Learning Interpretable Deep Representation via Attention Mechanism for ICU Clinical Prediction Tasks

Guohang Zeng , **Pauline Lin** , **Uwe Aickelin**

School of Computing and Information Systems,
The University of Melbourne

guohangz@student.unimelb.edu.au, {pauline.lin, uwe.aickelin}@unimelb.edu.au

Abstract

With the wide adoption of Electronic Health Record(EHR) systems in hospitals, a large amount of healthcare data has been stored and made deep learning-based medical data mining becomes possible. Recent studies showed that deep learning-based methods had superior performances on EHR data mining tasks, some of which provide interpretation on its predictions to assist doctors in evaluating the credibility of the black-box model. In this study, we develop an interpretable deep learning architecture named MA-LSTM(Multivariable Attention LSTM) for Intensive Care Unit(ICU) clinical prediction tasks. Our proposed model applies attention mechanism on multivariable time-series data and outperforms the channel-wise LSTM baseline. By evaluating local-Lipschitz constant and comparing with intrinsic feature importance measurement, experiments showed that local explanations provided by our model perform more stable and more faithful than post-hoc interpretability methods with less computation cost.

1 Introduction

For the past decade, there was an increase in the amount of medical data stored in Electronic Health Record(EHR) systems with the wide adoption of EHR systems in the hospital. EHR systems store a wide range of patient information related to healthcare, including diagnoses, laboratory results, patient conditions, proteomics, treatments, and clinical note[Shickel *et al.*, 2017]. Although EHR systems were originally designed for efficient information management for hospitals, the broad adoption of EHR systems has increased the amount of stored medical data as well as the possibility of developing medical data mining models to improve the quality of healthcare. The massive medical data are useful for a broad range of medical informatics applications, including phenotyping[Perros *et al.*, 2018][Zhao and Weng, 2011], medical concept extraction[Choi *et al.*, 2016a], and medication recommendation[Shang *et al.*, 2019].

EHR data usually can be formalized as time series sequences of clinical variables, where the sequences represent the documented content of medical records from each patient. It cares about inference patients' condition $P(Y|X)$ by learning from medical records, where patients condition $P(Y|X)$ usually represents the probability that the patient will have a particular disease(or mortality) in a specific time window. In the early stage of EHR research, traditional approaches rely on domain knowledge to define appropriate medical features for data mining, followed by white-box machine learning methods such as logistic regression or decision tree. The advantage of these methods is that its decision-making process can be trusted by doctors, while the performance of the accuracy of these models was not high enough to reach an acceptable level to clinical practice.

One challenge of EHR data mining is the performance of accuracy. Before medical AI systems can be widely applied in clinical practice, the accuracy of the model needs to be high enough to reach an acceptable level to people. Recently, deep neural networks have shown their excellent performance in computer vision and natural language processing. By training with the large amount the data via end-to-end approaches, deep neural networks learn discriminative and powerful representation for computer vision and natural language processing tasks and outperform traditional methods. With the wide adoption of Electronic Health Record(EHR) systems in hospitals, a large amount of healthcare data has been stored that made deep learning-based medical data mining becomes possible. It is natural to consider leveraging deep learning techniques to improve the performance of healthcare applications. In recent years, there is an increasing number of publications that are applying deep learning techniques to EHR research[Lipton *et al.*, 2015][Choi *et al.*, 2018][Choi *et al.*, 2016b], most of which outperformed traditional machine learning approaches by a large margin and have achieved high performance, especially in disease predictions[Li *et al.*, 2019a][Li *et al.*, 2019b]

Another challenge is interpretability. Although deep learning-based methods can achieve high accuracy, the black-box characteristic is the Achilles' Heel of deep learning, which makes it challenging to be understood and trusted by humans. The interpretability issue of the deep learning model has always been an important factor affecting the widespread adoption in medical practice. In general, performance and interpretability are two essential factors of machine learning models, while deep learning methods sacrifice its interpretability when pursuing accuracy. Since we must be able to

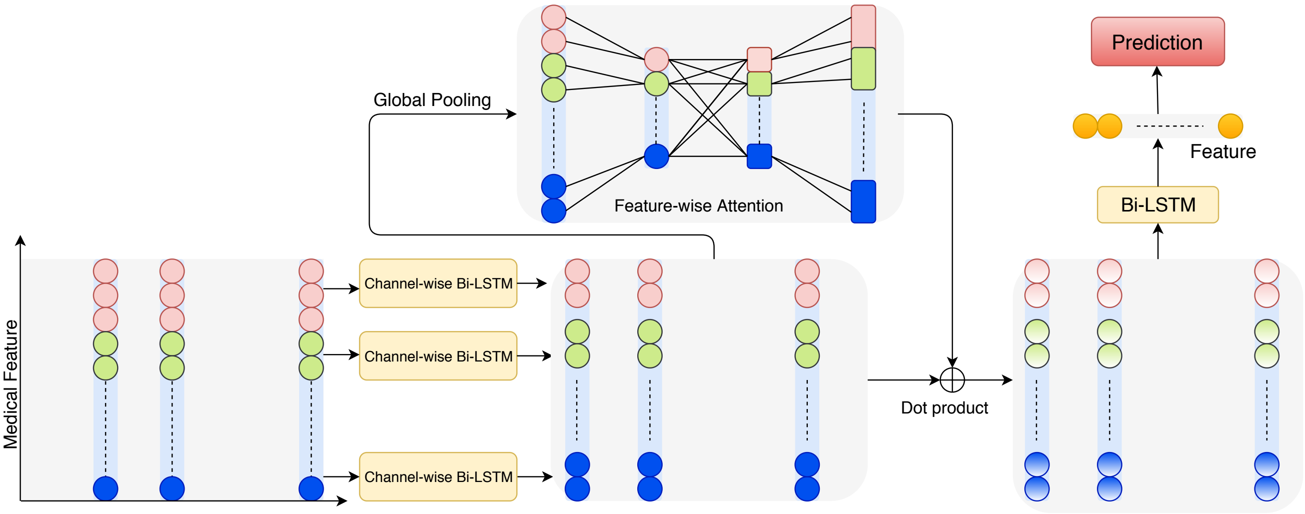

Figure 1: Our proposed MA-LSTM architecture: Attention mechanism can be applied on multivariable data after channel-wise Bi-LSTM was used to decouple the representations for different clinical features in the hidden states.

understand and trust the clinical predictions made by machine learning models, it is essential to establish an interpretable EHR model that can be widely applied and can be trusted by humans. In recent years, many interpretability methods have been proposed to explain the decision-making mechanism in black-box medical models, most of which rely on attention mechanisms to provide local explanations with visualization of the prediction to assist doctors in evaluating the correctness of the model[Kwon *et al.*, 2018][Zhang *et al.*, 2018][Perros *et al.*, 2018].

The other issue which has not been noticed in healthcare literature is the *reliability* of interpretability. Interpretability methods produce visualizations of prediction to assist doctors in evaluating the correctness of the model, while the correctness of interpretation has never been evaluated. Recent studies show that local explanations produced by interpretability methods might be fragile to adversarial attacks as well as being inconsistent with features that are really important to the model.

In this paper, we proposed our MA-LSTM(Multivariable Attention LSTM) architecture for Intensive Care Unit(ICU) clinical prediction tasks from MIMIC-III benchmarks[Harutyunyan *et al.*, 2019]. To the best of our knowledge, we are the first to introduce interpretability evaluation to healthcare research. The main contributions of this paper include:

- (1) We proposed a novel neural network architecture named MA-LSTM that applying attention mechanism on multivariable time-series clinical data. Our model outperforms the multi-channel LSTM baseline.
- (2) Local explanation(feature importance) can be obtained off-the-shelf while inferencing the neural network without additional computation cost.
- (3) Our model provides a more faithful and stable explanation compared with other post-hoc explanation methods.

2 Related Works

Deep learning-based EHR data mining: Instead of relying on defining appropriate clinical features with domain knowledge, deep neural networks automatically learn medical representation from large amounts of training data and outperform traditional approaches in EHR data mining. [Pham *et al.*, 2016] introduced a Long Short-Term Memory (LSTM) architecture with the attention mechanism for predictions of the onset of diabetes. [Zhang *et al.*, 2018] proposed Patient2Vec framework to learn interpretable deep representations of longitudinal EHR data and provides visualizations for its predictions. By building on the latest developments in NLP, [Li *et al.*, 2019b] proposed the BEHRT(BERT for EHR) trained on a dataset from 1.6 million records for disease predictions and show a remarkable improvement compared to other existing state-of-the-art EHR models. All these demonstrated deep neural networks' effectiveness in capturing temporal information for EHR data.

Interpretability Methods: Interpretability methods aim to provide local explanations(feature importance) for predictions produced by black-box models. Interpretability methods can be categorized into *post-hoc explanation methods* produced interpretation on trained models, and *self-explaining methods* learned interpretation within the model. Post-hoc explanations such as Saliency Map takes the magnitude of the partial derivative of the output of the model with respect to the input to estimate feature importance. Self-explaining methods such as attention mechanism learn importance weights for the model's input to improve the model's performance, while the attention weights are usually regarded as indications of feature importance.

Evaluation of Interpretability: Recent studies had shown that local explanation produced by interpretability methods might be fragile to adversarial attacks[Ghorbani *et al.*, 2019] as well as being inconsistent with features that are really important to the model[Jain and Wallace, 2019]. [Ghorbani *et al.*, 2019][Zheng *et al.*, 2019] generated adversarial perturbations that do not change the model outputs but cause very different interpretation that provided by interpretability methods. To this end, [Alvarez-Melis and Jaakkola, 2018a] argued that the stability of explanations is a crucial property of interpretability: an explanation should remain roughly constant in its vicinity, which can be evaluated by local-Lipshitz estimation. As for faithfulness of interpretability, recent works[Kim *et al.*, 2016][Kindermans *et al.*, 2017] have shown that post-hoc interpretation can be misleading. [Jain and Wallace, 2019] evaluated local explanation produced by attention mechanism and claimed that attention weights are frequently uncorrelated with other intrinsic feature importance measures and mean that the attention mechanism does not provide meaningful explanations.

3 Proposed Model :Multivariable Attention LSTM(MA-LSTM)

It's challenging to apply attention mechanisms to multivariable time-series data due to it is not satisfied with the precondition of vanilla attention mechanism. In this paper, we proposed an architecture named MA-LSTM(Multivariable Attention LSTM) that encodes the attention module for multivariable data and develops the interpretation for feature importance. The overall structure of MA-LSTM is illustrated in Figure 1.

3.1 Preliminaries

Long Short Term Memory(LSTM) is the most widely-used variant of Recurrent neural networks(RNNs) and has shown its superior performance on sequential modelling. It is composed of input gates, output gates, cell state vector, and forget gates that proposed to preserve long term dependencies among inputs to learn the representation for sequential input. The LSTM is formulated as:

$$\begin{aligned}
i_t &= \sigma(x_t W^{(xi)} + h_{t-1} W^{(hi)}),\\
f_t &= \sigma(x_t W^{(xf)} + h_{t-1} W^{(hf)}),\\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(x_t W^{(xc)} + h_{t-1} W^{(hc)} + b^{(c)}),\\
o_t &= \sigma(x_t W^{(xo)} + h_{t-1} W^{(ho)} + b^{(o)}),\\
h_t &= o_t \odot \sigma_h(c_t),
\end{aligned} \tag{1}$$

where i_t denotes the input gates, f_t denotes the forget gates, o_t denotes the output gates, $c + t$ denotes the cell state vector and h_t denotes the hidden state for position t. This complex architecture of LSTM was Intuitively designed for remember the importance information and forget unimportant to capture long-term dependencies in sequential data. In the following section, we will use $h_t = LSTM(x_t, h_{t-1})$ for short for the above equations.

Given sequential input $x = [x_1, x_2, ..., x_t]$, LSTM iteratively learns the temporal representation $h_t = LSTM(x_t, h_{t-1})$ which can be used as the feature vector for sequential modeling. For sequential data $x = [x_1, x_2, ..., x_t]$ where each input x_t is represented as word embedding with unique semantics, we can assign learnable attention weights α_t for h_t to imporve the generalization capability of LSTM, where α_t is computed by:

$$\alpha_i = \frac{e^{W_i h_i}}{\sum_{j=1}^{T} e^{W_j h_j}} \tag{2}$$

where W_i are parameters for learning attention weights, given hidden state h from LSTM, attention mechanism compute the importance weight $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_t\}$ for sequential inputs. Intuitively, the effect the attention weights is to put more attention on essential input features toward the final task to improve the performance of the model. On the other hand, the learned attention weights are also considered to be an indication of feature importance for interpreting the model.

3.2 Multivariable Representation with Attention Mechanism

In this paper, we are focus on mining with time-series EHR data, where each input x_t is concatenated by multi-variable clinical features $x_t = [x_t^1, x_t^2, ..., x_t^n,]^\intercal$. Notice that vanilla LSTM learns a mixture reprensentation $h_t = LSTM(x_t, h_{t-1})$ which cause attention mechanism is not applicable to learn a meaningful correlation among its inputs.

Here we proposed the MA-LSTM architecture: Multi-variable Attention LSTM to decouple the hidden representation for each semantic inputs and then learn attention weights for imporving the performance of baseline model as well as providing interpretability. Given time-series data $x = [x_1, x_2, ..., x_T]$ where each x_t is multivariable $x_t = [x_t^1, x_t^2, ..., x_t^N,]^\intercal$, we use multi-channel LSTM to learning seperatehidden state for each clinical features $h_t^n = LSTM(x_t^n, h_{t-1}^n)$. Notice that channel-wise feature mapping was originally proposed to speed up the inference of neural network[Howard *et al.*, 2017], here we utilize channel-wise computation to decouple the mixture semantics of hidden state and learning seperate representation for each clinical feature. The t-t hidden state h_t is concatenanted as the representation of each hidden features: $h_t = [h_t^1, h_t^2, ..., h_T^N]$, where $h_t^n = LSTM(x_t^n, h_{t-1}^n)$.

Here $h_t^n \in R^D$ is the hidden state for clinical feature x_n at time position t, and D denote the dimension of hidden states. In our model, we use global average pooling to obtain temporal-irrelated state $h^n = \sum_{t=1}^{T} h_t^n$ since the temporal correlation is irrelevant to the interpretability, which we are interested in. Notice that it is a challenge to apply vanilla attention mechanism to h_t^n considering that directly assigning attention weight α on each h does not learn meaningful feature importance. We derive a statistics $\sum_j^D \beta_j h_j$ for each multivariable hidden state via linear transformation. Thus, the overall attention weights for multivariable time-series data is computed by:

$$\alpha_i = \frac{\prod_{j=1}^{D} e^{W_i \beta_j h_j}}{\sum_{k=1}^{T} \prod_{j=1}^{D} e^{W_k h_k \beta_j}} \tag{3}$$

where $h_t = [h_t^1, h_t^2, ..., h_T^N]$ denotes the hidden states for clinical feature x_i which learned from channel-wise LSTM,

Metrics	Method			
	Logistic Regression	LSTM	Channel-wise LSTM	MALSTM
Task: In-hospital Motarlity				
AUROC	84.50%	85.4%	85.60%	**85.69%**
AUPRC	47.20%	51.6%	50.71%	**51.75%**
Task: Decompensation				
AUROC	87.03%	89.52%	89.87%	**90.34%**
AUPRC	21.37%	32.19%	32.95%	**33.29%**
Task: Length of Stay				
Kappa	0.402	0.436	**0.440**	**0.440**
MSE	63385	46517	44590	**43812**
MAPE	573.5	244.8	209.1	**189.8**
Task: Phenotyping				
Micro AUC	70.13%	75.51%	76.61%	**76.73%**
Macro AUC	74.17%	80.02%	81.72%	**81.80%**
Weigthed AUC	73.28%	74.99%	75.23%	**73.35%**

Table 1: Performance of four MIMIC-III ICU prediction tasks

β denotes that linear transformation factor for h_i and W_i are parameters for learninig attention weights. We learn attention weights $\alpha = [\alpha_1, \alpha_2, ..., \alpha_N]$ which can be regarded as the importance weights for each clinical features and provides intperpretability of the model. Given attention weights α, we use dot-product to derive final feature representation h.

3.3 Learning with Lasso Regularization

Since the learned attention weights are considered to be the feature importance of the model, it is desired to attend on important features while not attending its weight on unimportant features. Intuitively, it can be accomplished by sparsifying attention weights via Lasso regularization. Hence, the proposed model is jointly minimized by the loss function as below:

$$\mathcal{L} = \mathcal{L}_s + \lambda \sum_{i=1}^{L} |\alpha_i| \tag{4}$$

where λ is the sparse coefficient of Lasso regularization and $\mathcal{L}_s$ denotes the supervision loss depends on the specific task. We use the cross-entropy loss for binary classification and multi-label classification, and Euclidean distance loss for regression.

4 Experiment

4.1 Dataset

We conduct experiments using Multi-parameter Intelligent Monitoring in Intensive Care(MIMIC-III benchmarks)[Johnson *et al.*, 2016], which is the largest freely accessible de-identified Intensive Care Unit(ICU) database associated with 53,423 hospital admission for patients admitted to ICU from 2001 to 2012. The MIMIC-III database supports a wide range of healthcare research with heterogeneous clinical records, including laboratory measurements, medical notes charted by providers, diagnostic codes, and more. In order to provide a standardized benchmark for healthcare research based on MIMIC-III database, [Harutyunyan *et al.*, 2019] proposed a time series MIMIC-III benchmark Dataset for four kinds of clinical task: mortality prediction, physiologic decompensation detection, forecasting length of stay, and phenotyping. This benchmark contains some features common to medical data, including varying-length sequences, missing values, and highly skewed distributions. We employ MA-LSTM model to study the four MIMIC-III benchmark tasks. In-hospital mortality prediction is framed as a binary classification task that aims to predict whether a patient dies during hospital admission. The label distribution of the task is highly skewed, where the mortality rate is 13.23%. The decompensation task is formulated as a binary classification task to predict whether a patient will have decompensation and die within the next 24 hours. The distribution in Decompensation is very skew and only 4.2% data are positive in the dataset. The length of stay task aims to forecast the length of patients would stay in hospitals, which can be frame as a regression problem. Phenotyping task is a multi-label classification task that aims to predict possible disease conditions for patients.

Data preprocessing: The raw MIMIC-III dataset is heterogeneous data contains 17 numerical data and categorical data that sampled from irregular time intervals. Categorical variables are encoded as one-hot vectors, and numerical inputs are standardized to be normalized features ranging from 0 to 1. After normalization, the mixed-type data with 17 medical variables become a standardized feature of length 76. We re-sample the irregular time-series data into regular intervals with the value of the last measurements. As for missing value, we impute the missing value with pre-specific normal value if there is no recent measurement that exists. The processed data is converted into variable-length data $\in R^T \times D$, which can be fed into out MA-LSTM model, where D denotes the dimension of features, and T denotes the length of time.

4.2 Performance of MA-LSTM

In this section, we evaluate our MA-LSTM model on the four Mimic-III benchmarks and present comparisons to the baseline LSTM model. We use the receiver operator characteristic curve(AUROC) and the area under the precision-recall curve(AUPRC) as metrics to evaluate the performance of bi-

Feature	Type	Impute value
Capillary refill rate	categorical	’0’
Diastolic blood pressure	numerical	59
Fraction inspired oxygen	numerical	0.21
Glucose	numerical	128
Heart Rate	numerical	86
Height	numerical	170
Mean blood pressure	numerical	77
Oxygen saturation	numerical	98
Respiratory rate	numerical	19
Systolic blood pressure	numerical	118
Temperature	numerical	36.6
Weight	numerical	81
pH	numerical	7.4
Glascow coma scale eye opening	categorical	’4 spontaneously’
Glascow coma scale motor response	categorical	’6 obeys commands’
Glascow coma scale verbal response	categorical	’5 oriented’
Glascow coma scale total	categorical	’15’

Table 2: The 17 clinical features in the MIMIC-III benchmark. The ’Imputate value’ column is the (physiological) normal values for missing value imputation. *Glascow coma scale total* is the derivate categorical feature based on other three *Glascow coma scale* features.

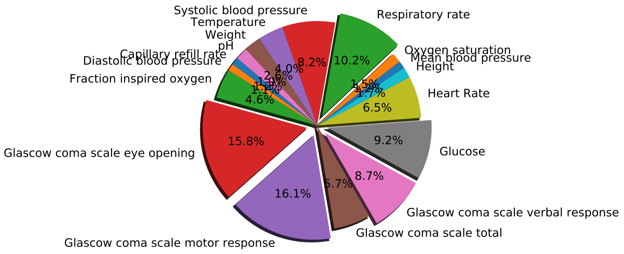

Figure 2: An illutration of local explanation provided by our model. The top-5 important features to be used to predict a patient’s possibility of in-hospital mortality: *Glascow coma scale motor response*, *Glascow coma scale eye opening*, *Respiratory rate*, *Gluscose* and *Glascow coma scale verbal response*.

nary classification tasks, the mean squared error(MSE) and mean absolute percentage error(MAPE) for regression task. As for multi-label classification tasks, we used the macro-averaged AUROC, micro-averaged AUROC, and weighted AUROC as metrics.

Experiment setting: We follow the same experiment setting as [Harutyunyan *et al.*, 2019], use 70% of patients’ records as the training set, 15% of patients’ records as the validation set, and test the models on the rest 15% data. The partition is fixed so we don’t use cross-validation in our experiment setting. For the LSTM-based model we used, the hyperparameters included the unit number of LSTM layer, dropout rate, and the number of layers that have been trained. We trained our model using ADAM optimization with 10^{-3} learning rate, $beta = 0.9$ and batch size = 8. Our best LSTM model for four prediction tasks has 2 LSTM layers used 16 and 32 units, respectively, dropout rate=0.3, and $\lambda = 0.01$. Table I shows that our model has a significant improvement on the baseline.

4.3 Obtain Local Explanation of predictions from Attention Weights

Attention mechanism in our model learns feature importance that integrated into the neural network so that we can obtain a local explanation for its prediction while we inferencing the model with no additional computation cost. Figure 2 shows an illustration of a local explanation produced by our model. It can provide feature importance with visualization for any given input instance.

5 Evaluation of Interpretation

[Alvarez-Melis and Jaakkola, 2018b] proposed three desiderata for explanations in general:

- **Explicitness:** *Are the explanations immediate and understandable?*
- **Faithfulness:** *Are relevance scores indicative of ”true” importance?*
- **Stability:** *How consistent are the local explanations for similar examples?*

Explicitness matters when raw features are not meaningful to humans. For example, pixels in images often lead to meaningless regions to visual perception. In cases where semantic of feature is not available, the basic concepts could be learnt instead.[**?**](For instance, given a neural network recognizing zebras, an explicit explanation should quantify the influence of learnt concepts ’striped’ and ’horse’ to the ’zebra’ prediction)

For tabular-based data, features usually correspond to concepts that humans can understand. Features in EHR data are doctor-provided clinical features(e.g. pH value, respiratory rate) where raw input features are understandable and transparent to human. To this study, explicitness is already satisfied in our case; we mainly focus on faithfulness and stability in our experiments of evaluation of interpretation.

Methods	Addinitional Computation Cost
Occlusion	inference model D times
Saliency	compute $\frac{\partial y}{\partial x_i}$
Attention	no additional cost needed
τ_{loo}	train D models then inference D times

Table 3: Comparison of computation cost for interpretability methods

5.1 Faithfulness Evaluation

Recent studies had shown that interpretability methods might provide local explanations that do not contribute to the model prediction[Jain and Wallace, 2019][Kindermans *et al.*, 2019]. In this section, we introduce interpretability evaluation to EHR domain. Inspired by [Jain and Wallace, 2019], we evaluate the faithfulness of local explanation by comparing it with an intrinsic feature sensitiveness measurement - degradation of model performance induced by leaving feature out(τ_{loo}).

Given a machine leanring model $f : X \to Y$, x_{-t} denotes a derived input with leaving feature x_t out, we can use $\Delta y_t = \|f(x) - f_{-x_t}(x_{-t})\|_2^2$ to estimate the importance the feature x_t since it represents the degree of perturbation of the output when feature x_d has been removed from the input, where f_{-x_t} denotes model trained without x_t. For every input data $x \in R^d$, we can use this approach to compute $\Phi = \{\Delta y_1, \Delta y_2, ..., \Delta y_d\}$ to represent feature importance. Notice that computing Φ is with huge computing load: we need to train additional D models then inference these model D+1 times to every test sample in inference phrase, which is infeasible to be an general approach to estimate feature importance. In our case where D is relatively small(D=17) so that we are able to compute Φ to be an indication of *ground true* feature importance to evaluate faithfulness for other inteprertability methods. Notice that evaluation of *ground true* of local explanation remains an open problem in interpretability research[Hooker *et al.*, 2018], we don't think of τ_{loo} as the *ground true* but as a better estimation of *ground true*.

Given a attention-based model, we can obtain attention interpretation $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_d\}$ from the parameters of the model and no additional computing load is needed. We evaluate whether α can really represent the feature importance by comparing α with feature sensitiveness measurement Φ. Formally, the reliability of attention interpretation is represented by the correlation between α and Φ, denoted as $c(\alpha, \Phi)$. In our study, we compare the local explanation produced by our model with other two widely-used post-hoc interpretability methods, Saliency Map and Occlusion:

- Saliency Map[Simonyan *et al.*, 2013]: compute gradient $\frac{\partial y}{\partial x_i}$ via back propagation to estimate the importance weight for feature x_i
- Occlusion[Zeiler and Fergus, 2014]: occlude feature x_i to derive a new output y_{x_i}, then compute the perturbation $y - y_{x_i}$ to represent the importance weight for x_i.

Table2 shows that comparison of computation cost for interpretability methods.

Notice that α is a probability distribution, while Φ is not a probability distribution as it represents the magnitude of feature perturbation. Probability-based metrics are not applicable to compute $c(\alpha, \Phi)$. In this study, we use two rank-based measures, Kendall correlation to compute $c(\alpha, \Phi)$. Kendall τ correlation is usually used to measure the rank correlation between two variables, ranging from -1 to 1. τ will have a high value if the two variables have a similar rank.

Algorithm 1 Faithfulness Evaluation

```
Require:
    input: $x \in R^d$, model: $f$,
    models trained by leaving feature out: $f_{-x_1}, f_{-x_2}, ..., f_{-x_d}$
Obtain local explanation via interpretability method: $e = \{e_1, e_2, ..., e_d\}$
Obtain original model output: $y = f(x)$
for every feature $x_i$ do
    obtain feature sensetiveness: $\Delta y_i = \|f(x) - f_{-x_i}(x_{-i})\|_2^2$
end for
Obtain feature sensetiveness measurement: $\Phi = \{\Delta y_1, \Delta y_2, ..., \Delta y_d\}$
return Faithfulness evaluation for local explanation: $c(e, \Phi)$
```

The experiment shows that the attention mechanism provide more faithfulness explanation compared with saliency map and occlusion.

5.2 Stability Evaluation

[Alvarez-Melis and Jaakkola, 2018a] claimed that stability(robustness) of explanations is a crucial property of interpretability: an explanation should remain roughly constant in its vicinity. Intuitively, when the input is modified slightly while it does not change the output of the model too much, we would expect that the explanation provided by the interpretability methods does not change too much either. On the other hand, previous studies[Ghorbani *et al.*, 2019][Zheng *et al.*, 2019] generated adversarial perturbations that do not change the model outputs but cause very different interpretation that provided by interpretability methods, which had shown that the vulnerability to adversarial attacks also appears in interpretability.

In this section, we investigate the stability of several interpretability methods with the notion of local Lipschitz continuity:

Definition 1. *$f : X \subseteq R^n \to R^m$ is **locally Lipschitz** if for every x_0 there exist $\delta > 0$ and $L \in R$ such that $\|x - x_0\| < \delta$ implies $\|f(x) - f(x_0)\| \leq L\|x - x_0\|$.*

where L is the local lipschitz constant that measures relative changes in the output with respect to the input. Lipschitz constants are commonly used to characterize the robustness of a function in terms of perturbations. For variable-length data, we used dynamic time warping[Berndt and Clifford, 1994] to compute similarity of time-series data. We are able to quantitively evaluate the robustness for interpretation methods with the notion of local lipschitz. Notice that the continuous notion of local vicinity δ might not be suitable for

Task	Local Lipschitz Estimation			
	Occlusion	Saliency	Attention	Leaving-feature Out
In-hospital Mortality	0.0117± 0.0047	0.0262 ± 0.0067	0.0099 ± 0.0021	**0.0065** ± **0.0017**
Decompensition	0.0183 ±0.0030	0.0264 ± 0.0055	0.0137 ± 0.0018	**0.0124** ± **0.0014**
Length of Stay	0.0201 ± 0.0032	0.0231 ± 0.0049	0.0092 ± 0.0028	**0.0052** ± **0.0011**
Phenotyping	0.2480 ± 0.0527	0.2579 ± 0.0704	0.1086 ± 0.0300	**0.0513** ± **0.0118**

Table 4: Local Lipschitz Estimation for interpretability methods (Mean ± Std.)

Task	Kendall Correlation		
	Occlusion	Saliency	Attention
In-hospital Motarlity	0.207	0.231	**0.424**
Decompensition	0.235	0.243	**0.417**
Length of Stay	0.190	0.223	**0.390**
Phenotyping	0.154	0.198	**0.256**

Table 5: Kendall correlation for three interpretability methods on MIMIC-III benchmarks

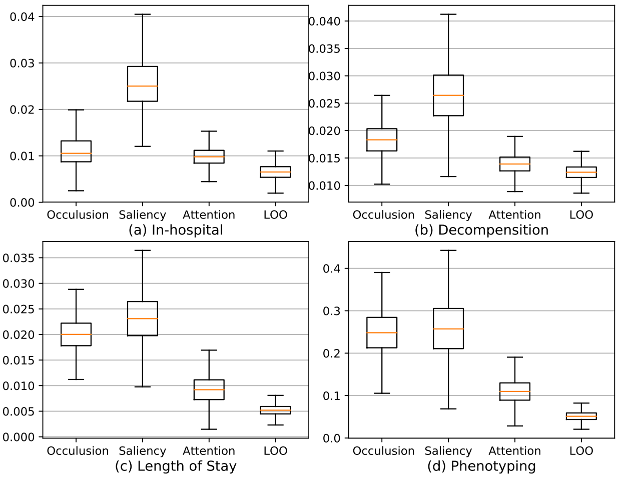

Figure 3: Box-plot of local lipschitz estimation for the four interpretability methods on MIMIC-III benchmarks. (Mean value)

our models with discrete and mix-type clinical data, we instead estimate local lipschitz constant for every test data x_i with with its k-nearest neighbors in the test dataset:

$$\tilde{L}_X\left(x_i\right)=\underset{x_j \in KNN(x_i)}{\operatorname{argmax}} \frac{\left\|f\left(x_i\right)-f\left(x_j\right)\right\|_2}{\left\|x_i-x_j\right\|_2} \tag{5}$$

where KNN(x_i) denotes k-nearest neighbors in terms of euclidean distance, function f denotes a interpretability method and $f(x_i)$ denotes the explanation for x_i provided by interpretability method f. We estimate the local lipschitz constant for 3 interpretablity methods on the MIMIC-III dataset. For k-nearest neighbors, we choose k=100.

The experiment(Table 4) show that our intepretability method is more robust than other methods.

6 Conclusion and Discussion

In this study, we proposed a novel architecture that applies attention mechanism on multivariable time-series EHR data, and its accuracy outperforms the baseline method in four ICU prediction tasks. In addition, we are able to obtain local explanation off-the-shelf when inferencing the model in test phrase. Compared with the other two widely-used post-hoc interpretability methods, our method provides more reliable local explanations with respect to two faithfulness and stability. On the other hand, we work may also indicate that self-explaining methods(models that learn local attention within the model, such as attention mechanism) can provide more reliable explanation than post-hoc explanations with less computation cost.

6.1 Limitations of this work

It is challenging to evaluate the faithfulness of the local explanation produced by interpretability methods. There are important limitations to this study and the conclusion in section 5.1 we draw from it. We used τ_{loo} as the indication of *ground ture* feature importance to evaluate the faithfulness of other methods. Actually, there is no ground truth. Defining a ground truth of feature importance remains an open problem in interpretable ML research[Hooker *et al.*, 2018]. If we were able to find a ground true feature importance, we would not need interpretability methods in the first place. We do not imply that τ_{loo} should be regarded as the ground truth.

There is a fundamental assumption in section 5.1: compared with other heuristic interpretability methods(e.g. attention mechanism, saliency map, occlusion), τ_{loo} produce intrinsic feature importance with mathematic insight, and it is assumed to be much closer to the ground truth.

Obtaining τ_{loo} is very computational expensive since there are D different models needed to be trained for each task., where D denotes the number of features. We are able to compute τ_{loo} in MIMIC-III benchmarks because D is relatively small(D=17) in our case. While in most cases, computing τ_{loo} is impossible due to the number of medical features is usually very large.

Finally, we have limited our evaluation to feature correlation, which actually quite common in medical datasets. A naive solution to consider feature correlation is to take feature combination into account, however the number of models needed to be trained will increase from D to 2^D, which is completely impossible to obtain. On the other hand, the experiment in section 5.2 shows that τ_{loo} perform more stable than other methods even with the limitation we discussed above, which also indicate that τ_{loo} is more ”reliable” from another aspect.

Acknowledgement

This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200.

References

[Alvarez-Melis and Jaakkola, 2018a] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

[Alvarez-Melis and Jaakkola, 2018b] David Alvarez-Melis and Tommi S Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7786–7795. Curran Associates Inc., 2018.

[Berndt and Clifford, 1994] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.

[Choi *et al.*, 2016a] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1495–1504. ACM, 2016.

[Choi *et al.*, 2016b] Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2):361–370, 2016.

[Choi *et al.*, 2018] Edward Choi, Cao Xiao, Walter Stewart, and Jimeng Sun. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4547–4557. Curran Associates, Inc., 2018.

[Ghorbani *et al.*, 2019] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019.

[Harutyunyan *et al.*, 2019] Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019.

[Hooker *et al.*, 2018] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. Evaluating feature importance estimates. *arXiv preprint arXiv:1806.10758*, 2018.

[Howard *et al.*, 2017] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[Jain and Wallace, 2019] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

[Johnson *et al.*, 2016] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[Kim *et al.*, 2016] Been Kim, Rajiv Khanna, and Sanmi Koyejo. Examples are not enough, learn to criticize! Criticism for interpretability. In *Advances in Neural Information Processing Systems*, 2016.

[Kindermans *et al.*, 2017] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. The (Un)reliability of saliency methods. *NIPS workshop on Explaining and Visualizing Deep Learning*, 2017.

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

[Kwon *et al.*, 2018] Bum Chul Kwon, Min-Je Choi, Joanne Taery Kim, Edward Choi, Young Bin Kim, Soonwook Kwon, Jimeng Sun, and Jaegul Choo. 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.

[Li *et al.*, 2019a] Yikuan Li, Shishir Rao, Jose Roberto Ayala Solares, Abdelaali Hassaïne, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi Khorshidi. BEHRT: transformer for electronic health records. *CoRR*, abs/1907.09538, 2019.

[Li *et al.*, 2019b] Yikuan Li, Shishir Rao, Jose Roberto Ayala Solares, Abdelaali Hassaine, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: Transformer for electronic health records. *arXiv preprint arXiv:1907.09538*, 2019.

[Lipton *et al.*, 2015] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

[Perros *et al.*, 2018] Ioakeim Perros, Evangelos E Papalexakis, Haesun Park, Richard Vuduc, Xiaowei Yan, Christopher Defilippi, Walter F Stewart, and Jimeng Sun. Sustain: Scalable unsupervised scoring for tensors and its application to phenotyping. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2080–2089. ACM, 2018.

[Pham *et al.*, 2016] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. Deepcare: A deep dynamic memory model for predictive medicine. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 30–41. Springer, 2016.

[Shang *et al.*, 2019] Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346*, 2019.

[Shickel *et al.*, 2017] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.

[Simonyan *et al.*, 2013] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[Zeiler and Fergus, 2014] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[Zhang *et al.*, 2018] Jinghe Zhang, Kamran Kowsari, James H Harrison, Jennifer M Lobo, and Laura E Barnes. Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record. *IEEE Access*, 6:65333–65346, 2018.

[Zhao and Weng, 2011] Di Zhao and Chunhua Weng. Combining pubmed knowledge and ehr data to develop a weighted bayesian network for pancreatic cancer prediction. *Journal of biomedical informatics*, 44(5):859–868, 2011.

[Zheng *et al.*, 2019] Haizhong Zheng, Earlence Fernandes, and Atul Prakash. Analyzing the interpretability robustness of self-explaining models. *arXiv preprint arXiv:1905.12429*, 2019.