Interpretable Deep Models for ICU Outcome Prediction

Zhengping Che[1], Sanjay Purushotham[1], Robinder Khemani[2], Yan Liu[1]
[1]University of Southern California, [2]Children's Hospital Los Angeles

Edward Chou, William Du, Kevin Looby, John Louie
Group 16

Introduction

- Surge in health care data (*e.g.*, longitudinal data from electronic health records (EHR), sensor data from intensive care unit (ICU)
- Deep learning models have been effectively applied to healthcare prediction tasks
- Deep models are difficult to interpret
- An interpretable predictive model should result in faster clinical adoption

Goal

Develop a data-driven solution satisfying:

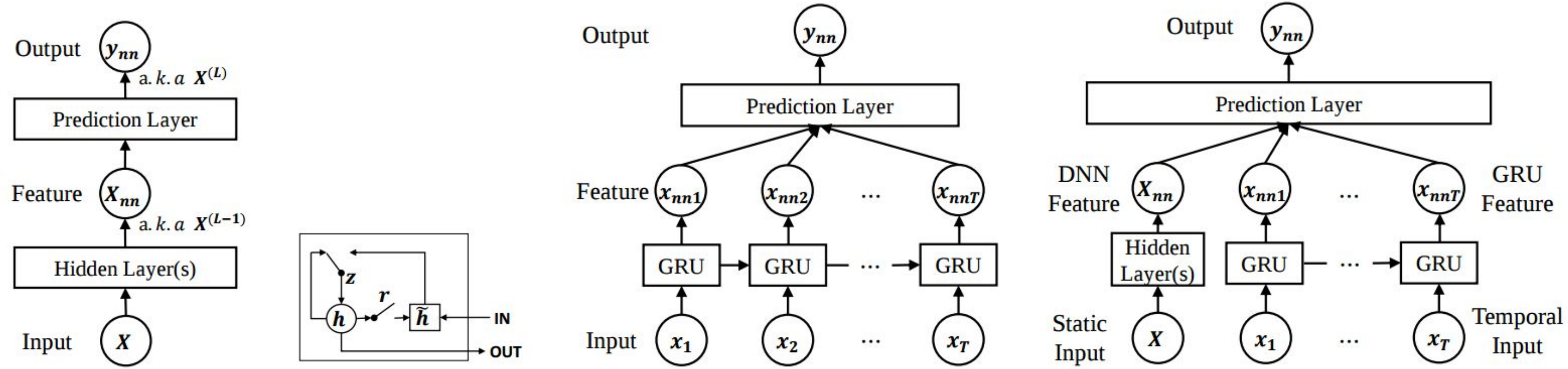
1. Achieves performance comparable to state-of-the-art deep learning models
2. Can be easily interpreted by healthcare professionals

Overview of Approach

- Employ mimic learning to learn an interpretable model
    - Knowledge distillation approach called “interpretable mimic learning”
    - Make use of gradient boosting trees (GBT) instead of standard shallow neural networks or kernel methods
- Investigate using feed-forward networks and recurrent neural networks for predicting mortality and ventilator free days (VFD) using a pediatric ICU dataset

Background: Deep Learning Models

- Feedforward networks + gated recurrent units



(a) DNN prediction model. (b) Gated Recurrent Unit (GRU). (c) GRU prediction model. (d) DNN and GRU combination model.

Interpretable Mimic Learning/Knowledge Distillation

- Knowledge distillation - Train large, slow, accurate model and transfer the model to a shallow fast model
- Soft labels learned from complex model used as the Y for the small model

Interpretable Mimic Learning Framework

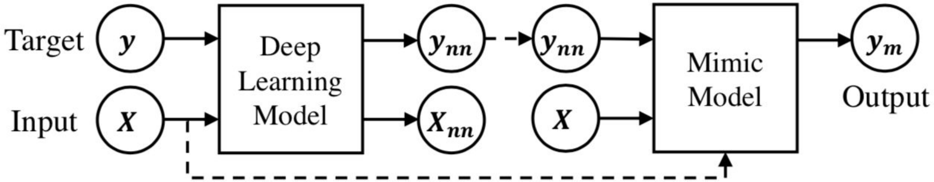


Figure 2: Illustration of mimic method training pipeline 1.

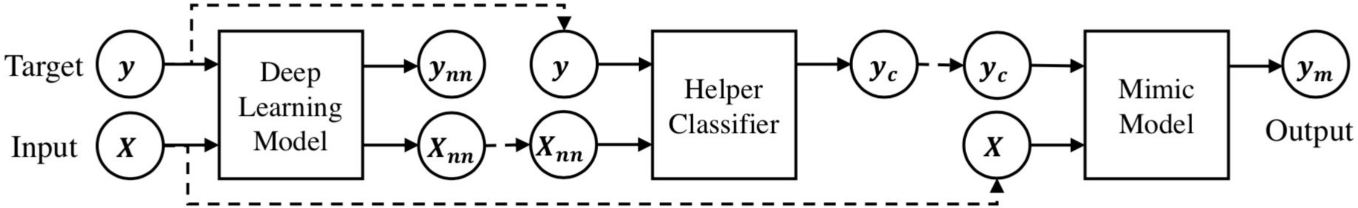


Figure 3: Illustration of mimic method training pipeline 2.

Interpretable Mimic Learning/Knowledge Distillation

- Eliminates potential noise and error in the student model
- Soft labels are more informative than original hard labels
- Implicit regularization on teacher model transfers over and prevents overfitting
- Shallow Models:
    - Shallow Neural Networks
    - Kernelized Methods (SVM)
    - Decision Trees

Visualizing Gradient Boosted Trees

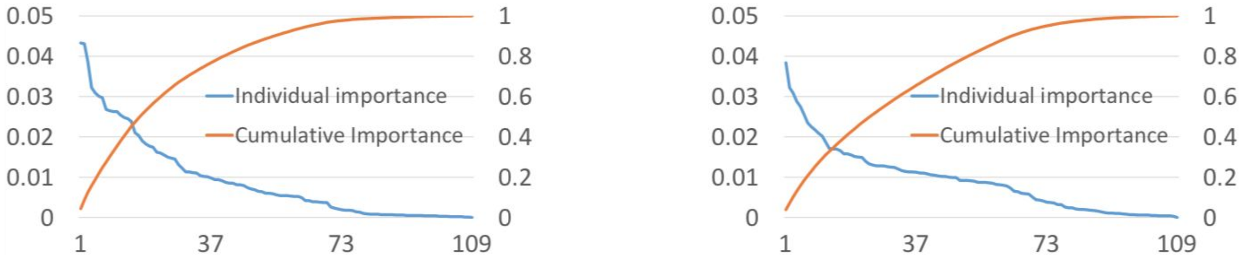


Figure 4: Individual (with left y-axis) and cumulative (with right y-axis) feature importance for MOR (top) and VFD (bottom) tasks. x-axis: sorted features.

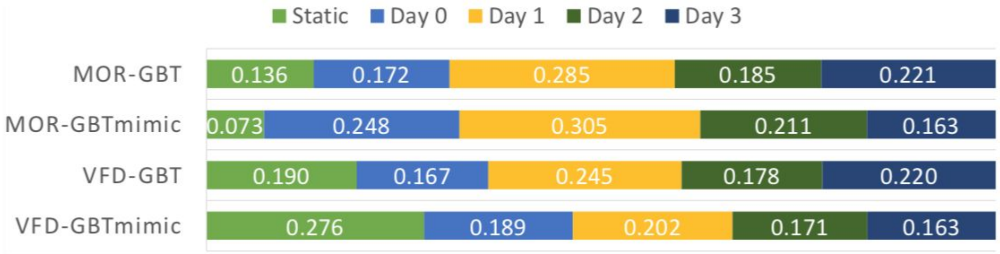


Figure 5: Feature importance for static features and temporal features on each day for two tasks.

Visualizing Gradient Boosted Trees

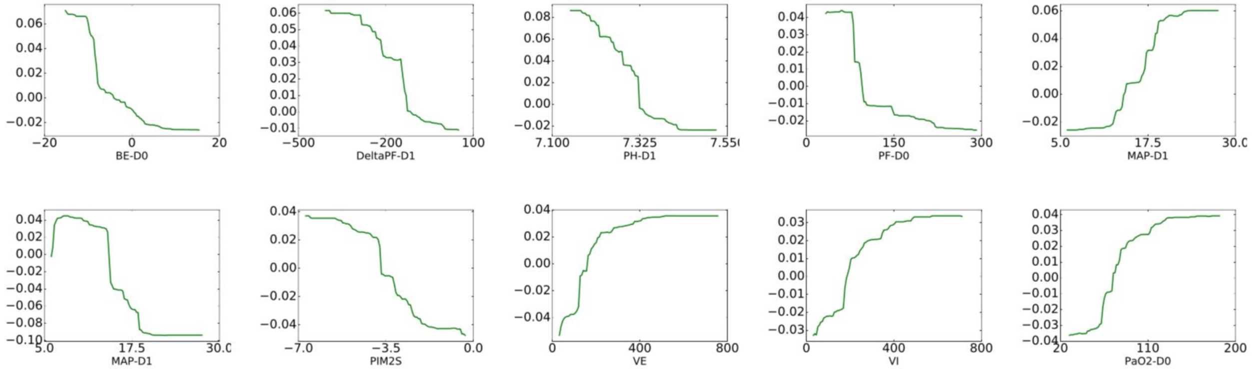


Figure 6: One-way partial dependence plots of the top features from GBTmimic for MOR (top) and VFD (bottom) tasks. x-axis: variable value; y-axis: dependence value.

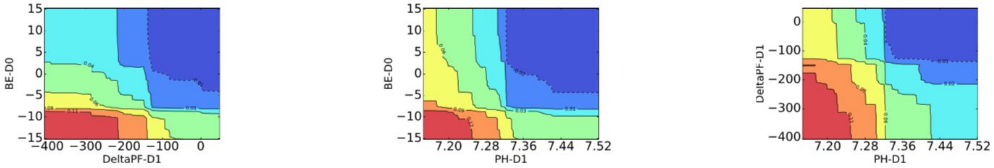


Figure 7: Pairwise partial dependence plots of the top features from GBTmimic for MOR (top) and VFD (bottom) tasks. Red: positive dependence; Blue: negative dependence.

Dataset

- Pediatric ICU dataset collected from Children's Hospital LA
    - Consists of 398 unique patients with acute lung injury
    - 27 static features, e.g.
        - Demographic information
        - Preliminary admission findings
    - 21 temporal features recorded over first 4 days (0 - 3) on a mechanical ventilator, e.g.
        - pH levels
        - Change in PaO2/FIO2 (PF) ratio
- Missing features filled via naive imputation

Experimental Design

- Dataset used for two binary classification tasks:
    1. Mortality (MOR)
    2. Ventilator Free Days (VFD)
- Experimental learning tasks:
    1. Baselines
    2. Deep neural networks
    3. Mimic learning models
- Each model/experiment run with 5 randomized trials with 5-fold CV

Results and Findings

- Deep models outperform all baseline models
- Best deep model:
    - Combination neural net with both standard non-linearity cells for static features and GRU for temporal features
- Mimic learning model attained comparable performance

Methods		MOR (Mortality)		VFD (Ventilator Free Days)	
		AUROC	AUPRC	AUROC	AUPRC
Baselines	SVM	0.6437 ± 0.024	0.3408 ± 0.034	0.7251 ± 0.023	0.7901 ± 0.019
	LR	0.6915 ± 0.027	0.3736 ± 0.038	0.7592 ± 0.021	0.8142 ± 0.019
	DT	0.6024 ± 0.013	0.4369 ± 0.016	0.5794 ± 0.022	0.7570 ± 0.012
	GBT	0.7196 ± 0.023	0.4171 ± 0.040	0.7528 ± 0.017	0.8037 ± 0.018
Deep Models	DNN	0.7266 ± 0.089	0.4117 ± 0.122	0.7752 ± 0.054	0.8341 ± 0.042
	GRU	0.7666 ± 0.063	0.4587 ± 0.104	0.7723 ± 0.053	0.8131 ± 0.058
	DNN + GRU	0.7813 ± 0.028	$\mathbf{0.4874 \pm 0.051}$	$\mathbf{0.7896 \pm 0.019}$	$\mathbf{0.8397 \pm 0.018}$
Best Mimic Model		$\mathbf{0.7898 \pm 0.030}$	0.4766 ± 0.050	$\mathbf{0.7889 \pm 0.018}$	0.8324 ± 0.016

Results and Findings

- Proposed model is highly interpretable:
  - Evaluate feature influence for tree based models
    - All temporal features are most influential
    - Most important static features: PRISM (Pediatric Risk of Mortality)
  - Evaluate one-way and two-way partial dependence
  - Obtain top decision tree rules, e.g.
    - MOR: Mean airway pressure on day 1, lung injury score (LIS), etc.
- Pipeline 1 produces better results than pipeline 2

Visualizing Gradient Boosted Trees

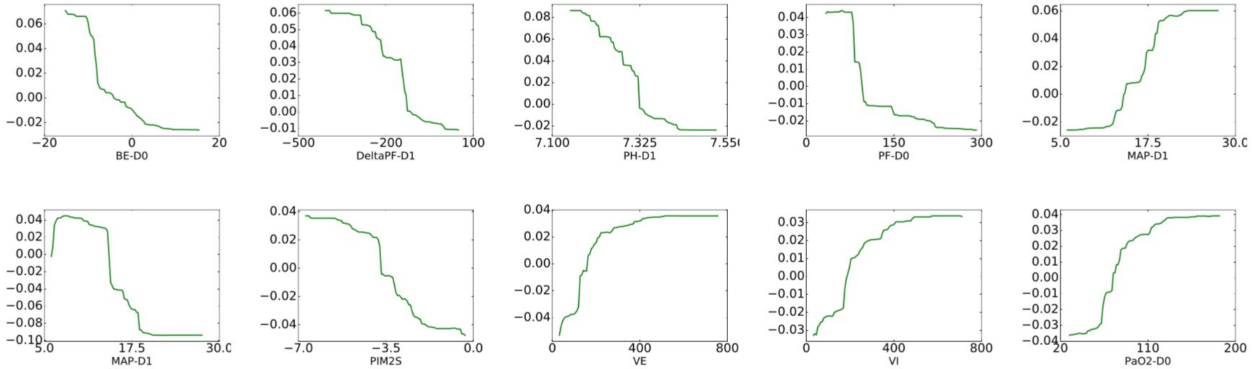


Figure 6: One-way partial dependence plots of the top features from GBTmimic for MOR (top) and VFD (bottom) tasks. x-axis: variable value; y-axis: dependence value.

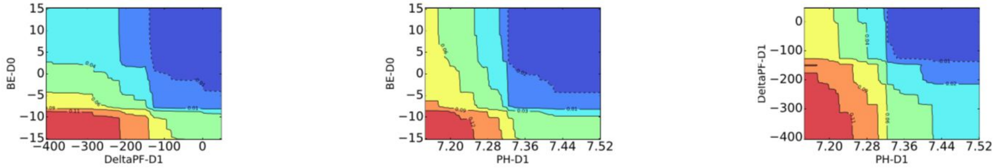


Figure 7: Pairwise partial dependence plots of the top features from GBTmimic for MOR (top) and VFD (bottom) tasks. Red: positive dependence; Blue: negative dependence.

Critique and Feedback

- Strengths:

  Achieves high performance with good interpretability

  Interpretability corresponds with empirical medical findings

  Well designed and explained mimc learning model

- Weaknesses:

  Data preprocessing - needs more sophisticated imputation methods

  Needs better range of temporal features

  Lacks thorough analysis of model's interpretability