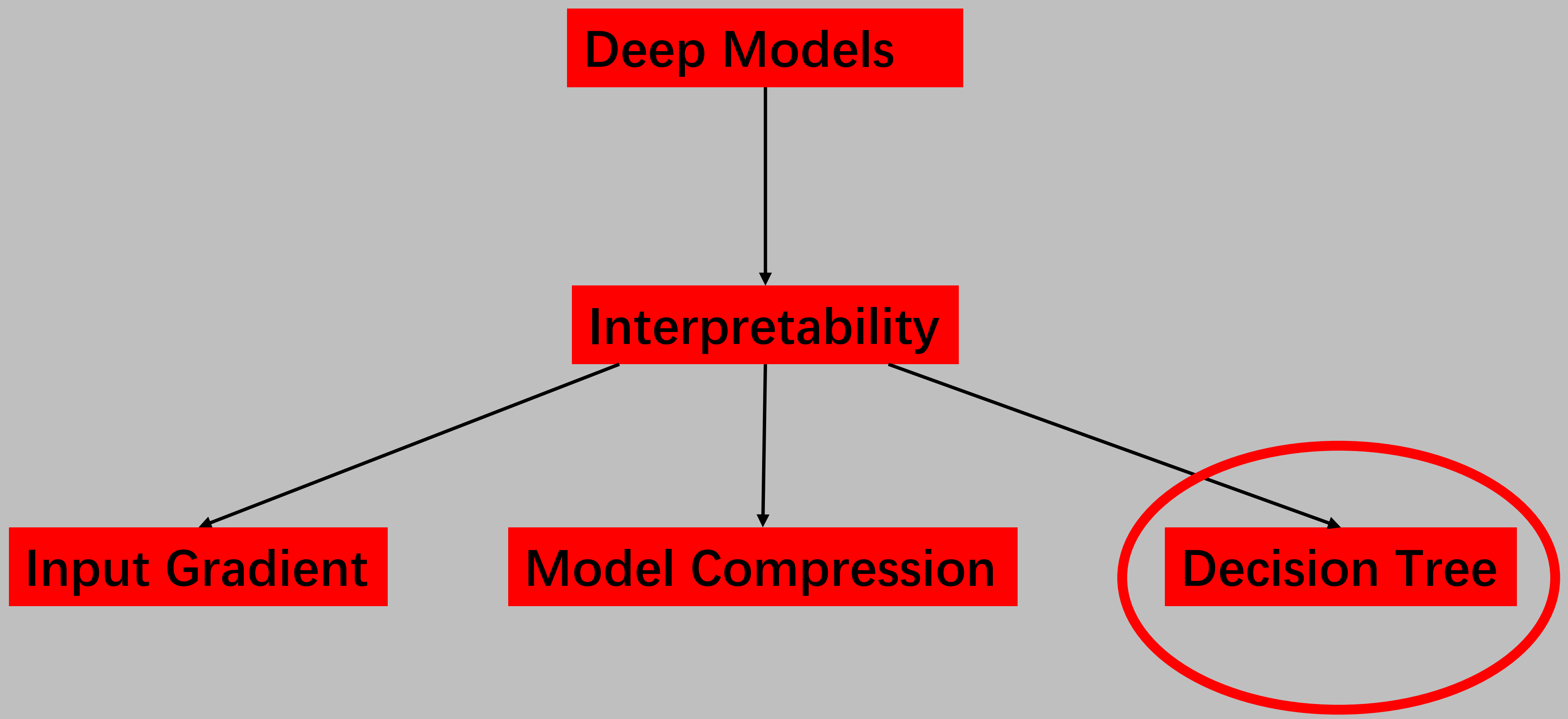

Deep Models
Interpretability
Input Gradient
Model Compression
Decision Tree

Beyond Sparsity:
Tree Regularization of Deep Models for Interpretability

MikeWu, Michael C.Hughes, Sonali Parbhoo, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez

Standford &&Harvard &&…

AAAI 2018

Outline

- Motivation
- Background
- Model
- Experiment
- Discussion

Motivation

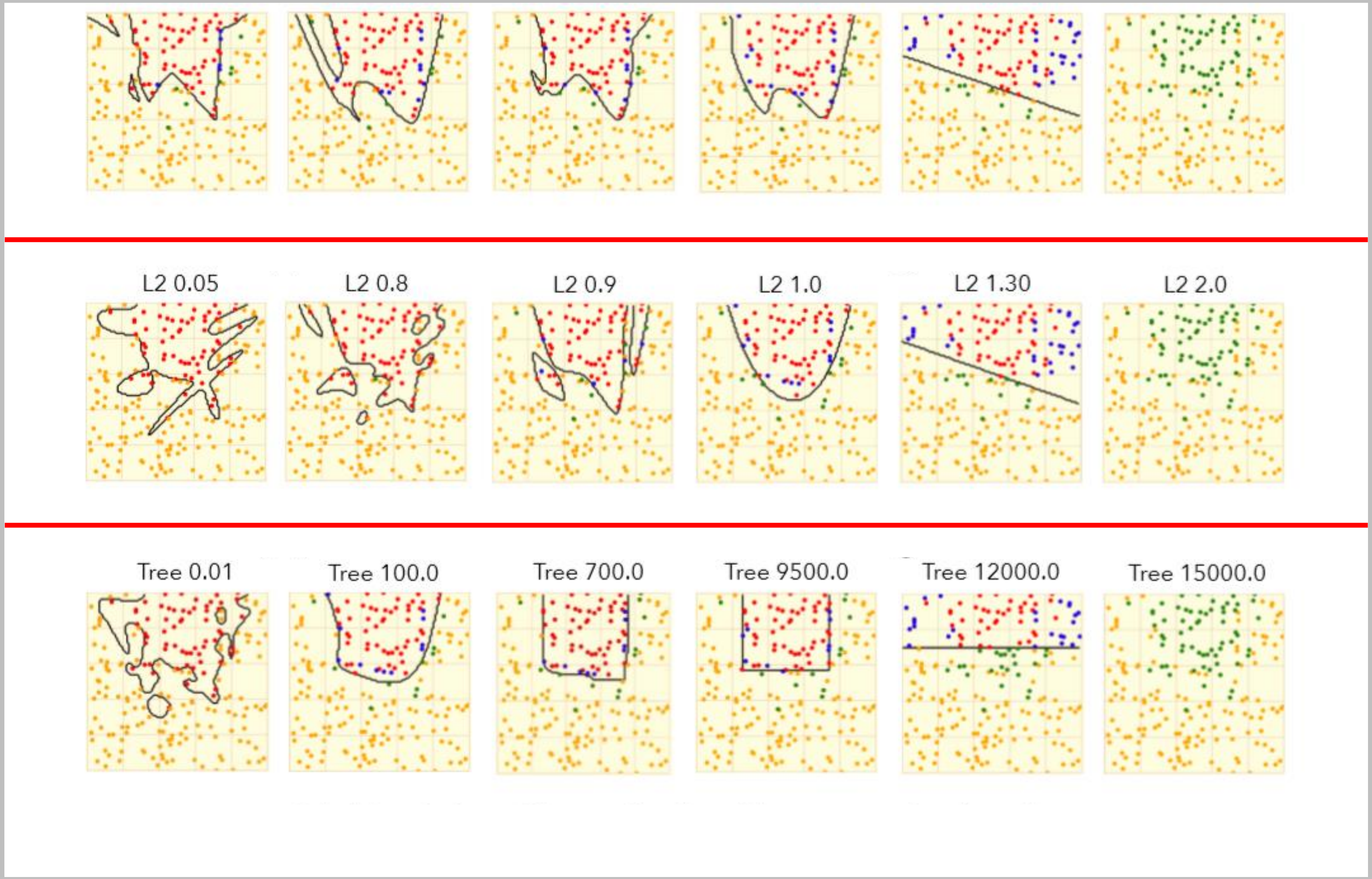
- ✓ Deep models are **difficult to interpret** → **interpretability**
- ✓ Deep models often have **multiple optima** of similar predictive accuracy → a **more interpretable** one
- ✓ Want a **human-simulatable** model → **small decision tree**

User can take in input data together with the parameters of the model to produce a prediction

Motivation

multiple optima

$$y = 5 * (x - 0.5)^2 + 0.4$$



Background

We can train deep models via the following **loss minimization** objective:

$$\min_{W} \lambda \boxed{\Psi(W)} + \sum_{n=1}^{N} \sum_{n=1}^{T_n} \text{loss}(y_{nt}, \hat{y}_{nt}(x_n, W))$$

L0 L1 L2

Background

- L0: 非零元素的个数 → 参数矩阵是稀疏的
- L1:向量各个元素的绝对值之和
- L2:向量各元素的平方和然后求平方根
- L???

Feature selection

Interpretability

Model

Tree Regularization for Deep Models:

$$\min_{W} \lambda \Psi(W) + \sum_{n=1}^{N} \sum_{n=1}^{T_n} \text{loss}(y_{nt}, \hat{y}_{nt}(x_n, W))$$

→ Want a small decision tree

→ Measure the complexity of this decision tree

→ The average decision path length

→ We use the DecisionTree model distributed in Python’s scikit-learn

Background

DecisionTree in Python's scikit-learn:

CART Tree:

For the whole dataset D, each feature A and all the K categories:

$$Gini(D, A) = \frac{D1}{D} Gini(D1) + \frac{D2}{D} Gini(D2)$$

$$Gini(D) = 1 - \sum_{k=1}^{K} \left(\frac{C_k}{D}\right)^2$$

Model

$$\min_{W} \lambda \Psi(W) + \sum_{n=1}^{N} \sum_{n=1}^{T_n} \mathrm{loss}(y_{nt}, \hat{y}_{nt}(x_n, W))$$

Algorithm 1 Average-Path-Length Cost Function

Require:

$\hat{y}(\cdot, W)$: binary prediction function, with parameters W

$D = \{x_n\}_{n=1}^{N}$: reference dataset with N examples

1: **function** $\Omega(W)$

2: tree $\leftarrow$ TRAINTREE($\{x_n, \hat{y}(x_n, W)\}$)

3: **return** $\frac{1}{N} \sum_n$ PATHLENGTH(tree, x_n)

Model

Making the Decision-Tree Loss **Differentiable:**

A **surrogate regulation function** $\Omega(W)$ map each W to the average-path-length

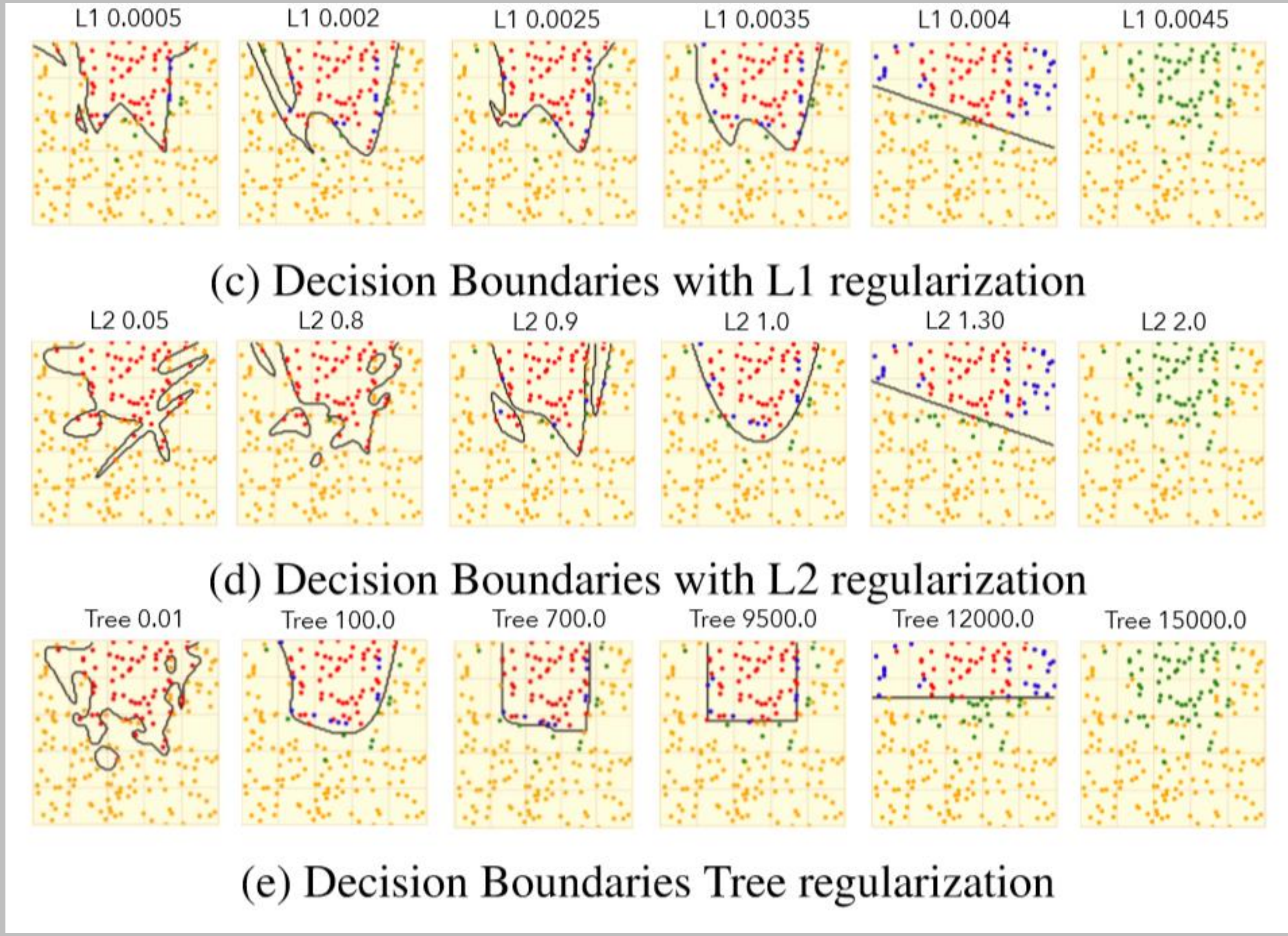
Multi-layer perceptron network → differentiable

$$\min_{\xi} \sum_{j=1}^{J} (\Omega(W_j) - \hat{\Omega}(W_j, \xi))^2 + \epsilon ||\xi||_2^2$$

Experiment 1

Tree-Regularized MLPs: A Demonstration

$$y = 5 * (x - 0.5)^2 + 0.4.$$



(c) Decision Boundaries with L1 regularization

(d) Decision Boundaries with L2 regularization

(e) Decision Boundaries Tree regularization

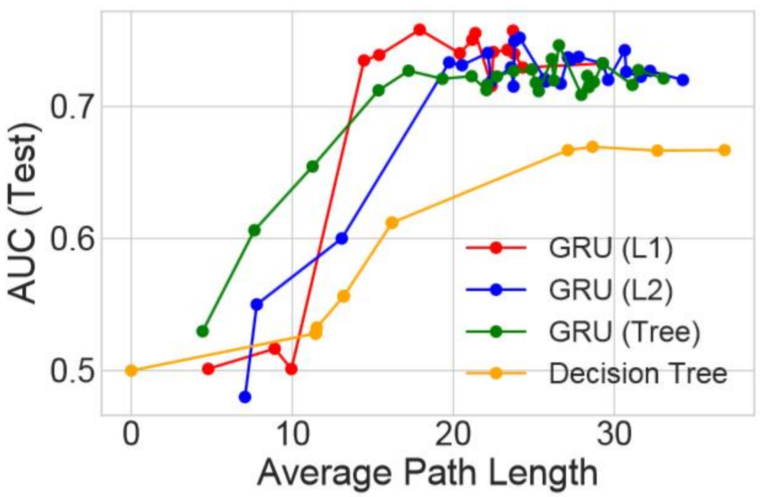
Experiment 2

Tree-Regularized Time-Series Models

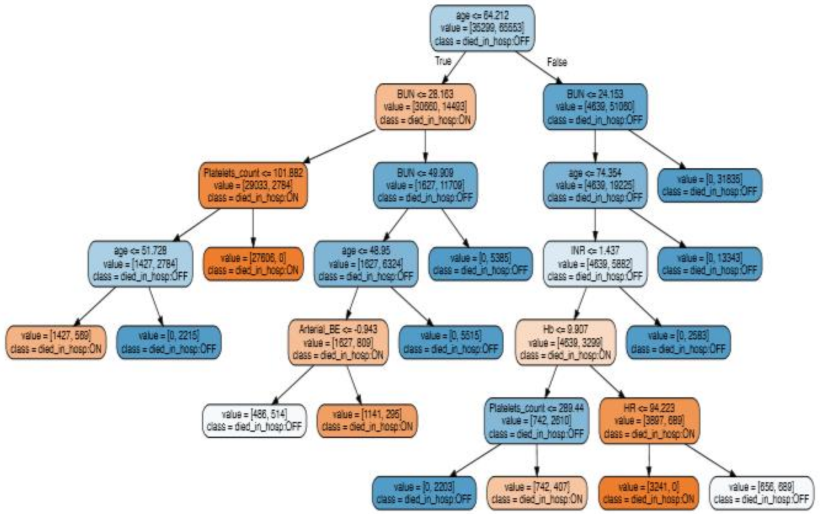
Sepsis Critical Care: input: **35** vital signs → output: **5** binary outcomes
HIV Therapy Outcome (HIV): **40** features(including blood counts, viral load measurements) → output: 15 binary labels
Phonetic Speech (TIMIT): input: 26 continuous features → output : stop phonemes or non-stops

Experiment 2

Sepsis Critical Care



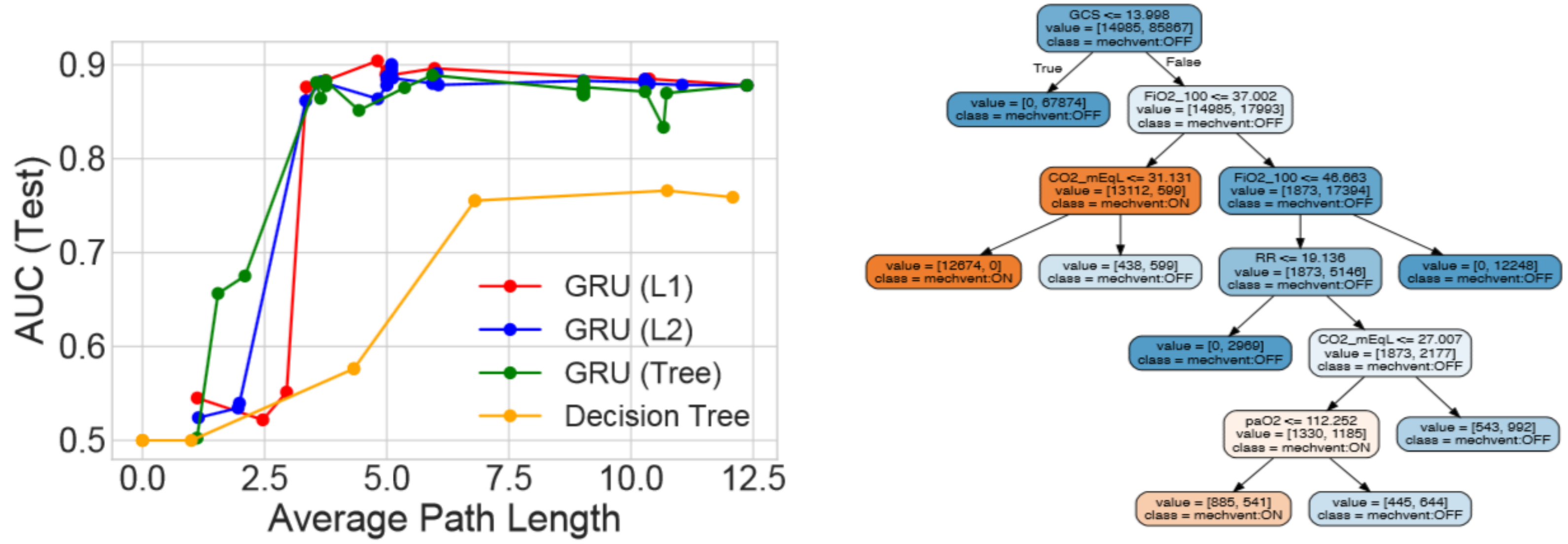
(a) In-Hospital Mortality



(b) In-Hospital Mortality

Experiment 2

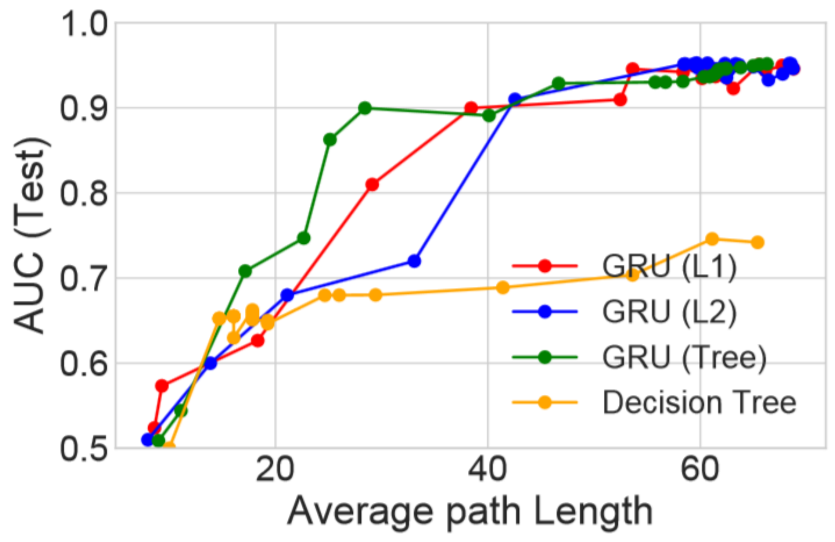
Sepsis Critical Care



(c) Mechanical Ventilation (d) Mechanical Ventilation

Experiment 2

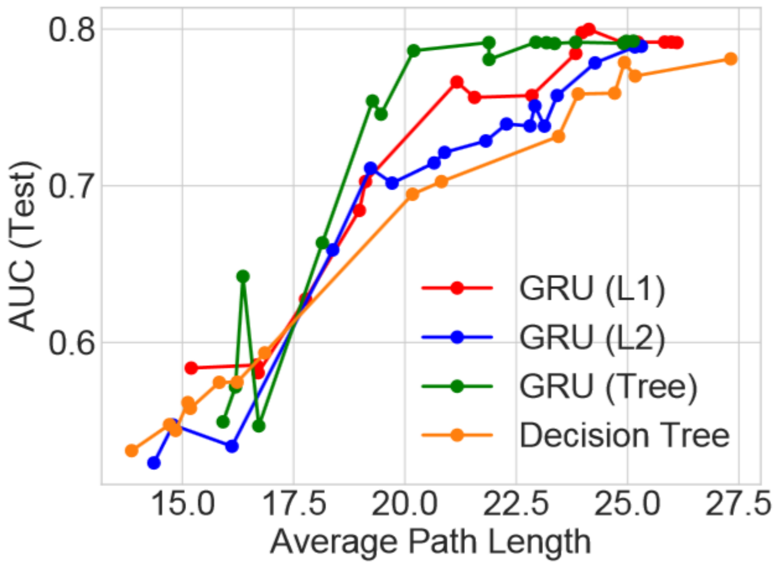
TIMIT Stop Phonemes



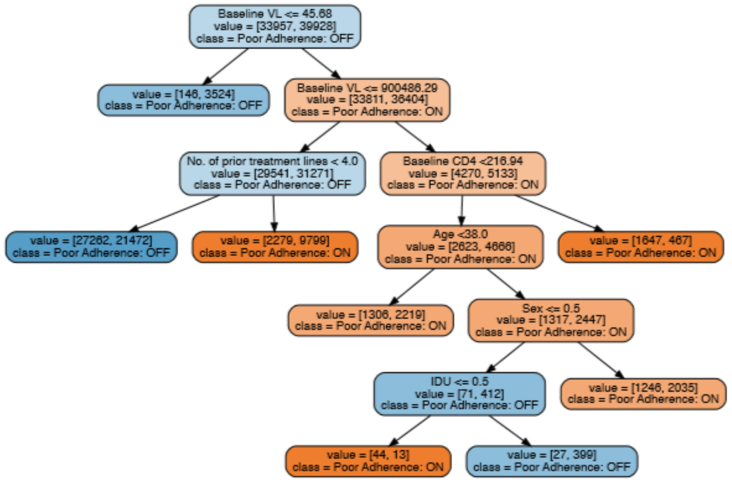
(a) TIMIT Stop Phonemes

Experiment 2

HIV Therapy Adherence



(c) HIV Therapy Adherence



(d) HIV Therapy Adherence

Discussion

The **limitations** of this kind of small decision tree ?

Thanks