TRACER: A Framework for Facilitating Accurate and Interpretable Analytics for High Stakes Applications

Kaiping Zheng, Shaofeng Cai, Horng Ruey Chua,
Wei Wang, Kee Yuan Ngiam, Beng Chin Ooi

Speaker: Kaiping Zheng

Outline

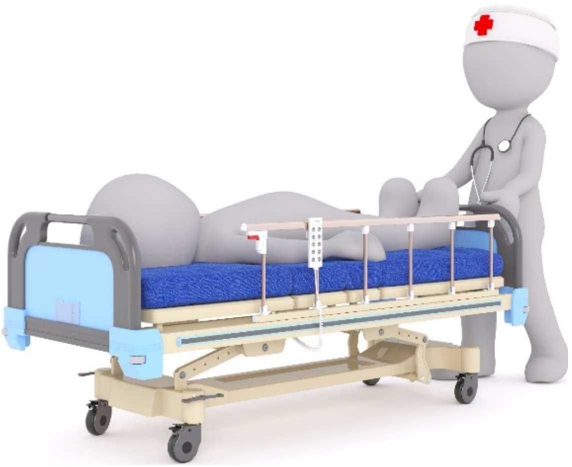
- Introduction
- TRACER Framework
- TITV Model
- Evaluation
- Conclusion

Introduction

- *Healthcare analytics refers to data analytics on a selected cohort of patients for tasks like diagnosis, prognosis, etc*
- *Neural network based models have emerged to improve the accuracy over traditional machine learning models*
- *An accurate analytic model helps healthcare workers and organizations make effective decisions on patient management and resource allocation, and thus reduces healthcare cost*
- *However, accuracy alone is not sufficient*

Introduction

- ***If train an accurate model for in-hospital mortality prediction***



"Our model predicts this patient has a 26% probability of mortality."

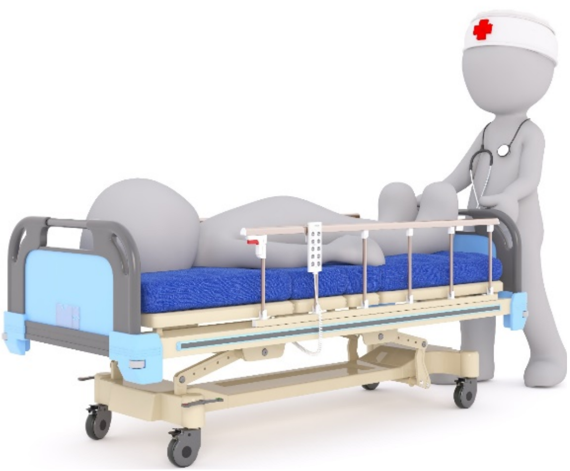
Introduction

- ***If train an accurate model for in-hospital mortality prediction***

"Our model predicts this patient has a 26% probability of mortality."

Introduction

- ***If train an accurate model for in-hospital mortality prediction***



"Our model predicts this patient has a 26% probability of mortality."

- ***This is unacceptable to doctors***
- Cannot trust our model if there is no explanation of the prediction results

- ***Essential to devise a model which can derive interpretable as well as medically meaningful results***

Introduction

- ***Feature - "time-invariant" and "time-variant" feature importance***
- Exhibit a kind of time-invariant influence on a patient over the whole time series
- Its influence also has some variations in different time periods or visits

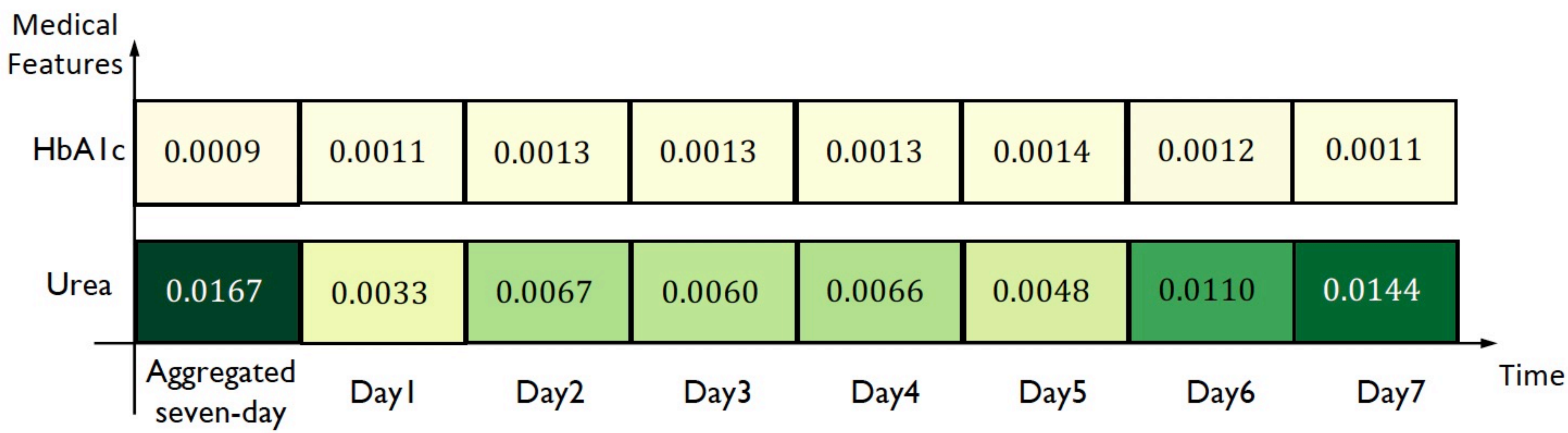


Figure: The normalized coefficients in both an LR model trained on the aggregated seven-day data (leftmost) and seven LR models trained separately. We illustrate with two representative laboratory tests HbA1c and Urea.

Introduction

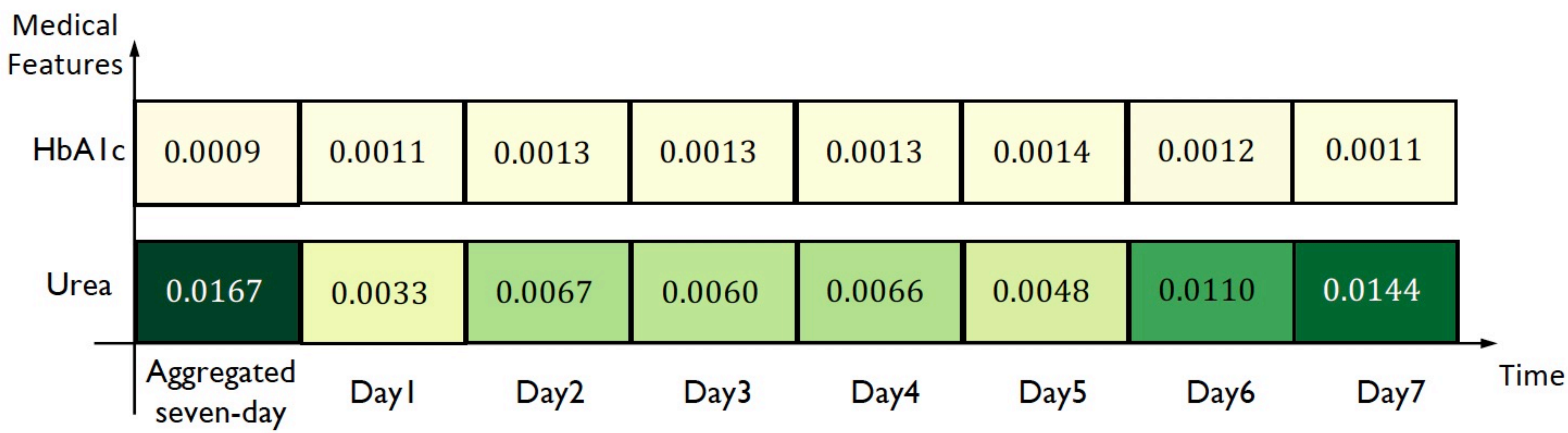


Figure: The normalized coefficients in both an LR model trained on the aggregated seven-day data (leftmost) and seven LR models trained separately. We illustrate with two representative laboratory tests HbA1c and Urea.

HbA1c →
Urea →
Risk of developing Acute Kidney Injury (AKI)

Introduction

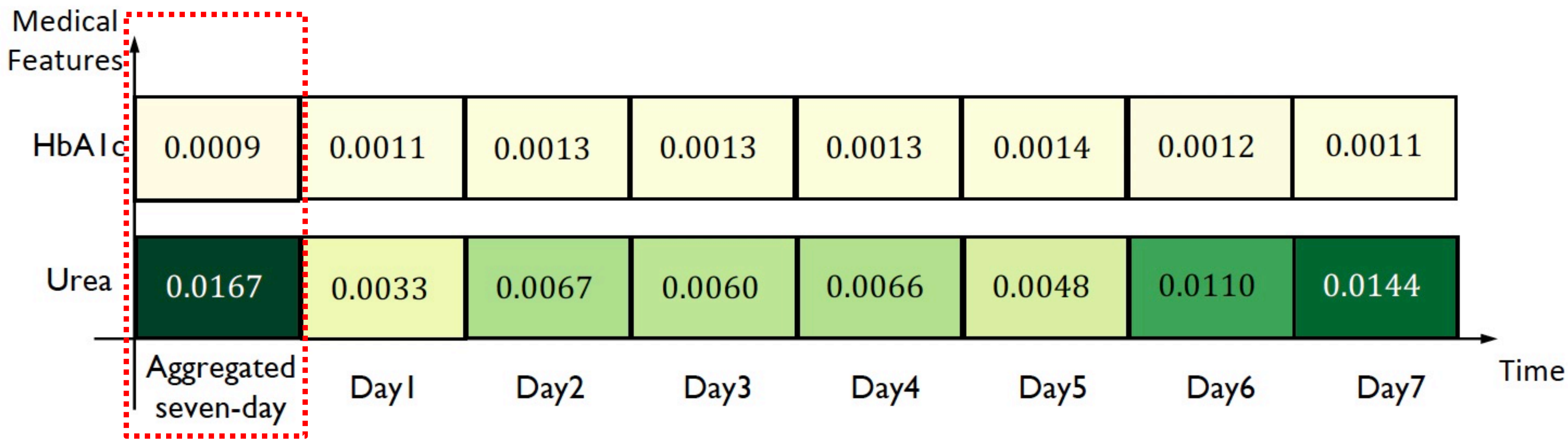


Figure: The normalized coefficients in both an LR model trained on the aggregated seven-day data (leftmost) and seven LR models trained separately. We illustrate with two representative laboratory tests HbA1c and Urea.

Time-Invariant Feature Importance

Introduction

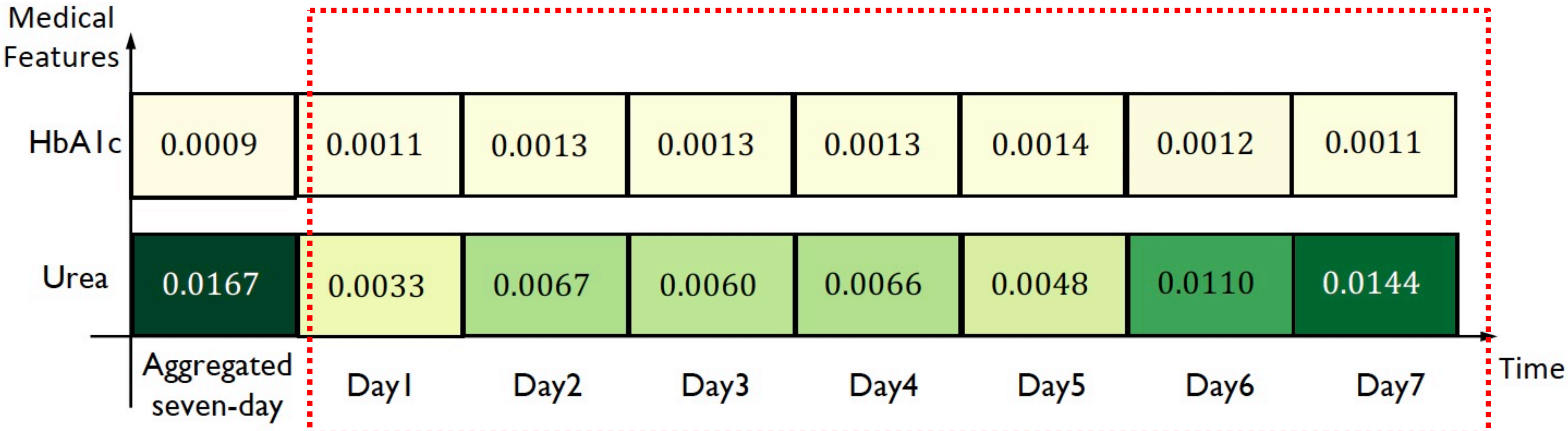


Figure: The normalized coefficients in both an LR model trained on the aggregated seven-day data (leftmost) and seven LR models trained separately. We illustrate with two representative laboratory tests HbA1c and Urea.

Time-Variant Feature Importance

Introduction

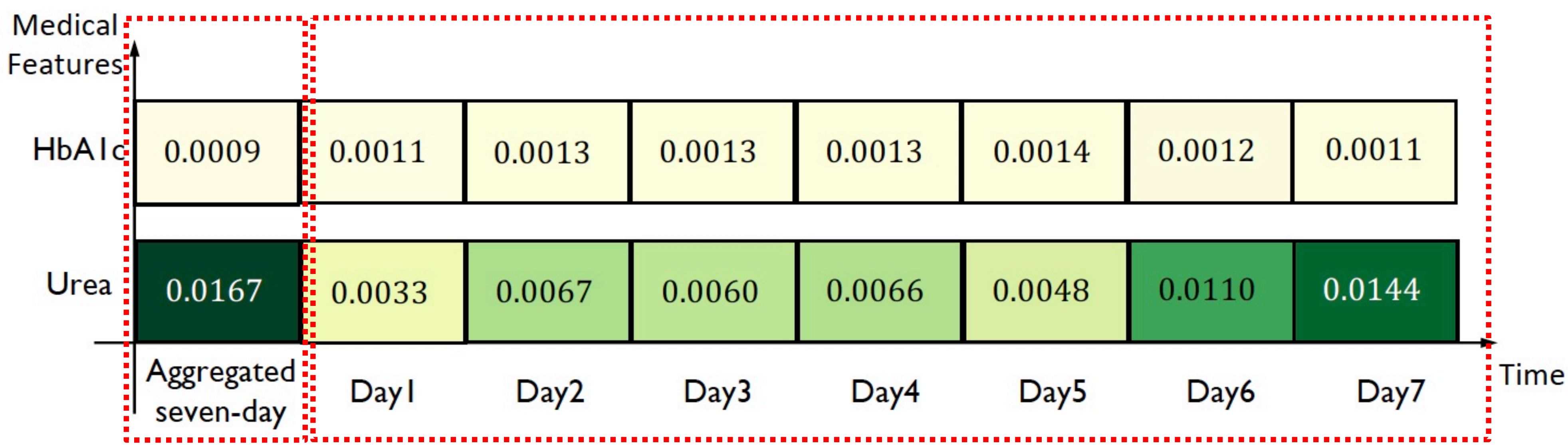


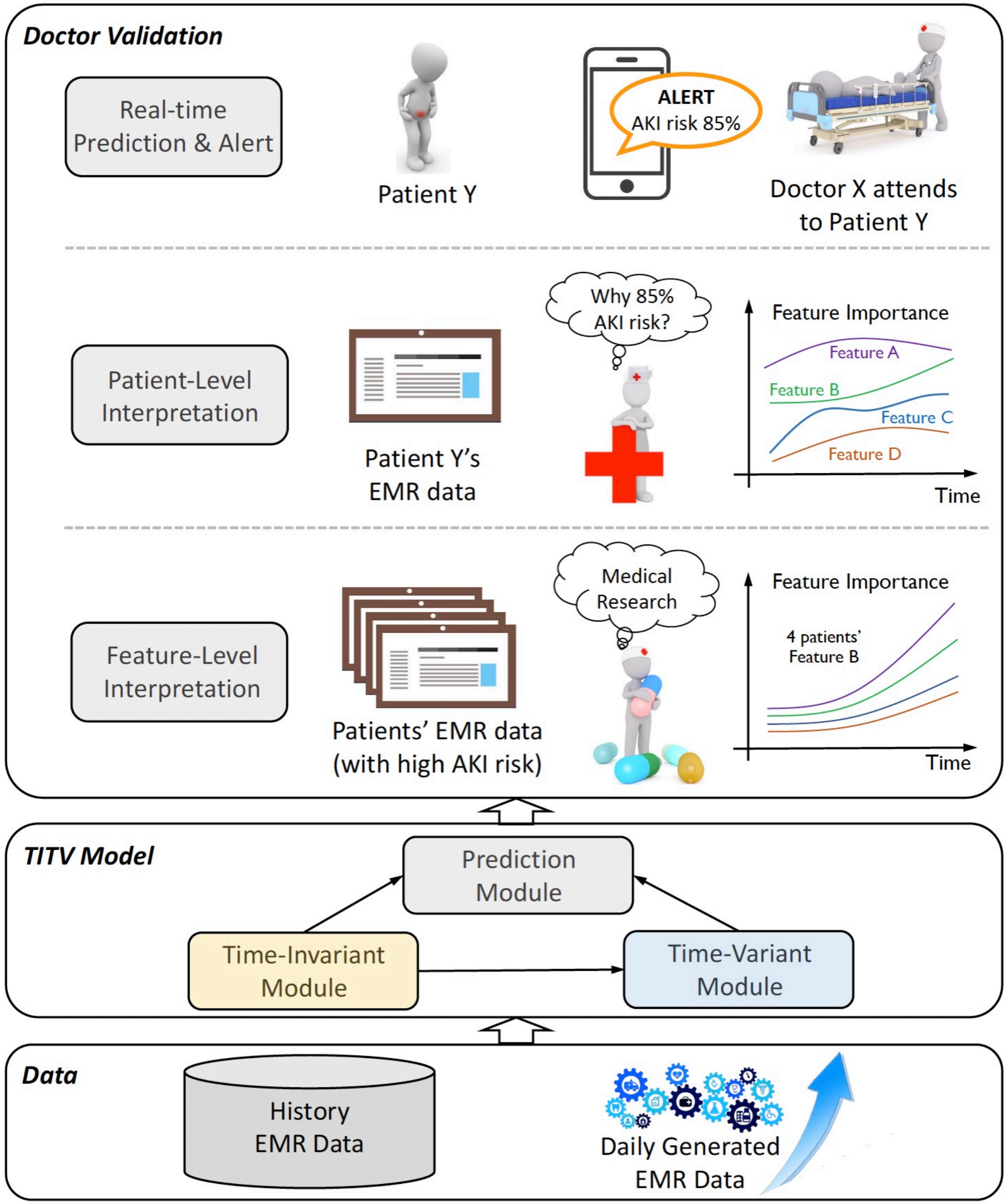
Figure: The normalized coefficients in both an LR model trained on the aggregated seven-day data (leftmost) and seven LR models trained separately. We illustrate with two representative laboratory tests HbA1c and Urea.

- Existing approaches do not differentiate time-invariant and time-variant feature importance (e.g., Choi et al. 2016; Ma et al. 2017; Sha et al. 2017)

Outline

- Introduction
- TRACER Framework
- TITV Model
- Evaluation
- Conclusion

TRACER Framework



Outline

- Introduction
- TRACER Framework
- TITV Model
- Evaluation
- Conclusion

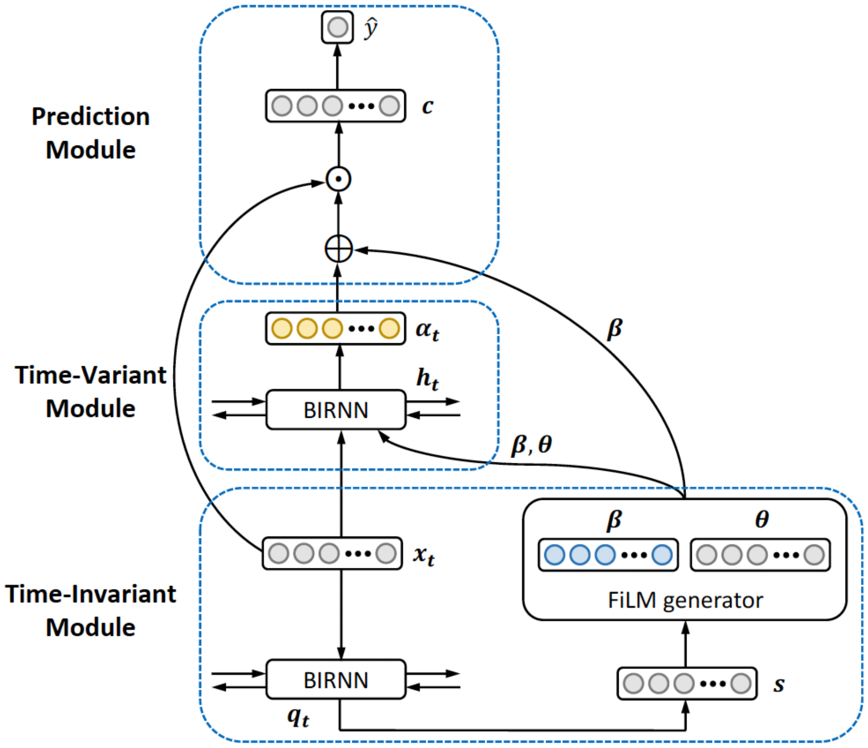
TITV Model

- ***TITV: an interpretable model capturing both time-invariant and time-variant feature importance for each sample***

- Time-Invariant Module
  - → time-invariant feature importance
  - via FiLM mechanism

- Time-Variant Module
  - → time-variant feature importance
  - via self-attention mechanism

- Prediction Module
  - → derive TITV's final prediction



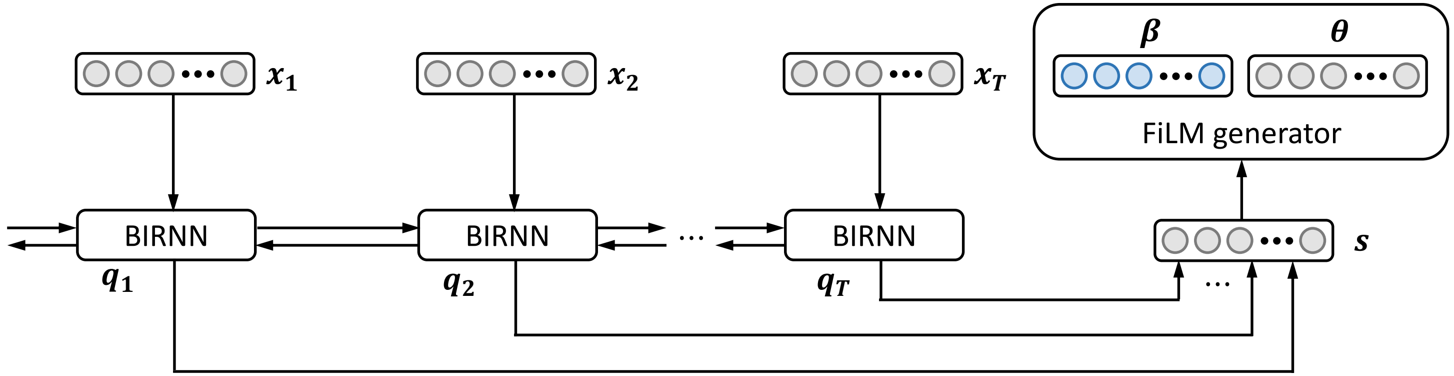
Time-Invariant Module

- ***Aim: model the time-invariant feature importance shared across time where data in all time windows are exploited***

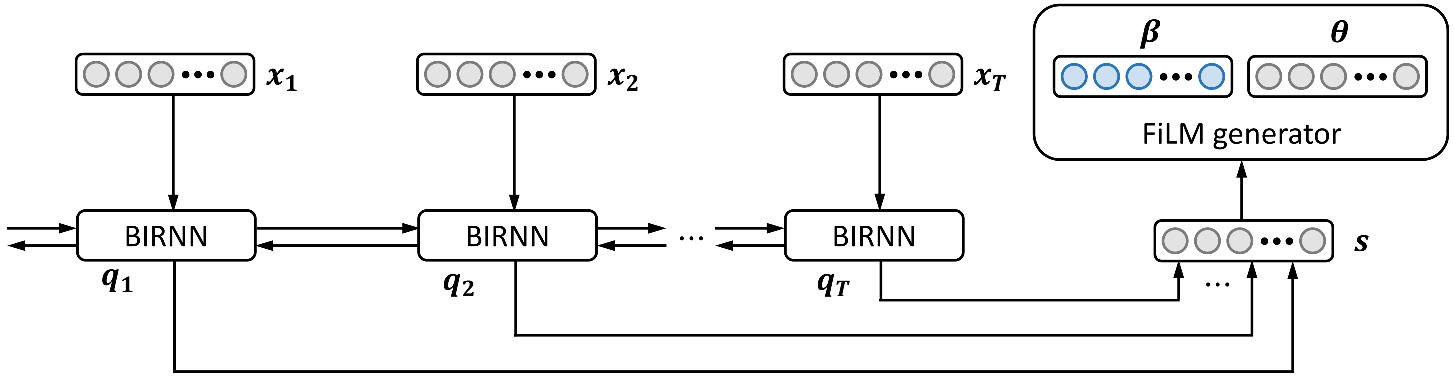
- ***FiLM - feature-wise linear modulation***
- → good at modelling feature importance

  (Dumoulin et al. 2018, Kim et al., 2017, Perez et al., 2018)

- ***Integrate FiLM in Time-Invariant Module***



Time-Invariant Module



- Bi-directional RNN computation → capture both the forward and the backward temporal relationship

$$(\boldsymbol{q}_1, \cdots, \boldsymbol{q}_t, \cdots, \boldsymbol{q}_T) = BIRNN(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t, \cdots, \boldsymbol{x}_T)$$

- Summary vector computation → utilize all available data in all time windows.

$$\boldsymbol{s} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{q}_t$$

- FiLM generator → compute scaling parameter $\boldsymbol{\beta}$ and shifting parameter $\boldsymbol{\theta}$

$$\boldsymbol{\beta} = \boldsymbol{W}_{\boldsymbol{\beta}}\boldsymbol{s} + \boldsymbol{b}_{\boldsymbol{\beta}}$$

$$\boldsymbol{\theta} = \boldsymbol{W}_{\boldsymbol{\theta}}\boldsymbol{s} + \boldsymbol{b}_{\boldsymbol{\theta}}$$

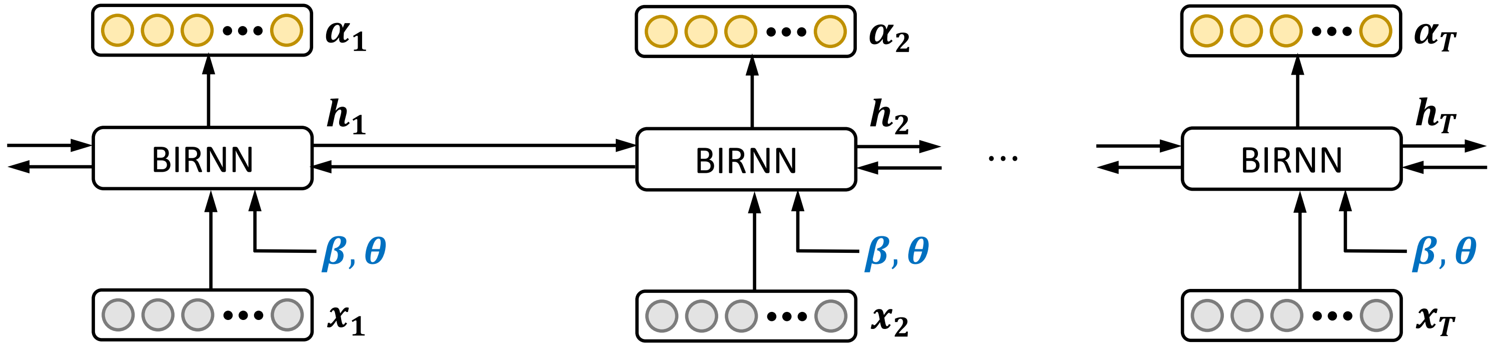
Time-Variant Module

- ***Aim: differentiate the influence of different features in different time windows***

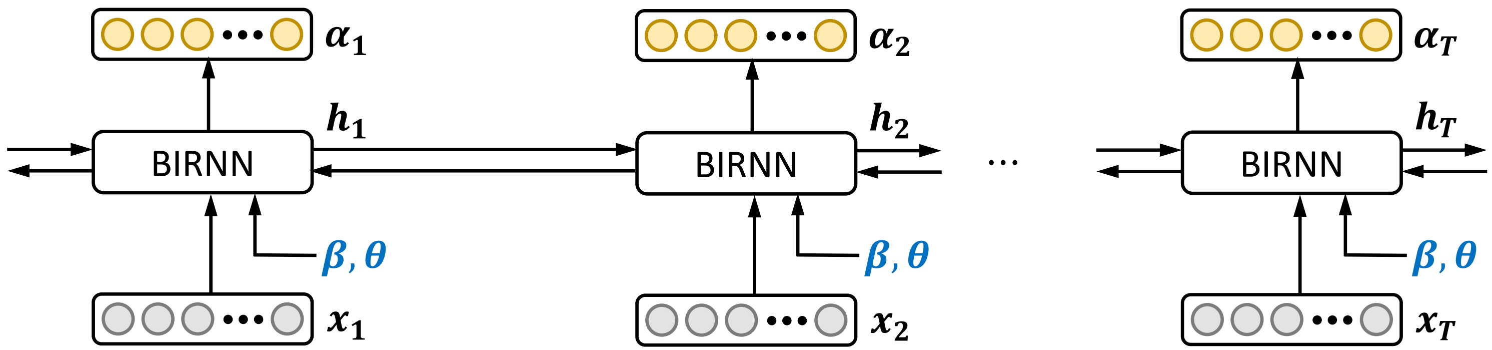
- ***Self-attention mechanism***
- → successfully applied for many similar tasks

(Cheng et al. 2016, Xu et al., 2015)

- ***Integrate self-attention mechanism in Time-Variant Module***



Time-Variant Module



- Process time-series input data via $BIRNN_{FiLM}$

$$(\boldsymbol{h_1}, \cdots, \boldsymbol{h_t}, \cdots, \boldsymbol{h_T}) = BIRNN_{FiLM}(\boldsymbol{x_1}, \cdots, \boldsymbol{x_t}, \cdots, \boldsymbol{x_T}; \boldsymbol{\beta}, \boldsymbol{\theta})$$

- $BIRNN_{FiLM}$ computation, with $FiLM(\boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\theta}) = \boldsymbol{\beta} \odot \boldsymbol{x} + \boldsymbol{\theta}$

$$\boldsymbol{z_t} = \sigma(FiLM(\boldsymbol{W_z x_t}; \boldsymbol{\beta}, \boldsymbol{\theta}) + \boldsymbol{U_z h_{t-1}})$$

$$\boldsymbol{r_t} = \sigma(FiLM(\boldsymbol{W_r x_t}; \boldsymbol{\beta}, \boldsymbol{\theta}) + \boldsymbol{U_r h_{t-1}})$$

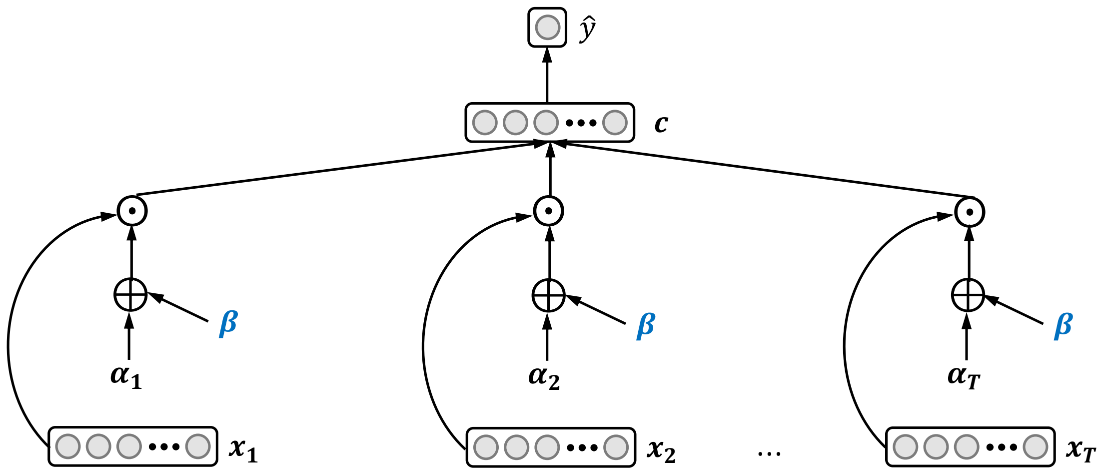
$$\widetilde{\boldsymbol{h_t}} = \tanh\left(FiLM\left(\widetilde{\boldsymbol{W}}\boldsymbol{x_t}; \boldsymbol{\beta}, \boldsymbol{\theta}\right) + \boldsymbol{r_t} \odot \widetilde{\boldsymbol{U}}\boldsymbol{h_{t-1}}\right)$$

$$\boldsymbol{h_t} = (\boldsymbol{1} - \boldsymbol{z_t}) \odot \widetilde{\boldsymbol{h_t}} + \boldsymbol{z_t} \odot \boldsymbol{h_{t-1}}$$

- Self-attention mechanism

$$\boldsymbol{\alpha_t} = \tanh(\boldsymbol{W_\alpha h_t} + \boldsymbol{b_\alpha})$$

Prediction Module



- Obtain the overall influence from time-invariant and time-variant feature importance

$$\boldsymbol{\xi}_t = \boldsymbol{\beta} \oplus \boldsymbol{\alpha}_t$$

- Compute context vector by summarizing information at each time window t

$$\boldsymbol{c} = \sum_{t=1}^{T} \boldsymbol{\xi}_t \odot \boldsymbol{x}_t$$

- Derive final predicted label

$$\hat{y} = \sigma(\langle \boldsymbol{w}, \boldsymbol{c} \rangle + b)$$

Feature Importance $FI(\hat{y}, x_{t,d})$

- ***Risk of a sample falling into the positive class $\hat{y}$***

$$\hat{y} = \sigma\left(\sum_{t=1}^{T} \langle \boldsymbol{w}, (\boldsymbol{\beta} \oplus \boldsymbol{\alpha}_t) \odot \boldsymbol{x}_t \rangle + b\right)$$

- ***$x_{t,d}$'s Feature Importance to TITV's predicted label $\hat{y}$***

$$FI(\hat{y}, x_{t,d}) = (\beta_d + \alpha_{t,d}) \cdot w_d$$

- ***All appearing features collaboratively contribute to $\hat{y}$***

$$\hat{y} = \sigma\left(\sum_{t=1}^{T} \sum_{d=1}^{D} FI(\hat{y}, x_{t,d}) \cdot x_{t,d} + b\right)$$

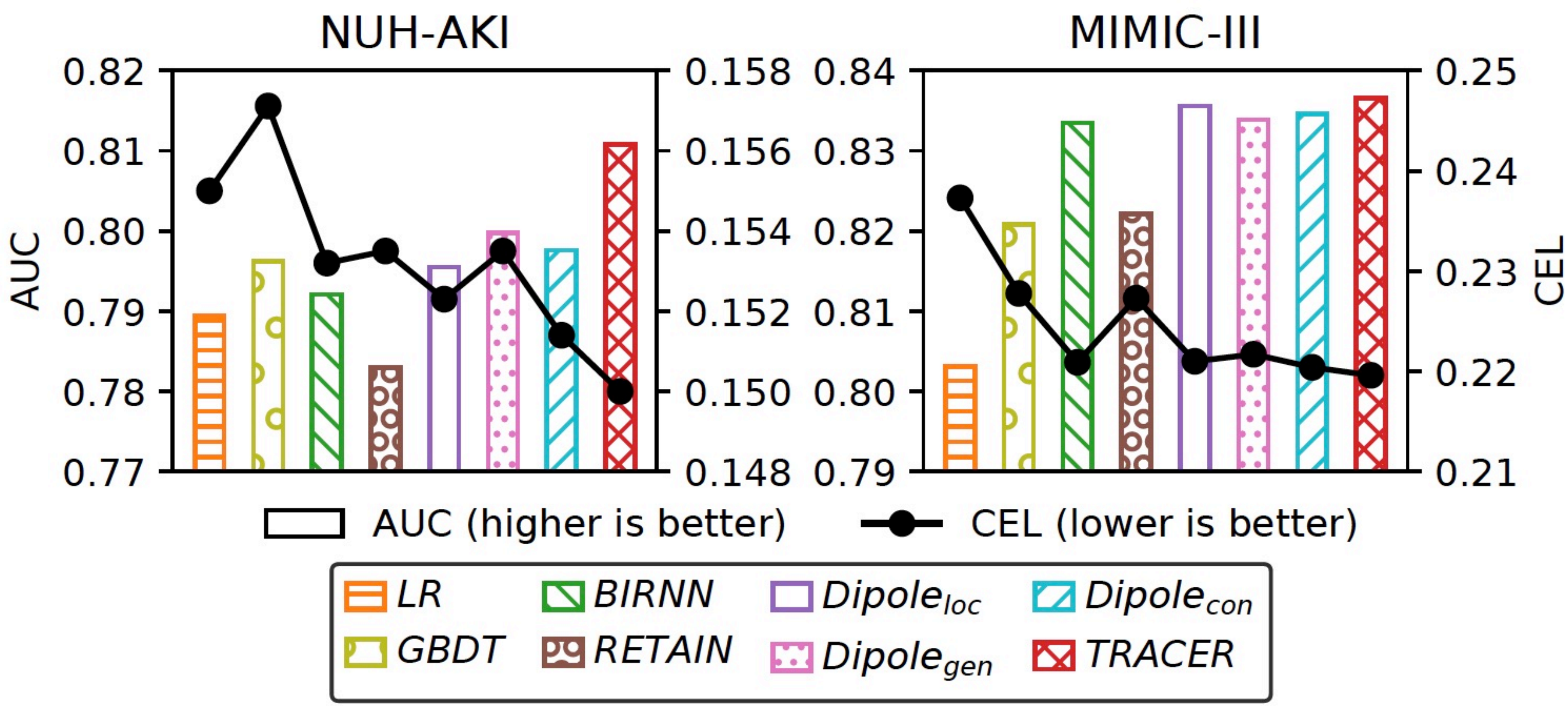
Outline

- Introduction
- TRACER Framework
- TITV Model
- Evaluation
- Conclusion

Evaluation

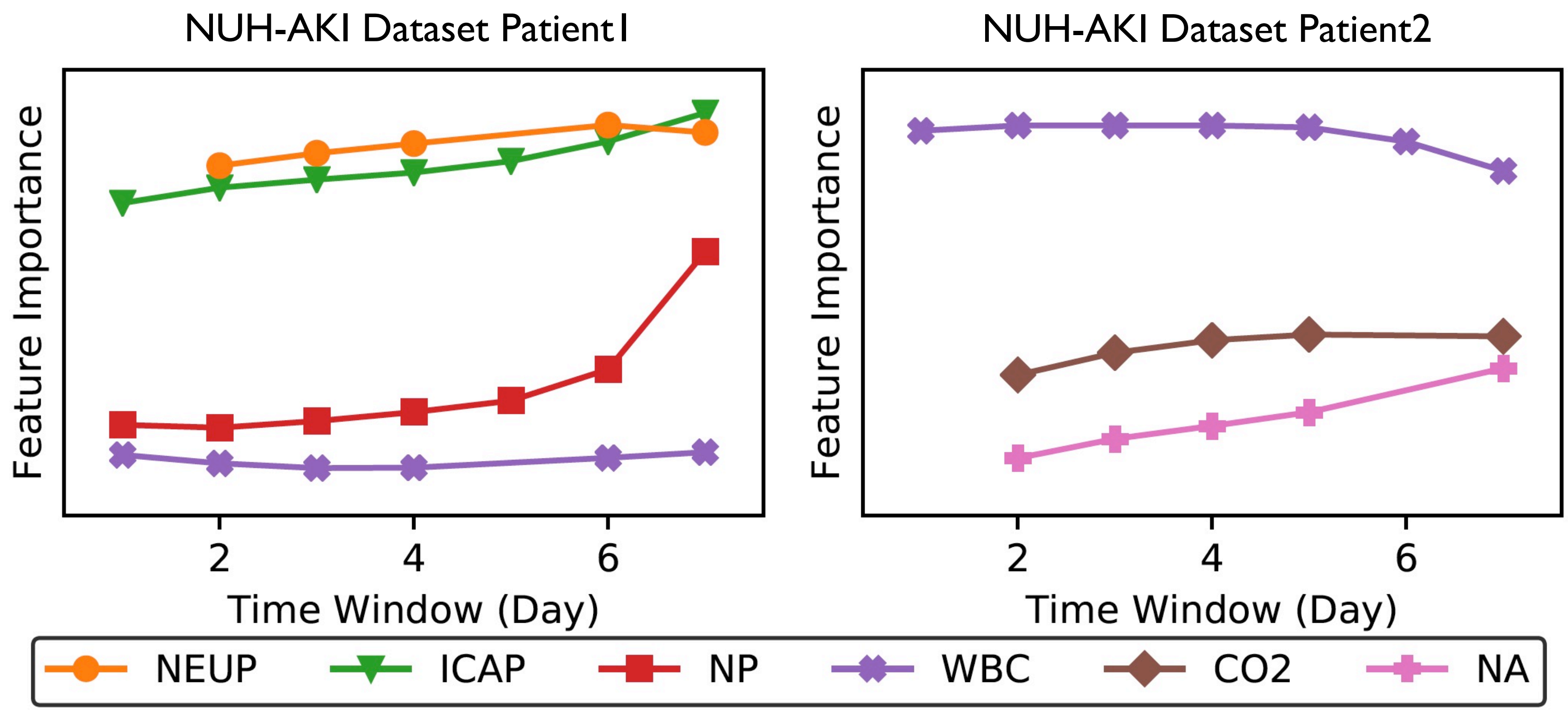
- ***Datasets and Applications***
  - NUH-AKI dataset - hospital-acquired AKI prediction
  - MIMIC-III dataset - in-hospital mortality prediction
- ***Baselines***
  - LR
  - GBDT
  - BIRNN
  - RETAIN (Choi et al. 2016)
  - Dipole (Dipole_{loc}, Dipole_{gen}, Dipole_{con}) (Ma et al. 2017)
- ***Prediction Results***
  - comparison results in terms of AUC and CEL
- ***Interpretation Results***
  - patient-level interpretation & feature-level interpretation

Evaluation



- TRACER outperforms LR and GBDT
- TRACER outperforms RETAIN
- TRACER achieves better prediction performance than BIRNN and Dipole

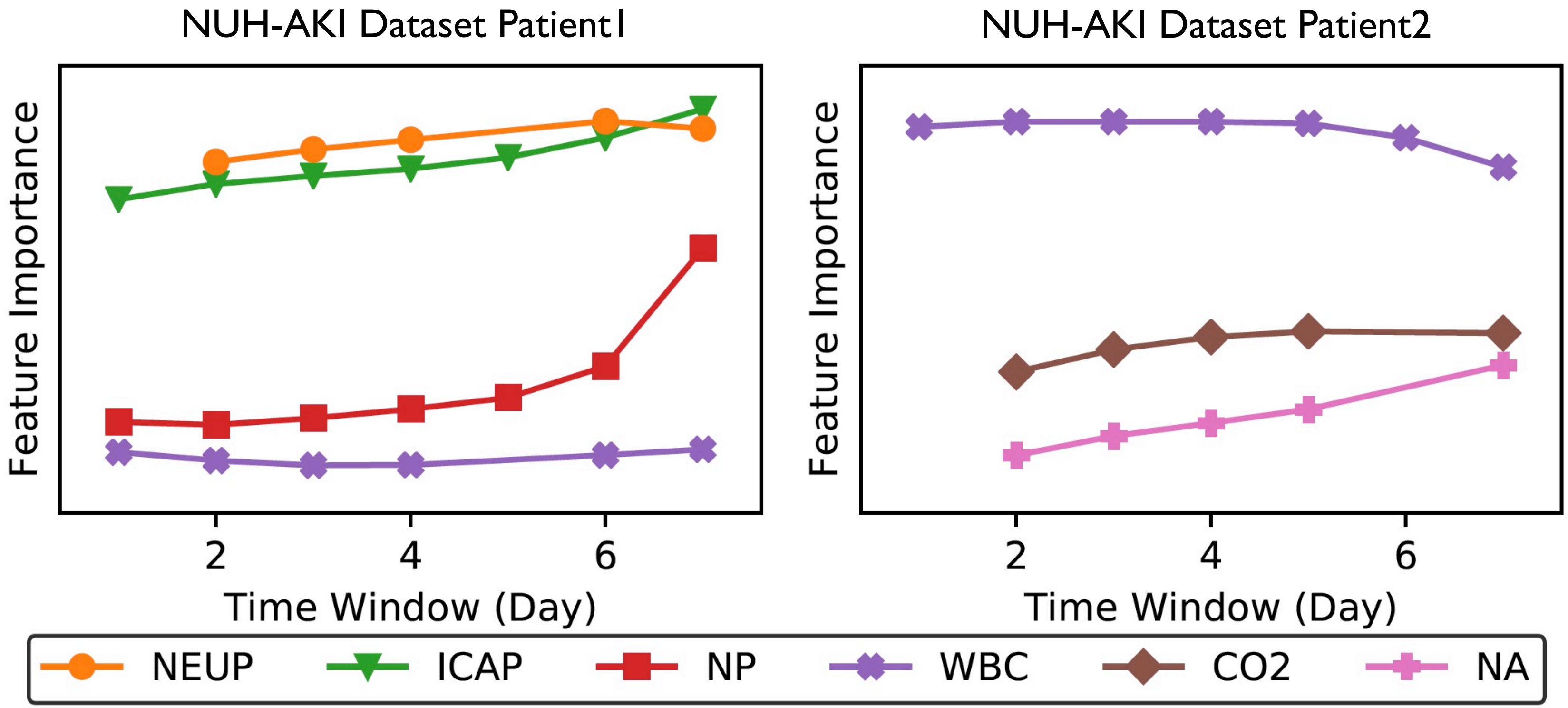
Patient-Level Interpretation



Features involved: "Neutrophils %" (NEUP), "Ionised CA, POCT" (ICAP), "Sodium, POCT" (NP), "White Blood Cell" (WBC), "Carbon Dioxide" (CO2) and "Serum Sodium" (NA).

- Patient1
  - NEUP and WBC: worsening infection
  - ICAP and NP: worsening electrolyte imbalance

Patient-Level Interpretation

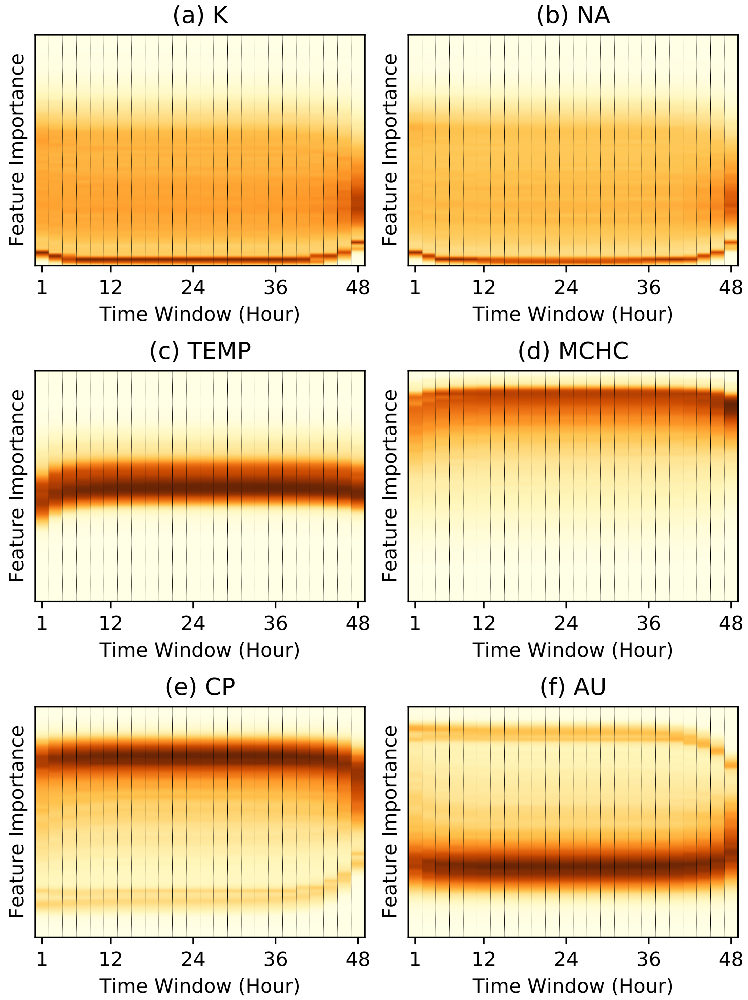


Features involved: "Neutrophils %" (NEUP), "Ionised CA, POCT" (ICAP), "Sodium, POCT" (NP), "White Blood Cell" (WBC), "Carbon Dioxide" (CO2) and "Serum Sodium" (NA).

- Patient2
  - WBC: presence of inflammation or infection
  - CO2: acidosis that builds up with progressive kidney dysfunction
  - NA: progressive NA-fluid imbalance and worsening kidney function

Feature-Level Interpretation

MIMIC-III Dataset



Features involved: "Serum Potassium" (K), "Serum Sodium" (NA), "Temperature" (TEMP), "Mean Corpuscular Hemoglobin Concentration" (MCHC), "Cholesterol, Pleural" (CP) and "Amylase, Urine" (AU).

- ***Low Feature Importance detected for common features which are not generally highly related to mortality***
  - K & NA
- ***High Feature Importance detected for common features that are generally highly related to mortality***
  - TEMP & MCHC
- ***Same feature's diverging patterns indicate different patient clusters***
  - CP & AU

Outline

- Introduction
- TRACER Framework
- TITV Model
- Evaluation
- Conclusion

Conclusion

- ***Capture the feature importance in two aspects***
- Time-invariant feature importance: overall influence of feature shared across time
- Time-variant feature importance: time-related influence varying along with time
- ***Propose TRACER framework***
- provide accurate and interpretable clinical decision support to doctors
- ***Devise an interpretable model TITV in TRACER***
- Time-invariant feature importance via FiLM mechanism
- Time-variant feature importance via self-attention mechanism
- ***Evaluate the effectiveness of TRACER***
- Prediction performance
- Interpretation capability: both patient-level and feature-level

Thank you!

NUS
National University
of Singapore

NUHS
National University
Health System