What is Interpretable? Using ML to Design Interpretable Decision-Support Systems

Owen Lahav[1], Nicholas Mastronarde[2], and Mihaela van der Schaar[1,3,4]

[1]University of Oxford, [2]University at Buffalo, [3]University of California Los Angeles (UCLA), [4]Alan Turing Institute

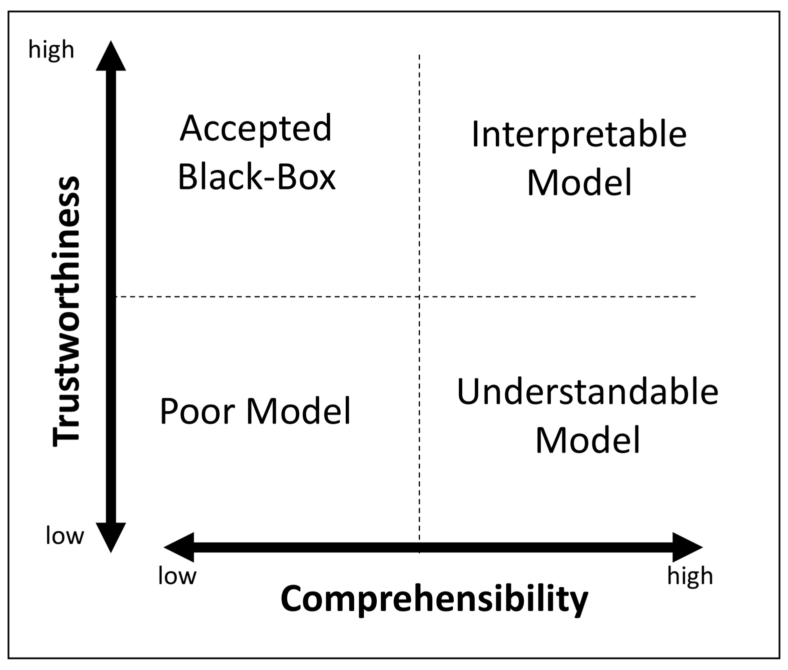
Introduction

The need for interpretability

- **Machine learning models** can accurately predict medical outcomes
- However, clinicians cannot professionally or ethically utilize **black-box** models without understanding and trusting them
- As a result, we need **interpretability**

Intrepretability in clinical settings

- ML interpretability has focused on user comprehension - **interpretability modules** presented with the ML model's outputs
- However, comprehensibility is insufficient
- Clinicians must also **trust** models before they can use them



Solution: Ask doctors!

- Use **reinforcement learning** to design comprehensible, trustworthy systems
- Present supplementary information to **clinicians**, and learn from their responses

Decision-Support System

MAGGIC data-set

- 30,389 heart-failure (HF) patients
- 31 features: patient characteristics, symptoms, medications, etc.
- Average 1-year mortality rate of 18.8%

Machine Learning Model

- Predict 1-year mortality risk after HF
- Simple **Deep Neural Network (DNN)** with 2 layers of 100 and 20 nodes
  - Outperforms MAGGIC Risk Score used by clinicians

Model	AUC-ROC	AUC-PR
Linear Regression	0.573 ± 0.0078	0.250 ± 0.0023
Random Forest	0.731 ± 0.0046	0.328 ± 0.0105
Gradient Boosting Machine	0.710 ± 0.0031	0.373 ± 0.0116
XGBoost	0.711 ± 0.0041	0.371 ± 0.1110
Neural Network	0.725 ± 0.0054	0.376 ± 0.0060
MAGGIC Risk Score	0.693 ± 0.0071	0.324 ± 0.0121

Model Evidence

- Collated a large set of possible evidence to present to users
  - **Model Details:** data set, training, accuracy, DNN approximation methods
  - **Interpretability Modules:** linear approximations, local decision-tree, feature sensitivity
- Consulted medical experts to reduce evidence space and inform design

Reinforcement Learning Model

- Multi-armed bandit using UCB1 algorithm
  - Arms = evidence sequences
- **Any RL method** could be utilized for identifying optimal sequence of evidence

Experimental Design

We designed a **RL-based clinical decision-support system (DSS)** around the neural network model, in the form of an online survey.

Below are screenshots showing some of the model evidence presented (counter-clockwise): A patient scenario, local linear model, local decision-tree model, and a feature sensitivity sample.

- **Patient Betty** is an 86-year-old non-Caucasian female suffering from heart failure
- Betty has a BMI of 21.6
- Betty exhibits rales and shortness of breath at rest
- Our model predicts the probability of Betty dying within 1 year is **83.5%**

Examine how the following features might impact Betty's risk score, based on our neural network model:

New York Risk Association Score: IV

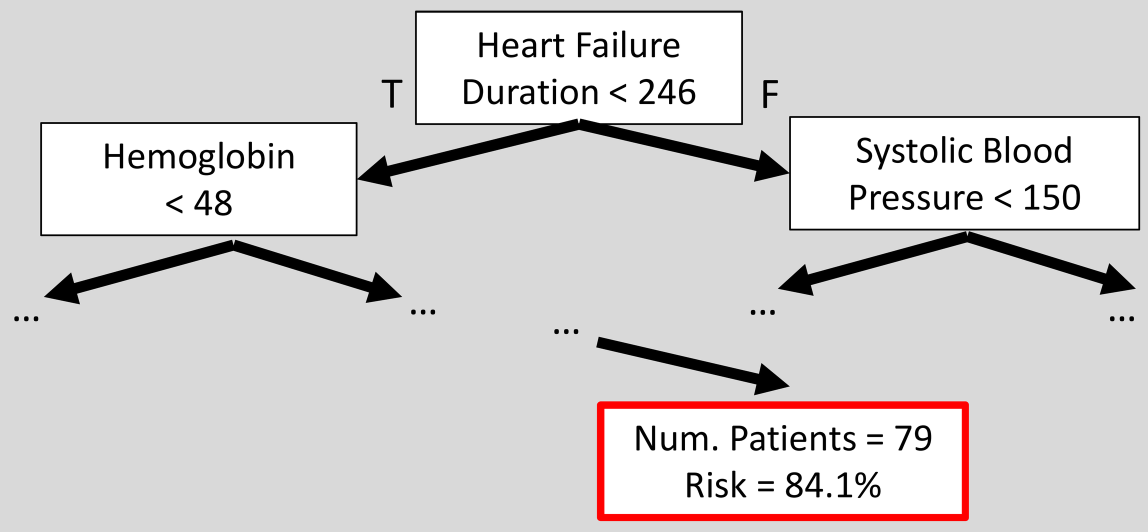
Prescription:
No prescription
ACE Inhibitors
Beta Blockers

Predicted Risk Score:
77.2%

Local linear approximation for 80-100% risk strata:
Significant coefficients:

Patient Characteristic	80-100% Risk
New York Heart Association Score	0.548
ACE Inhibitors or ARB perscribed	-0.452
Beta Blockers perscribed	-0.263
Shortness of breath at rest	0.248
Ethnicity (Caucasian)	-0.241
Rales	0.211
Diabetes	0.138
Gender	0.057
Age	0.051

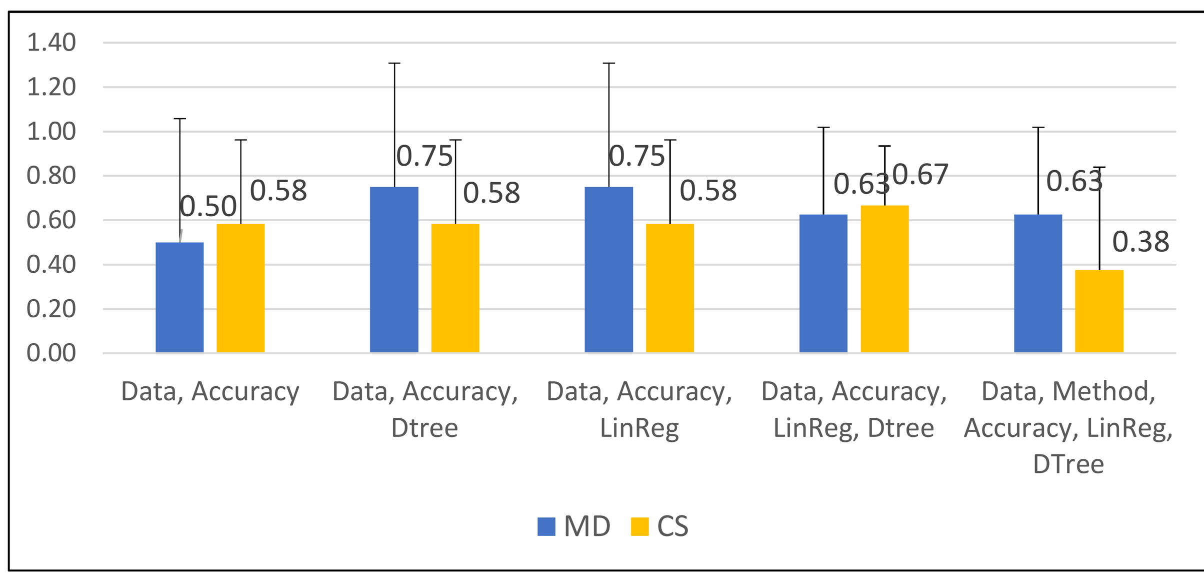
Local decision-tree approximation for 80-100% risk strata:



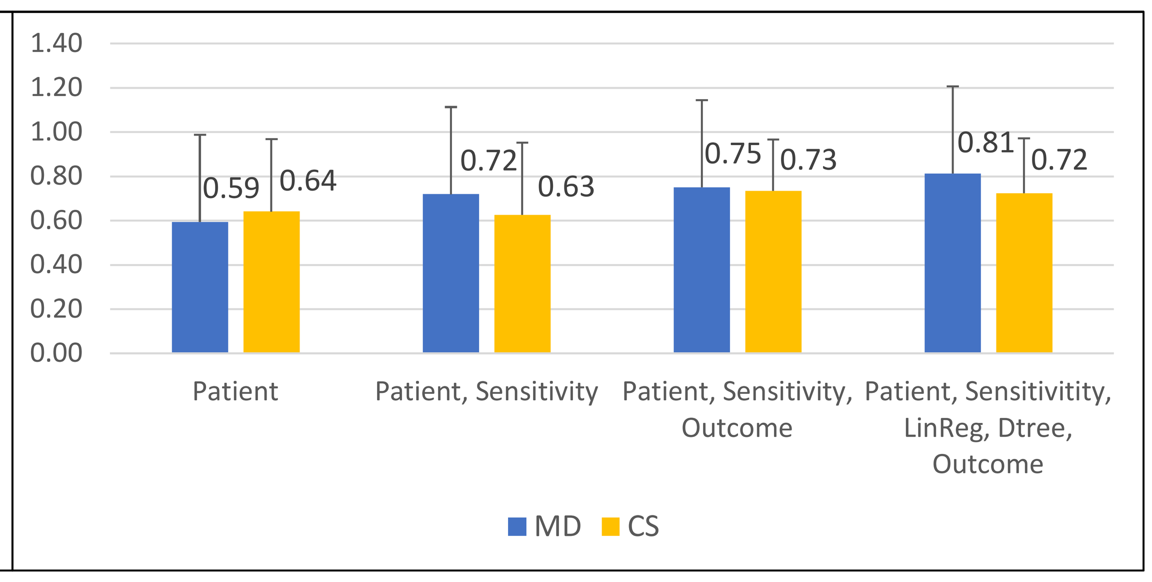
Main Results

- We surveyed **14 doctors** who rated their confidence in the model based on evidence shown
- We also surveyed **30 ML experts** who predicted the average doctor's confidence in the model

The average ratings provided by doctors and ML experts for each evidence sequence are below:



(a) General Model Evidence Sequences



(b) Patient Scenario Sequences

Key Findings

- Machine learning experts appear **unable** to predict which interpretability modules will best engender **doctor trust**
- Evidence is not super-additive: more information may not increase confidence, possibly due to **information overload**
- Doctors must be consulted to create ML-driven DSSs that are truly useful in healthcare settings

Take our Survey!

Contact the research team for details!

Contact:
owen.lahav@gtc.ox.ac.uk

Future Work

Our proposed framework utilizing **reinforcement learning** to design **comprehensible, trustworthy** systems based on ML models can be extended:

- Test different ML models, data-sets, or contexts in medicine and beyond
- Test the effectiveness of a wide variety of **interpretability modules**, including LIME, DeepLIFT, associative classifiers, feature rankings, and more
- Test different **RL algorithms**, including contextual bandits and deep RL

Next step: improved, larger scale survey

- Fewer arms, more doctors + ML experts
  - Statistically significant results
- Contextualize clinicians by specialization, years in practice, familiarity with ML, etc.

*This study was reviewed by the Ethics Committee of the University of Oxford's Department of Computer Science, 2018