. PMLR, 2020.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pp. 5491–5500. PMLR, 2020.

Kwon, B. C., Choi, M.-J., Kim, J. T., Choi, E., Kim, Y. B., Kwon, S., Sun, J., and Choo, J. Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE transactions on visualization and computer graphics*, 25(1): 299–309, 2018.

Lahav, O., Mastronarde, N., and van der Schaar, M. What is interpretable? using machine learning to design interpretable decision-support systems. *arXiv preprint arXiv:1811.10799*, 2018.

Lambden, S., Laterre, P. F., Levy, M. M., and Francois, B. The sofa score—development, utility and challenges of accurate assessment in clinical trials. *Critical Care*, 23 (1):1–9, 2019.

Lasko, T. A., Denny, J. C., and Levy, M. A. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6):e66341, 2013.

Li, D., Lyons, P. G., Lu, C., and Kollef, M. Deepalerts: deep learning based multi-horizon alerts for clinical deterioration on oncology hospital wards. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 743–750, 2020.

Li, O., Liu, H., Chen, C., and Rudin, C. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Mincu, D., Loreaux, E., Hou, S., Baur, S., Protsyuk, I., Seneviratne, M., Mottram, A., Tomasev, N., Karthikesalingam, A., and Schrouff, J. Concept-based model explanations for electronic health records. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 36–46, 2021.

Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10, 2016.

Payrovnaziri, S. N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J. H., Liu, X., and He, Z. 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.

Ras, G., Xie, N., van Gerven, M., and Doran, D. Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73:329–397, 2022.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Shickel, B., Loftus, T. J., Adhikari, L., Ozrazgat-Baslanti, T., Bihorac, A., and Rashidi, P. Deepsofa: a continuous acuity score for critically ill patients using clinically interpretable deep learning. *Scientific reports*, 9(1):1–12, 2019.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMLR, 2017.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*. Citeseer, 2014.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Thomas, D. R., Zhu, P., Zumbo, B. D., and Dutta, S. On measuring the relative importance of explanatory variables in a logistic regression. *Journal of Modern Applied Statistical Methods*, 7(1):4, 2008.

Tonekaboni, S., Joshi, S., McCradden, M. D., and Goldenberg, A. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pp. 359–380. PMLR, 2019.

Waljee, A. K. and Higgins, P. D. Machine learning in medicine: a primer for physicians. *Official journal of the American College of Gastroenterology— ACG*, 105(6): 1224–1226, 2010.

Yao, Z., Liu, P., Lei, L., and Yin, J. R-c4. 5 decision tree model and its applications to health care dataset. In *Proceedings of ICSSSM'05. 2005 International Conference on Services Systems and Services Management, 2005.*, volume 2, pp. 1099–1103. IEEE, 2005.

Yu, S. C., Shivakumar, N., Betthauser, K., Gupta, A., Lai, A. M., Kollef, M. H., Payne, P. R., and Michelson, A. P. Comparison of early warning scores for sepsis early identification and prediction in the general ward setting. *JAMIA open*, 4(3):ooab062, 2021.

Zhang, J., Kowsari, K., Harrison, J. H., Lobo, J. M., and Barnes, L. E. Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record. *IEEE Access*, 6:65333–65346, 2018.

You can have as much text here as you want. The main body must be at most 8 pages long. For the final version, one more page can be added. If you want, you can use an appendix like this one, even using the one-column format.