Towards Better Interpretability of Sepsis Prediction by Deep Neural Networks with Variable-wise Attribution Maps

P.-E Thiboud, V Wargnier-Dauchelle, M Lefort, N Duchateau, M Sdika

TOWARDS BETTER INTERPRETABILITY OF SEPSIS PREDICTION BY DEEP NEURAL NETWORKS WITH VARIABLE-WISE ATTRIBUTION MAPS

P.-E. Thiboud[1,2,3] *V. Wargnier-Dauchelle*[1] *M. Lefort*[2] *N. Duchateau*[1,4] *M. Sdika*[1]

[1] INSA-Lyon, UCBL, CNRS, Inserm, CREATIS UMR 5220, U1294, Lyon, France
[2] Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS UMR 5205, Villeurbanne, France
[3] PREVIA MEDICAL, Lyon, France
[4] Institut Universitaire de France (IUF)

ABSTRACT

Because of the multi-symptomatic nature of sepsis, its prediction is challenging as it requires considering subtle changes in multiple monitored variables across time. Recent works based on deep neural networks improved the prediction performance, but still suffer from poor interpretability. Critical for healthcare applications, we propose to improve this aspect by separating time variables in the convolution layers of a sepsis prediction network. We reveal the improvement in interpretability capacity with the use of gradient-based attributions on high-level intermediate features and through a metric correlating the variable attribution with the prediction for perturbed pathologic samples. With 171,945 patients from the MIMIC-IV database, we demonstrate that our method not only maintains classification performances at similar network parameters count, but also substantially improves the faithfulness of per-variable attributions. This enhanced interpretability has the potential to improve clinical decision-making by enabling practitioners to swiftly identify critical variables, which could streamline patient monitoring workflows.

***Index Terms*—** Sepsis, deep learning, interpretability.

1. INTRODUCTION

Sepsis is a life-threatening condition characterized by a dysregulated immune response to infection, leading to tissue and organ damage that can result in shock, multiple organ failure, and death. It remains a major cause of morbidity and mortality, with $\sim$49 million cases and 11 million sepsis-related deaths occurring worldwide in 2017, accounting for $\sim$20% of all-cause deaths globally [1]. Early detection and treatment of sepsis are crucial for improving patient outcomes. For every hour delay in antibiotic administration, there is an estimated 9% increase in mortality [2]. Automated screening tools can continuously monitor patient data from electronic health records, and therefore could decrease diagnostic delays and increase screening accuracy. Prediction of sepsis is commonly approached as a multi-channel time series classification problem. Recurrent Neural Networks (RNNs) have traditionally been used for such tasks, as demonstrated in the MGP-RNN approach [3]. Temporal Convolutional Networks (TCNs) have proven to be more effective than RNNs for processing long sequences [4]. Building on this, MGP-TCN [5] combined the MGP approach (a more refined imputation strategy) with TCNs, outperforming previous RNN-based models in sepsis prediction tasks.

Besides, explainability of prediction models used is a clear asset for their adoption by clinicians, especially in the context of sepsis detection [6]. Gradient-based attributions are widely used post-hoc methods to interpret deep learning models. For example, Input*Gradient [7] simply multiplies the input by the gradient of the output with respect to the input, while Integrated Gradients [8] integrates gradients along a straight path from a baseline input to the actual input.

The interpretability problem has also been approached for temporal health signals, with RETAIN (Reverse Time Attention model) [9] and the Attention-based Temporal Convolutional Network (AttTCN) [10]. Both methods focus on the most relevant time steps and features for prediction using attention maps. By using these maps at inference, they can detect influential past visits and significant clinical variables within those visits. The difference between the two methods lies mainly in the architecture used for the embedding before the attention maps: a linear layer followed by a RNN or a TCN. Nevertheless, in addition to a complex architecture, like the attributions methods described above, it is difficult to extract the importance of each temporal signal in the decision as the high-level features, which are most relevant to the decision, are an entanglement of all input signals.

Contributions We propose a novel and simple approach to enhance interpretability in sepsis prediction by introducing variable-wise convolutions in TCNs, namely convolutions respecting a clear separation between the input time series of each variable. This modification establishes, by construction, a direct relationship between each input variables and inter-

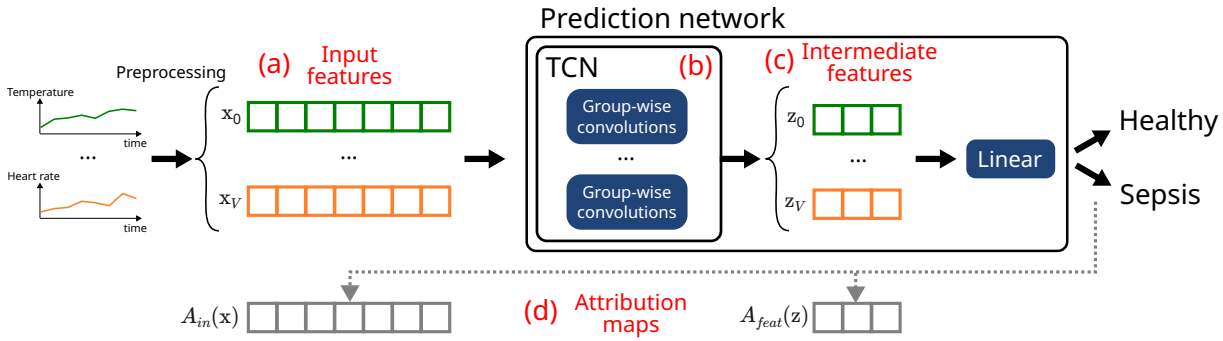

Fig. 1. Overview of the processing pipeline we propose. Preprocessing of raw and irregularly sampled time series data creates V hour-binned input features $\mathbf{x} \in \mathbb{R}^D$ (a) with D the window size, which are then processed with a TCN model (b) to generate V intermediate feature vectors $\mathbf{z} \in \mathbb{R}^d$ of dimensionality d (c). Our proposed modification on the TCN convolutions allows generating variable-wise attribution maps (d) from both input features $A_{in}(\mathbf{x}) \in \mathbb{R}^D$ and intermediate feature vectors $A_{feat}(\mathbf{z}) \in \mathbb{R}^d$.

mediate feature vectors and, as we establish experimentally, enables more accurate attribution of feature importance. Our method allows for *(1)* the computation of attribution methods directly on variable-wise intermediate feature vectors, and *(2)* direct gradient-based attributions availability through layer's weights. Experimental results demonstrate that our approach maintains classification performance while improving the correlation between variable attribution and effective pathological prediction.

2. METHOD

This section details the key stages of our pipeline (Fig.1) to classify samples as healthy or sepsis: the processing of input signals to obtain intermediate features (Sec.2.1), and the way we foster interpretability (Sec.2.2 and 2.3).

2.1. Temporal Convolutional Network

First introduced by [4] and adapted for sepsis prediction by [5], TCN models utilize dilated causal convolutions [11] to effectively capture long-range dependencies in time series data while maintaining causality. Predictions at any time step only depend on past and present inputs. For a single prediction, the last value of each channel is selected and passed through a linear layer. The output corresponds to features that should better represent the input time series: we build upon this TCN architecture with a rather simple modification aiming at the interpretability of sepsis prediction (Sec.2.2).

2.2. Variable-wise interpretability

In standard convolutional networks, all channels are mixed from the very first layer. This way, the model learns highly complex interactions between variables but all subsequent feature maps become hardly interpretable. In our case, the input data are time series of a set of variables (temperature, heart rate, ...). We propose to separate the processing of individual time variables within the network's convolutions to get a direct relation between input variables and intermediate feature vectors. This is achieved through the use of grouped convolutions [12]. With grouped convolutions with V groups (V being the number of variables), intermediate representations have V blocks of channels where each block is a representation of the corresponding variable only. When post-hoc attributions are computed at intermediate levels, the values of each of the V attribution blocks can directly be related to the corresponding input variable. This enables a direct interpretation of the intermediate representation within the network. Of note, an interesting consequence of grouped convolutions is a strong reduction of the number of parameters of the network.

2.3. Intrinsic attribution maps

Following previous works using TCN models for sepsis prediction [5], intermediate features are processed by a linear layer to give the final prediction. The simplicity of this layer makes gradient-based attribution methods on these intermediate features directly readable through the layer's weights.

Input*Gradient [7] Attribution maps on intermediate features can be reduced to a scaled version of the Hadamard product between intermediate feature vectors $\mathbf{z}$ and the corresponding linear layer weights $\mathbf{w}$, as: $A_{feat}(\mathbf{z}) = (\mathbf{z} \odot \mathbf{w})\ \sigma'(\mathbf{w}^T\mathbf{z})$ where σ is the sigmoid function at the end of the network.

Integrated Gradients [8] Attributions are computed as:

$$A_{in}(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}') \odot \int_{\alpha=0}^{1} \frac{\partial F}{\partial \mathbf{x}} (\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}'))\, d\alpha, \quad (1)$$

where $\mathbf{x}'$ is the baseline, F stands for the prediction network and α is a path integration parameter. When computed just before the last linear layer, it becomes:

$$A_{feat}(\mathbf{z}, \mathbf{z}') = ((\mathbf{z}-\mathbf{z}')\odot\mathbf{w}) \int_{\alpha=0}^{1} \sigma'\left(\mathbf{w}^T\left(\mathbf{z}' + \alpha(\mathbf{z}-\mathbf{z}')\right)\right) d\alpha. \tag{2}$$

3. EXPERIMENTS

Our method was evaluated on a dataset of 171,945 patients (with 80-10-10 train/valid/test splits) extracted from MIMIC-IV [13] with both Intensive Care Unit (ICU) and Emergency Department (ED) [14] patients. Sec.3.1 describes the data used, including the time variables extracted and the preprocessing steps applied, while Sec.3.2 focuses on the metrics used to assess the classification performances and increased interpretability provided by our method, and Sec.3.3 describes various implementation details.

3.1. Data & Preprocessing

For each patient, 15 time variables were extracted: 6 vital signs (heart rate, systolic and diastolic blood pressure, respiration rate, temperature, and oxygen saturation), 8 laboratory values (partial pressure of oxygen - PaO2, fraction of inspired oxygen - FiO2, lactate, platelets, white blood cells - WBC, creatinine, bilirubin, and diuresis) and 1 level of consciousness assessment score (Glasgow Coma Scale - GCS). Sepsis cases were retrospectively identified through ICD-10 codes. Sepsis onset was defined as the first instance of a 2-point increase in SOFA score, a simplified version of Sepsis-3 [1].

For data preparation, irregularly sampled time series were hour-binned, taking the last available measurement for each hour. Then, missing values were handled through a carry-forward imputation. 12-hour time windows were created, starting from 11 hours before onset and extending to the onset time. For non-sepsis patients, a random onset was selected over the entire stay.

3.2. Metrics analyzed

Classification metrics Following previous works on sepsis detection [3, 5], we evaluate the models performances on the test split in terms of AuPRC, Precision, and Recall, which are better suited for unbalanced data as in many sepsis datasets. AuROC is also included for comparison purposes.

Perturbation-correlation metric To determine the faithfulness of an attribution map to the inner workings of a model, the Infidelity metric [15] proposes to perturb inputs and measure the difference in model prediction. Building on this, we define a strong perturbation of a subset of input features of a pathological case as the replacement of a time variable by the same variable from a random sample of the healthy class. Then the Pearson correlation coefficient is computed between the attribution value of this variable ($A_{in}(\mathbf{x})$ or $A_{feat}(\mathbf{z})$) and the prediction drop, averaged over multiple healthy variable replacements.

However, attribution maps have the same dimensionality as their corresponding features, and may remain hard to interpret. To get a single value from the attribution maps of each variable, we devised two *aggregation* methods: summing attribution values per variable before computing the correlation (*Sum*), or taking the maximum value of feature-by-feature correlations attained by variable (*Max*). For models without variable-wise convolutions, intermediate features were randomly grouped to mimic per-variable intermediate feature vectors of variable-wise models and provide a comparison point. Models with and without variable-wise convolutions are compared on both input and intermediate features, with Input*Gradient and Integrated Gradients attribution methods using both aggregation methods (*Sum* and *Max*).

3.3. Implementation details

Following [5], in the TCN part of the network, we keep the same number of convolution filters C for every convolution layer, with 4 temporal blocks each composed of 2 convolution blocks and a single skip connection skipping both convolutions. Each convolution block is composed of a convolution layer (kernel size 4), followed by a ReLU, InstanceNorm and dropout (0.1). The dilation rate increases exponentially for each temporal block, from 1 in the first block to 8 in the

Variable-wise	# filters	# parameters	Recall	Precision	AuROC	AuPRC
No (Raw-TCN [5])	15	7336	0.948 (±0.007)	0.527 (±0.021)	0.99 (±0.001)	0.859 (±0.01)
No	30	27751	0.935 (±0.004)	0.611 (±0.02)	0.991 (±0.0)	**0.878 (±0.01)**
Yes (ours)	4×15	7621	0.944 (±0.015)	0.572 (±0.018)	0.988 (±0.003)	0.812 (±0.026)
	8×15	28681	**0.954 (±0.005)**	**0.654 (±0.009)**	**0.993 (±0.0)**	0.872 (±0.011)

Table 1. Classification-related metrics for models with and without variable-wise convolutions for varying number of filters used, with respective number of parameters. For variable-wise models, the number of filters is given per variable (with $V = 15$).

last block. All models were implemented using PyTorch and trained using Adam optimizer with a default learning rate of 10^{-3}. The batch size was set to 256 and a null baseline was used for Integrated Gradients.

For our baseline models with normal convolutions, we used as many convolution filters as the number of variables, similarly to the Raw-TCN architecture [5] ($C = V = 15$), and tested a larger version with twice as many filters ($C = 30$). For variable-wise models, filters count depends on the number of variables and was chosen to reach a number of parameters similar to their non-variable-wise counterparts, so $C = 4 \times 15$ and $C = 8 \times 15$ for smaller and larger model sizes respectively. Classification results presented in Sec.4.1 are provided for the 4 variations, while interpretability results in Sec.4.2 are only given for the largest models (8×15 and 30 filters, respectively, for models with and without variable-wise convolutions).

4. RESULTS

4.1. Classification performance

Table 1 compares models with and without variable-wise convolutions. Notably, separating variable convolutions does not decrease classification performance, with both models achieving comparable classification performance at equivalent parameters count.

4.2. Perturbation-correlation interpretation

In terms of faithfulness, our proposed modification substantially improves the correlation between the attribution of a variable, at the intermediate features level, and the corresponding drop in prediction output when this variable is perturbed (Fig.2). While both models with and without variable-wise convolutions do not exceed a correlation coefficient of 0.25 for attributions computed on input features, variable-wise models attain a coefficient over 0.65 for intermediate features attributions. In contrast, correlation coefficients computed on intermediate features attributions for models without variable-wise convolutions are close to 0. The attribution method, Input*Gradient or Integrated Gradients, has a marginal effect on the correlation on both input and intermediate features level.

While correlations are slightly improved by the variable-wise convolutions on input features level with the *Max* aggregation method, the results are reversed for the *Sum* aggregation, where variable-wise models have slightly lower correlations than non variable-wise models.

5. DISCUSSION AND CONCLUSION

In this paper, we introduced a novel approach to enhance the interpretability of deep learning models for sepsis prediction using variable-wise convolutions. Our method substantially improves the correlation between model behavior and attributed variables at the intermediate features level, while maintaining classification performance. This opens the way for more faithful and granular interpretations of model decisions in critical care settings.

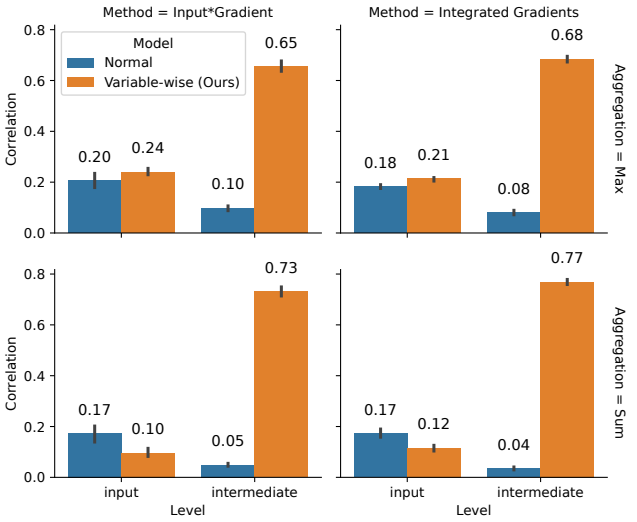

Fig. 2. Perturbation correlation metric comparisons between models with and without variable-wise convolutions, at input and intermediate features level, for both *Sum* and *Max* aggregation methods

The notable increase in correlation coefficients from input to intermediate level for variable-wise models demonstrates the relevance of our approach in preserving variable-specific information at intermediate features level. Even though the time dimension is lost, the successive convolution layers enable richer representation usable for the prediction, easing the attribution maps generation. However, the lower or higher correlations observed for input feature attributions in variable-wise models, depending of the aggregation method used, warrant further investigation of these aggregations, their theoretical implications and the end-user understanding.

The intermediate features and their direct relation to the input variables allow for a better interpretability of the models through attribution methods, but these intermediate features are not directly interpretable as such. Further work should also improve the readability of such features, in particular for interactions with clinicians. In any case, our results are promising, and should be confirmed on a broader range of datasets, potentially for different clinical applications.

Compliance with ethical standards This research study was conducted retrospectively using human subject data made available in open access[1]. Ethical approval was not required

[1] https://physionet.org/content/mimiciv/3.0

as confirmed by the license attached with the open access data. Confidentiality and safety of MIMIC-IV data are ensured by the recommendations of the French "Commission Nationale de l'Informatique et des Libertés" (CNIL), with the Reference Methodology MR-004 of the CNIL.

Acknowledgments This work was supported by the ANRT (Association nationale de la recherche et de la technologie) and PREVIA MEDICAL through a CIFRE fellowship granted to P.-E. Thiboud. It was performed within the framework of the LABEX PRIMES (ANR-11-LABX-0063) of Université de Lyon, within the program "Investissements d'Avenir" operated by the French National Research Agency (ANR) and using HPC resources from GENCI-IDRIS (Grant 2023-AD011014795).

6. REFERENCES

[1] M Singer, CS Deutschman, CW Seymour, et al., "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *JAMA*, vol. 315, 2016.

[2] VX Liu, V Fielding-Singh, JD Greene, et al., "The timing of early antibiotics and hospital mortality in sepsis," *Am J Respir Crit Care Med*, vol. 196, pp. 856–63, 2017.

[3] J Futoma, S Hariharan, and K Heller, "Learning to Detect Sepsis with a Multitask Gaussian Process RNN Classifier," *Proc. ICML*, vol. 70, 2017.

[4] C Lea, MD Flynn, R Vidal, et al., "Temporal Convolutional Networks for Action Segmentation and Detection," *Proc. CVPR*, 2017.

[5] M Moor, M Horn, B Rieck, et al., "Early Recognition of Sepsis with Gaussian Process Temporal Convolutional Networks and Dynamic Time Warping," *Proc. MLR*, vol. 106, 2020.

[6] KE Henry, R Adams, C Parent, et al., "Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing," *Nat Med*, vol. 28, 2022.

[7] A Shrikumar, P Greenside, A Shcherbina, et al., "Not Just a Black Box: Learning Important Features Through Propagating Activation Differences," *Proc. ICML*, vol. 70, 2017.

[8] M Sundararajan, A Taly, and Q Yan, "Axiomatic Attribution for Deep Networks," *Proc. ICML*, vol. 70, 2017.

[9] E Choi, MT Bahadori, JA Kulas, et al., "RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism," *Proc. NeurIPS*, 2016.

[10] M Rosnati and V Fortuin, "MGP-AttTCN: An interpretable machine learning model for the prediction of sepsis," *PLOS One*, vol. 16, 2021.

[11] A van den Oord, S Dieleman, H Zen, et al., "WaveNet: A Generative Model for Raw Audio," *Speech Synthesis Workshop*, 2016.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 84–90, 2017.

[13] AEW Johnson, L Bulgarelli, L Shen, et al., "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, 2023.

[14] A Johnson, L Bulgarelli, T Pollard, et al., "MIMIC-IV-ED," *PhysioNet*, 2023.

[15] Chih-Kuan Yeh, Cheng-Yu Hsieh, and Arun Sai et al. Suggala, "On the (In)fidelity and Sensitivity for Explanations," *Proc. NeurIPS*, 2019.