**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 insufficent
- 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 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.

**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:

**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

**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

**TAKE OUR SURVEY!**

Contact the research team for details!

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

Contact the research team for details!