Self-explaining Hierarchical Model for Intraoperative Time Series

Dingwen Li*, Bing Xue*, Christopher King†, Bradley Fritz†, Michael Avidan†, Joanna Abraham†, Chenyang Lu*
*McKelvey School of Engineering, Washington University in St. Louis
†School of Medicine, Washington University in St. Louis
{dingwenli, xuebing, christopherking, bafritz, avidanm, joannaa, lu}@wustl.edu

***Abstract*—Major postoperative complications are devastating to surgical patients. Some of these complications are potentially preventable via early predictions based on intraoperative data. However, intraoperative data comprise long and fine-grained multivariate time series, prohibiting the effective learning of accurate models. The large gaps associated with clinical events and protocols are usually ignored. Moreover, deep models generally lack transparency. Nevertheless, the interpretability is crucial to assist clinicians in planning for and delivering postoperative care and timely interventions. Towards this end, we propose a hierarchical model combining the strength of both attention and recurrent models for intraoperative time series. We further develop an explanation module for the hierarchical model to interpret the predictions by providing contributions of intraoperative data in a fine-grained manner. Experiments on a large dataset of 111,888 surgeries with multiple outcomes and an external high-resolution ICU dataset show that our model can achieve strong predictive performance (i.e., high accuracy) and offer robust interpretations (i.e., high transparency) for predicted outcomes based on intraoperative time series.**

I. Introduction

Major postoperative complications are devastating to surgical patients with increased mortality risk, need for care, length of postoperative hospital stay and costs of care [1]. With electronic intraoperative data and machine learning, some of these complications are potentially preventable via early predictions [2]. Intraoperative data comprise long and fine-grained multivariate time series, such as vital signs and medications. Furthermore, there are large gaps consisting of consecutive missing values, as shown in 1. These gaps are often associated with surgical procedures or clinical events that require different variables to be monitored at different stages of the surgery.

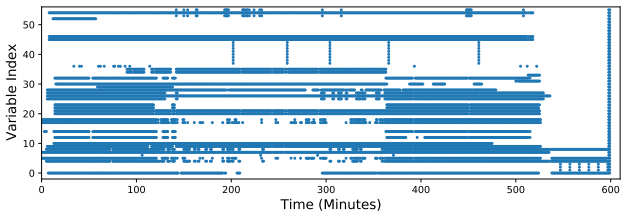

Fig. 1: An example of long intraoperative time series with large gaps. Blue dots represent measurements collected from a surgical case.

Corresponding author: Chenyang Lu (lu@wustl.edu).

It is challenging to learn representations from long time series. Recurrent neural networks (RNNs) have been widely used for learning dynamics from sequential inputs. A common approach to handle long input sequences is to add convolutional layers before recurrent layers [3], [4], but it increases the complexity leading to vanishing gradients. While the attention approach can capture data patterns by skipping recurrent connections and avoid the vanishing gradient issue [3], [5]–[7], pure attention models cannot exploit long-term progression patterns of intraoperative time series, which are informative given the physiological changes during a surgery.

The second challenge imposed by intraoperative time series is associated with the large data gaps commonly observed in intraoperative time series. While imputation has been investigated extensively to estimate missing values, they cannot preserve the information carried by data gaps. The information may be exploited by predictive models given their potential association with surgical procedures and clinical events.

Finally, the interpretability of machine learning models, as explaining which and how input variables contribute to the predictive outcomes, is crucial to clinicians. A good explanation helps clinicians understand the risk factors, thus knowing how to plan for and deliver postoperative care and timely interventions. Despite the invention of model-agnostic explanation methods [8], [9], attribution methods tailored for deep models [10], [11] and self-explaining models [3], [12], it remains challenging to generate accurate explanations identifying important data segments in fine-grained time series.

In this paper, we propose a novel **S**elf-**E**xplaining **H**ierarchical **M**odel (**SEHM**) to learn representations from long multivariate time series with large gaps and generate accurate explanations pinpointing the clinically important data points. The hierarchical model comprises a kernelized local attention and a recurrent layer, which effectively 1) captures local patterns while reducing the size of intermediate representations via the attention and 2) learns long-term progression dynamics via the recurrent module. To make the model end-to-end interpretable, we design a linear approximating network parallel to the recurrent module that models the behavior of a recurrent module locally.

We evaluate SEHM on a perioperative dataset from Barnes Jewish Hospital to predict three postoperative complications. For generality we also apply SEHM to the public High time Resolution ICU Dataset (HiRID) [13] to predict circu-

latory failure. SEHM outperforms state-of-the-art models in predictive performance while achieving higher computational efficiency. We evaluate the model interpretability through both quantitative evaluation and clinician reviews of exemplar surgical cases. Results suggest the advantage of SEHM over existing model interpretation approaches in identifying data points in the input time series with potential clinical importance.

II. Related Work

Handling Long Sequential Data. Traditional RNN models are ineffective when dealing with long sequential data due to the vanishing gradient issue and the computation cost. Temporal convolutional network (TCN) outperforms RNNs in various problems, particularly when input sequences are long [14]. However, TCN models rely on deep hierarchy to ensure the causal convolutions, which incurs significant computation cost for inference. Efficient attention models have been proposed for learning representations from long sequential data, which mainly focus on replacing the quadratic dot-product attention calculation with more efficient operations [6], [7]. SEHM's kernelized local attention captures important local patterns and reduces the size of intermediate representation, while the higher-level RNN model learns long-term dynamics.

Missing Values. Missing values are prevalent in clinical data, which provide both challenges and information for predicting clinical outcomes. RNN-based imputation models demonstrate better performance when learning on sequential data with missing values [15], [16]. However, the recurrent nature of these models makes it difficult to perform imputation and predictions on long sequences. Alternatively, models like MGP-RNN [17] and Latent ODE [18] accommodate the irregularity by creating evenly-sampled latent values. However, these models are computationally prohibitive for long sequences as they either operate with a very large covariance matrix or forward intermediate values to an ODE solver numerous times. The aforementioned approaches increase the uncertainty of estimation for consecutive missing values and ignore the information from data gaps. We overcome this by taking advantage of kernelized local attention and using 0s to encode missing values.

Interpretability. Model-agnostic explanation methods approximate the behavior of deep models while treating them as black boxes [8], [9]. There are also feature attribution approaches designed for interpreting neural networks [10], [11]. Attention models are self-explaining, which allow predictions to be interpreted using the attention matrices directly [3]. Self-Explaining Neural Networks [12] have relevance parametrizers for interpretability, which can be optimized jointly with the classification objective. However, these self-explaining models are not interpretable *end-to-end*. The explanations are generated for concept bases [12] or intermediate representations [3], instead of raw inputs. In contrast, our SEHM is specifically designed to provide end-to-end interpretability by generating decomposed data contribution matrices associated with raw inputs in a linear way.

III. Self-Explaining Hierarchical Model

SEHM comprises three key components: 1) *kernelized local attention* that captures important local patterns and preserves information about data gaps; 2) a *recurrent layer* that learns long-term dynamics; 3) a *linear approximating network* for interpreting the recurrent layer locally, as shown in Figure 2.

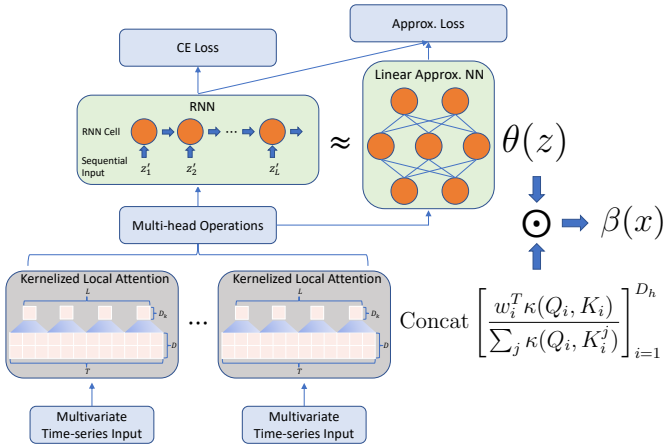

Fig. 2: The overview of SEHM

A. Kernelized Local Attention

Assume we have a two-dimensional multivariate time-series input $x \in \mathbb{R}^{T \times D}$, reshaped as $\tilde{x} \in \mathbb{R}^{L \times C \times D}$, such that $T = L \times C$. This essentially enforces attention weights attending to neighboring input points with a length of C and outputs L weighted sums. Self-attention allows each time step to interact with all its neighbors reducing the information decay. The attention matrix is formulated as a positive-definite kernel $\kappa(q_i, k_j)$, such that q_i and k_j are the i-th vector in the query and j-th vector in key calculated from $\tilde{x}$. We define the kernelized attention as an expectation over an inner product of a randomized feature map $\phi : \mathbb{R}^D \to \mathbb{R}_+^R$ as $R > 0$:

$$\kappa(q_i, k_j) = \mathbb{E}_{\omega \sim \mathcal{D}}[\phi(q_i)^T \phi(k_j)] \tag{1}$$

where $\mathcal{D}$ is a distribution from which ω is sampled i.i.d. Thus the attention can be formulated as a weighted sum over the latent dimension (usually the temporal dimension):

$$a_i = \sum_{j=1}^{C} \frac{\kappa(q_i, k_j)}{\sum_{j'=1}^{C} \kappa(q_i, k_{j'})} v_j = \frac{\mathbb{E}[\phi(q_i)^T \sum_{j=1}^{C} \phi(k_j) v_j]}{\mathbb{E}[\phi(q_i)^T \sum_{j'=1}^{C} \phi(k_{j'})]}. \tag{2}$$

After reordering products and reusing $\sum_{j=1}^{C} \phi(k_j) v_j$ and $\sum_{j'=1}^{C} \phi(k_{j'})$ for each i, the time and memory complexity can be reduced to $O(C)$ [7]. The kernel function in Eq.(2) unbiasedly approximates the exponential of the dot product in softmax attention by drawing feature vectors from a zero-mean Gaussian distribution $\omega \sim \mathcal{N}(0, I_D)$

$$\begin{aligned} &\exp(q_i^T k_j) = \mathbb{E}_{\omega \sim \mathcal{N}(0, I_D)}[\phi(q_i)^T \phi(k_j)], \\ &\text{s.t. } \phi(z) = \exp(\omega^T z - \frac{\|z\|^2}{2}), \ z = q_i \text{ or } k_j. \end{aligned} \tag{3}$$

When constructing random feature samples ω to be exact orthogonal, the softmax attention can be accurately approximated by having exponentially small and sharper bounds on regions where the attention values after softmax are small [7].

The attention output $A = \{a_i\}_{i=1}^{C}$ further shrinks to aggregate the learned information among neighbors, such that $w^T A$, where $w \in \mathbb{R}^C$ is a learnable parameterized vector. Then we have the multi-head version of above attention:

$$H = \text{Concat}(w_1^T A_1, ...w_h^T A_h, ...w_H^T A_H)W^O \tag{4}$$

where each $A_h^T w_h$ denotes the attention output of head h, $W^O \in \mathbb{R}^{HD \times D_o}$. The learned compact representation H will be used as the input to the recurrent layer.

In contrast to imputation or generative approaches, we propose to directly use original multivariate time series $\tilde{x}$ as the "value" component v in Eq.(2) and encode missing values as zeros. Zero encoding along with the special structure of the localized attention can effectively utilize the information conveyed by missing values at no additional computation costs.

Proposition 1 *Zero-encoding enables the kernelized local attention to output 0 for the measurement gaps $C_g \geq 2C-1$, where C is the size of neighborhood.*

Assume there is a gap in one input variable $\tilde{x} \in \mathbb{R}^{L \times C}$ that has a length $C_g \geq 2C-1$. Since $C_g \geq 2C-1$, there is always at least one row in $\tilde{x}$ that contains all zeros. Without loss of generality, we assume the l-th row has all zeros, denoted as $\tilde{\boldsymbol{x}}_l = \mathbf{0}$. Hence, for the attention output corresponds to the l-th row, we can easily verify that it is equal to 0 by

$$a_l = \sum_{j=1}^{C} \frac{\sum_{i=1}^{C} w_i \kappa(q_{li}, k_{lj})}{\sum_{j'=1}^{C} \kappa(q_{li}, k_{lj'})} \tilde{x}_{lj} = 0 \tag{5}$$

where a_l is the attention output corresponding to the l-th row vector. This design enables the attention to capture gaps that are large than $2C-1$ and preserve the information of gap in the attention output.

B. Self-explaining Model with Linear Approximation for RNN

In order to achieve end-to-end interpretability, a self-explaining linear approximation is introduced in parallel with the recurrent layer. Assume the intermediate input to the recurrent layer is denoted as $z = h(x)$, where $h(x)$ is the output of kernelized local attention. For simplicity, in the following context we use $g(z)$ to represent the explanation model, such that

$$f(z) \approx g(z) = \theta(z)^T z \tag{6}$$

where $\theta(z)$ denotes the parameters of explanation model to be learned. The goal is to locally approximate the actual probabilistic output $f(z)$ of RNN by $g(z)$. Other than the accurate approximation, we also want the explanation to be robust against local perturbation. If $g(z)$ is differentiable at z, the gradient of $g(z)$ can be decomposed as

$$\nabla_z g(z) = \theta(z)J + \nabla_z \theta(z) z \tag{7}$$

where J is an all-one matrix. In order to make $g(z)$ locally behave like a linear function and be close to the probabilistic output $f(z)$, $\theta(z)J$ should approximate $\nabla_z f(z)$, and $\nabla_z \theta(z)$ should approach $\mathbf{0}$. With these goals, we propose a loss $\mathcal{L}_\theta$ to ensure the local linearity as well as stability

$$\mathcal{L}_\theta = \|\theta(z)J - \nabla_z f(z)\|_2 + \lambda \|\nabla_z \theta(z)\|_1 \tag{8}$$

where λ is a balancing coefficient. However, this loss is hard to optimize in practice, since $\nabla_z \theta(z)$ has to be calculated in the loss function. We derive an upper bound of $\mathcal{L}_\theta$, which is a surrogate loss that can be calculated efficiently with the same goal of achieving local linearity and approximating accuracy.

Proposition 2 *For any Multi-Layer Perceptron (MLP) implementing $\theta(z)$ with 1-Lipschitz activation functions (e.g., ReLU, Leaky ReLU, SoftPlus, Tanh, Sigmoid, ArcTan or Softsign) [19], the upper bound of Eq. (8) is*

$$\hat{\mathcal{L}}_\theta = \|\theta(z)J - \nabla_z f(z)\|_2 + \lambda \sqrt{d} \prod_{k=1}^{K} \|W_k\|_2 \tag{9}$$

where W_k is the parameter of the k-th layer in the MLP implementing $\theta(z)$, d is the dimension of $\nabla_z \theta(z)$. From L1-L2 norm inequality, we have

$$\|\nabla_z \theta(z)\|_1 \leq \sqrt{d} \|\nabla_z \theta(z)\|_2. \tag{10}$$

Without loss of generality, we assume that the linear approximation parameter $\theta(z)$ is realized by a nested Multi-Layer Perceptron (MLP) with $\theta_k = a_k(g_k(\theta_{k-1}))$, where g_k is the k-th layer perceptron, a_k is the k-th 1-Lipschitz activation function, θ_{k-1} is the output from the last preceding layer. The k-th layer perceptron takes an affine transformation on the input data, such that $g_k(\theta_{k-1}) = W_k \theta_{k-1} + b_k$. The chain rule implies that the gradient of $\theta(z) = \theta_K$ can be derived as

$$\nabla_z \theta(z) = a'_K g'_K \nabla \theta_{K-1}. \tag{11}$$

where a'_k and g'_k represent the Jacobian matrices, K denotes the last layer in MLP. Take the 2-norm of both sides, then we get

$$\|\nabla_z \theta(z)\|_2 = \|a'_K g'_K \nabla \theta_{K-1}\|_2 \leq \tag{12}$$

$$\leq \|a'_K\|_2 \|g'_K \nabla \theta_{K-1}\|_2 \leq \|a'_K\|_2 \|g'_K\|_2 \|\nabla \theta_{K-1}\|_2 \tag{13}$$

where the norm for matrices is the induced 2-norm, $\|\nabla \theta_{K-1}\|_2$ can be further expanded via chain rule until reaching the input z. Since $\{a_k\}_{k=1}^{K}$ functions are all 1-Lipschitz activation functions, it implies that $\|a'_k\|_2 \leq 1$. Each layer in MLP is an affine transformation, which yields the magnitude of g'_k to be $\|W_k\|_2$. Thus, we have

$$\|\nabla_z \theta(z)\|_2 \leq \prod_{k=1}^{K} \|W_k\|_2 \tag{14}$$

assuming g_k is an affine function and a_k is a 1-Lipschitz activation function. With Eq.(10) and Eq.(14), we obtain an upper bound $\hat{\mathcal{L}}_\theta$ of the original loss $\mathcal{L}_\theta$, such that

$$\mathcal{L}_\theta \leq \|\theta(z)J - \nabla_z f(z)\|_2 + \lambda \sqrt{d} \prod_{k=1}^{K} \|W_k\|_2 = \hat{\mathcal{L}}_\theta \tag{15}$$

We randomly sample instances around z uniformly within a small distance. Thus, we obtain a perturbed set of $z' \in \mathbb{Z}$, which is used for approximating $f(z)$ locally.

IV. Experimental Evaluation

The experiments were conducted on a large dataset collected from 111,888 operations performed on adults at Barnes Jewish Hospital. We evaluated predictive performance for three types of complication including delirium, pneumonia and acute kidney injury (AKI). These complications were identified to be essential for postoperative care based on a recent stakeholder-based study with clinicians. We also performed an external evaluation on HiRID [13] with high-resolution data from 36,098 admissions to validate the generality on modeling other high-resolution clinical time series.

A. Dataset and Preprocessing

1) Postoperative Complication Prediction: There were 56 time-series variables in total with a maximum sampling rate of every minute. We included all observations from 600 minutes prior to the end of surgery. Missing values were handled by either built-in imputation method or zero-encoding according to different models. The label was defined as the onset of a particular postoperative complication. After preprocessing, we obtained three datasets for evaluating the model's performance on predicting delirium, pneumonia and AKI respectively.

2) Circulatory Failure Prediction: In HiRID, clinical time series were recorded at a frequency of one measurement every 2 minutes. The task is to predict circulatory failure 8 hours prior to the first occurrence [1]. We excluded admissions that were shorter than 8 hours, resulting in 134,362 samples. 37 time-series variables with the overall availability $>1\%$ were selected. For each admission with circulatory failure, we extracted all data of these 37 variables from 16 hours to 8 hours prior to the first occurrence of circulatory failure, yielding a positive sample with a maximum of 480-minute data. For the time period from the start of the admission to the 16-th hour prior to the first occurrence, we segmented it into multiple 8-hour consecutive chunks. We applied a sliding window with a stride of 8 hours to extract data from each chunk, yielding negative samples. For each admission without circulatory failure, we applied the same procedure as described above to extract negative samples, except the window slides along the whole admission. Same techniques were conducted for missing values.

B. Evaluation Setting

The datasets were split as 75% of the samples were used for training and the rest 25% were used for testing. Within the training set, we further designated 10% of them to form a validation set for hyperparameter tuning. For all models used in the evaluation, we tuned the batch size from a set of choices, such as 16, 32, 64, 128, 256. We also tuned the learning rate of Adam optimizer from 0.0001 to 0.01. For other hyperparameters specific to each model, we applied Bayesian optimization to select an optimal set of hyperparameters based on the validation set. Each predictive accuracy evaluation was run repeatedly for 10 times. The code[2] is available.

[1]We used the same definition of circulatory failure that was originally proposed by [13]

C. Predictive Performance Benchmark

In this experiment, we evaluate the predictive performance of SEHM in comparison to a set of existing models including state-of-the-art models designed for long multivariate time series. We use the area under the receiver operating characteristic curve (AUROC) and the area under the precision recall curve (AUPRC) as performance metrics.

The models included in our performance evaluation can be classified into three categories. The first category includes RNN variants: **GRU**, **BRITS*** [16][3], **GRU-D*** [15], **Latent-ODE** [18]. The second category includes Transformer-type attention models for handling sequential data: **SAnD** [5], **Informer** [6], **Performer** [7]. The third category includes existing deep hierarchical models and our proposed SEHM: **Conv GRU** [4], **Multi-scale CNN** [20], **TCN** [14], **RAIM** [3], **SEHM(GRU)**. The results of the predictive performance evaluation are shown in Table I. We have following observations from Table I. (1) SEHM outperforms vanilla RNN (GRU), which shows the advantage of using kernelized local attention to capture important local patterns and shorten the inputs to latter RNN models. (2) SEHM yields better results than BRITS, GRU-D and Latent-ODE, which suggests using zero encoding to represent missing values is beneficial. (3) SEHM demonstrates better performance than pure attention models, which indicates the necessity of incorporating RNN models for learning long-term dynamics. (4) When comparing SEHM with other hierarchical models using convolution as the first layer, we observe the locality of attention is a better way of learning local patterns than convolution for intraoperative time series. (5) The consistent results across different prediction tasks suggest the approach is generalizable for predicting different postoperative complications. All results from the comparison are statistically significant ($p < 0.05$).

D. Ablation Study

We performed ablation study on the delirium prediction to evaluate the effect of locality, zero-encoding and kernelization in terms of predictive performance and model inference speed. The inference speed is measured as the average time (in milliseconds) of completing a forward inference with a batch size of 64 samples. We have following observations from Table II. (1) Locality reduces inference time significantly while yielding better predictive results. This is because the temporal size of input to latter RNN model is reduced and local attention exploits useful information from temporal neighbors. (2) The introduction of zero encoding improves the predictive performance without additional computation overhead. (3)

[2]https://github.com/WU-CPSL/sehm

[3]The inputs to BRITS and GRU-D are down-sampled by a factor of 20. The original raw inputs cause slow training and the performance is sub-optimal when compared to models trained with down-sampled inputs

TABLE I: Predictive performance $mean(\sigma)$ reported for different complication prediction tasks.

	Delirium		Pneumonia		AKI		HIRID	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
GRU	0.7182(0.0013)	0.7369(0.0025)	0.8409(0.0017)	0.1155(0.0033)	0.7560(0.0021)	0.1921(0.0013)	0.8611(0.0035)	0.4156(0.0222)
BRITS*	0.7438(0.0010)	0.7684(0.0016)	0.8509(0.0018)	0.1384(0.0023)	0.7815(0.0006)	0.2182(0.0023)	0.9181(0.0015)	0.5354(0.0080)
GRU-D*	0.7386(0.0018)	0.7605(0.0020)	0.8510(0.0016)	0.1349(0.0038)	0.7698(0.0026)	0.2100(0.0012)	0.9068(0.0103)	0.5143(0.0077)
Latent-ODE	0.7294(0.0021)	0.7551(0.0019)	0.8406(0.0038)	0.1314(0.0050)	0.7663(0.0049)	0.2068(0.0032)	0.8876(0.0134)	0.5009(0.0065)
SAnD	0.7274(0.0042)	0.7575(0.0052)	0.8215(0.0053)	0.1121(0.0032)	0.7565(0.0056)	0.1938(0.0073)	0.8963(0.0074)	0.4539(0.0053)
Informer	0.7351(0.0009)	0.7627(0.0024)	0.8347(0.0034)	0.1206(0.0049)	0.7597(0.0016)	0.1955(0.0006)	0.9078(0.0086)	0.5220(0.0055)
Performer	0.7301(0.0033)	0.7581(0.0025)	0.8383(0.0056)	0.1192(0.0044)	0.7532(0.0015)	0.1888(0.0049)	0.9043(0.0044)	0.5178(0.0057)
Conv GRU	0.7369(0.0015)	0.7586(0.0014)	0.8503(0.0008)	0.1388(0.0015)	0.7763(0.0005)	0.2080(0.0014)	0.9150(0.0021)	0.5319(0.0023)
Multi-scale CNN	0.7397(0.0013)	0.7652(0.0010)	0.8504(0.0016)	0.1411(0.0018)	0.7769(0.0019)	0.2123(0.0031)	0.8952(0.0035)	0.4973(0.0065)
TCN	0.7369(0.0013)	0.7552(0.0019)	0.8401(0.0023)	0.1148(0.0064)	0.7444(0.0010)	0.1915(0.0015)	0.8908(0.0117)	0.4917(0.0022)
RAIM	0.7228(0.0038)	0.7509(0.0039)	0.8423(0.0005)	0.1314(0.0009)	0.7644(0.0008)	0.2045(0.0028)	0.9039(0.0076)	0.5034(0.0064)
SEHM(GRU)	**0.7571(0.0015)**	**0.7795(0.0011)**	**0.8610(0.0009)**	**0.1505(0.0026)**	**0.8116(0.0024)**	**0.2378(0.0033)**	**0.9265(0.0012)**	**0.5628(0.0020)**

TABLE II: Ablation study $mean(\sigma)$

Locality	Zero encoding	Kernel-ization	AUROC	AUPRC	Time (ms)
			0.7233(0.0034)	0.7450(0.0019)	37.5(3.1)
	✓		0.7342(0.0033)	0.7542(0.0012)	37.1(2.8)
✓			0.7342(0.0025)	0.7615(0.0029)	17.9(2.5)
		✓	0.7214(0.0018)	0.7435(0.0033)	32.4(3.7)
✓	✓		0.7565(0.0017)	0.7789(0.0030)	18.5(3.1)
✓		✓	0.7398(0.0014)	0.7651(0.0007)	12.1(1.8)
	✓	✓	0.7330(0.0023)	0.7547(0.0028)	33.9(2.3)
✓	✓	✓	0.7571(0.0015)	0.7795(0.0011)	11.8(2.2)

The kernelization further increases the model inference speed and achieves comparable predictive performance as original softmax function.

E. Evaluations on Interpretability

The model explanation methods used in the evaluations are: (1) model-agnostic explanation methods, e.g., **LIME** [8], **Kernel/Deep SHAP** [9]; (2) feature attribution methods, e.g., **Integrated Gradient** [10], **DeepLift** [11]; (2) a self-explaining deep model, e.g., **RAIM** [3][4].

TABLE III: Quantitative evaluation of model explanation approaches $mean(\sigma)$

	Local Accuracy ↓ (MSE)	Faithfulness ↑ (AOPC@2k)	Stability ↓ (est. Lipschitz)
LIME	0.2957(0.0374)	0.1934(0.0005)	12.3944(0.0114)
KernelSHAP	0.3241(0.0215)	0.1141(0.0035)	10.1523(0.3258)
DeepSHAP	0.3837(0.0084)	0.1118(0.0112)	8.7104(0.2246)
Int. Grad.	0.3178(0.0145)	0.2749(0.0041)	5.7964(0.1762)
DeepLift	0.2648(0.0027)	0.3321(0.0060)	8.9637(0.2714)
RAIM	–	0.1513(0.0025)	5.3167(0.2988)
SEHM	**0.2327(0.0118)**	**0.5583(0.0057)**	**3.5498(0.0811)**

1) Quantitative Evaluation: We propose three evaluation metrics for comparing the explanations generated by different approaches. *Local accuracy* is defined as the mean square error between the aggregated explanations generated by the model explanation approach and the probabilistic outputs of the original predictive model[5]. *Faithfulness* is achieved by evaluating the area over the most relevant first perturbation curve (AOPC), which assesses the ability of model assigning high values to those input variables that have high influence to final predictive outcomes. We report the AOPC of top 2,000 input data points ranked by model explanation methods. *Stability* is evaluated by the estimated Lipschitz continuity [12], which reflects the extent of changes in explanation when applying small perturbation to the input that does not change the predictive outcome. Smaller estimated Lipschitz continuity means more stable explanation.

We observe that SEHM significantly outperforms other baselines in the three quantitative evaluations ($p < 0.05$), as detailed in Table III. Since SEHM utilizes approximation to model the behavior of RNN, better local accuracy of SEHM can be interpreted as more accurate approximation. The evaluation on the faithfulness confirms that SEHM is better at identifying important data points in the intraoperative time series by ranking the most relevant data points correctly. This is a very promising result, since SEHM is able to provide clinicians with more faithful explanations and avoid wrong explanations that may trigger false alarms.

2) Case Studies: We visualize the explanations provided by SEHM, DeepLift, and KernelSHAP in a surgical case and have an anesthesiologist specializing in perioperative care review the explanations from a clinical perspective. DeepLift and KernelSHAP are chosen as they are representatives of feature attribution methods and model-agnostic explanation methods, respectively. The self-explaining models are not applied to the case studies as they cannot provide end-to-end explanations that attribute contributions to original clinical data. The visualizations shown in Figure 3 are generated from 100 consecutive minutes till the end of surgery. We select 3 intraoperative variables that are commonly available during the operation and intuitive to the readers. The selected variables include heart rate (HR), respiratory rate (RR) and non-intrusive blood pressure (sBP non).

For the surgical case shown in Figure 3, SEHM marks the duration around the 20-th minute as highly important. Based on medical records, medications affecting blood pressure and heart rate were administrated to the patient at that time. However, both DeepLift and KernelSHAP miss the critical

[4]Additional evaluations are available at https://arxiv.org/abs/2210.04417

[5]There is no local accuracy evaluation for RAIM, since it is not an additive feature attribution method.

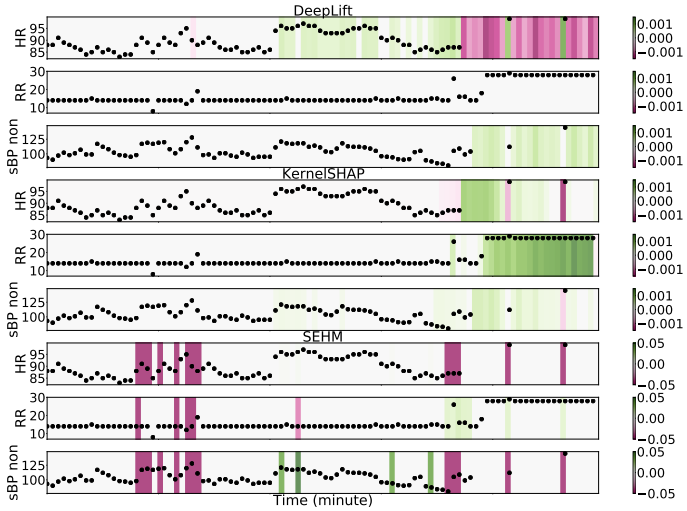

Fig. 3: Visualization of explanations generated for data points in the last 100 minutes of a surgery.

time period associated with a medication event. In addition, SEHM identifies a number of high values in the measurements with potential clinical significance. KernelSHAP attributes importance to the sequence of high RR values at the end of the case, even though the measurements are likely artifacts caused by an instrument issue.

Notably, both DeepLift and KernelSHAP assign high contributions to the end of surgery when few measurements were collected. The end-of-case sparse data issue is difficult for baseline methods to interpret because they tend to focus on missing data points, but missingness between observations is completely normal in that context. In contrast, SEHM avoids assigning importance to the end of surgery. This may be attributed to the design of SEHM that utilizes parameters learnt from global patterns during the training.

V. Conclusion

This paper presents SEHM, a self-explaining hierarchical model specifically designed for high-resolution clinical time series. SEHM integrates kernelized local attention and RNN to handle long time series. Furthermore, it provides end-to-end interpretability that identifies input variables and time windows in which the data are highly correlated to the final outcomes. Experiments on real-world datasets demonstrate SEHM's superior performance in predicting important clinical outcomes and generating clinical meaningful explanations.

VI. Acknowledgement

This work was supported, in part, by the Fullgraf Foundation.

References

[1] B. Xue, D. Li, C. Lu, C. R. King, T. Wildes, M. S. Avidan, T. Kannampallil, and J. Abraham, "Use of Machine Learning to Develop and Evaluate Models Using Preoperative and Intraoperative Data to Identify Risks of Postoperative Complications," *JAMA Network Open*, vol. 4, pp. e212240–e212240, 03 2021.

[2] G. B. Weller, J. Lovely, D. W. Larson, B. A. Earnshaw, and M. Huebner, "Leveraging electronic health records for predictive modeling of post-surgical complications," *Statistical Methods in Medical Research*, vol. 27, no. 11, pp. 3271–3285, 2018.

[3] Y. Xu, S. Biswal, S. R. Deshpande, K. O. Maher, and J. Sun, "RAIM: recurrent attentive and intensive model of multimodal patient monitoring data," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pp. 2565–2573, 2018.

[4] Q. Tan, A. J. Ma, M. Ye, B. Yang, H. Deng, V. W.-S. Wong, Y.-K. Tse, T. C.-F. Yip, G. L.-H. Wong, J. Y.-L. Ching, F. K.-L. Chan, and P. C. Yuen, "Ua-crnn: Uncertainty-aware convolutional recurrent neural network for mortality risk prediction," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, p. 109–118, 2019.

[5] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[6] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Conference*, vol. 35, pp. 11106–11115, 2021.

[7] K. M. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Q. Davis, A. Mohiuddin, L. Kaiser, D. B. Belanger, L. J. Colwell, and A. Weller, "Rethinking attention with performers," in *International Conference on Learning Representations*, 2021.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144, 2016.

[9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, pp. 4765–4774, 2017.

[10] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, p. 3319–3328, 2017.

[11] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, p. 3145–3153, 2017.

[12] D. Alvarez-Melis and T. S. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, p. 7786–7795, 2018.

[13] S. L. Hyland, M. Faltys, M. Hüser, X. Lyu, T. Gumbsch, C. Esteban, C. Bock, M. Horn, M. Moor, B. Rieck, M. Zimmermann, D. Bodenham, K. Borgwardt, G. Rätsch, and T. M. Merz, "Early prediction of circulatory failure in the intensive care unit using machine learning," *Nature Medicine*, vol. 26, pp. 364–373, Mar 2020.

[14] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *CoRR*, vol. abs/1803.01271, 2018.

[15] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific Reports*, vol. 8, p. 6085, Apr 2018.

[16] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, and Y. Li, "Brits: Bidirectional recurrent imputation for time series," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[17] J. Futoma, S. Hariharan, and K. Heller, "Learning to detect sepsis with a multitask Gaussian process RNN classifier," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*, pp. 1174–1182, 06–11 Aug 2017.

[18] Y. Rubanova, R. T. Q. Chen, and D. K. Duvenaud, "Latent ordinary differential equations for irregularly-sampled time series," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[19] A. Virmaux and K. Scaman, "Lipschitz regularity of deep neural networks: analysis and efficient estimation," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[20] Z. Cui, W. Chen, and Y. Chen, "Multi-scale convolutional neural networks for time series classification," *CoRR*, vol. abs/1603.06995, 2016.