Elucidating Discrepancy in Explanations of Predictive Models Developed using EMR

Aida Brankovic| 9 July 2023

Aida Brankovic Wenjie Huang David Cook

Sankalp Khanna Konstanty Bialkowski

(Co)Author(s)

Australian e-Health Research Centre

Australia's National Science Agency
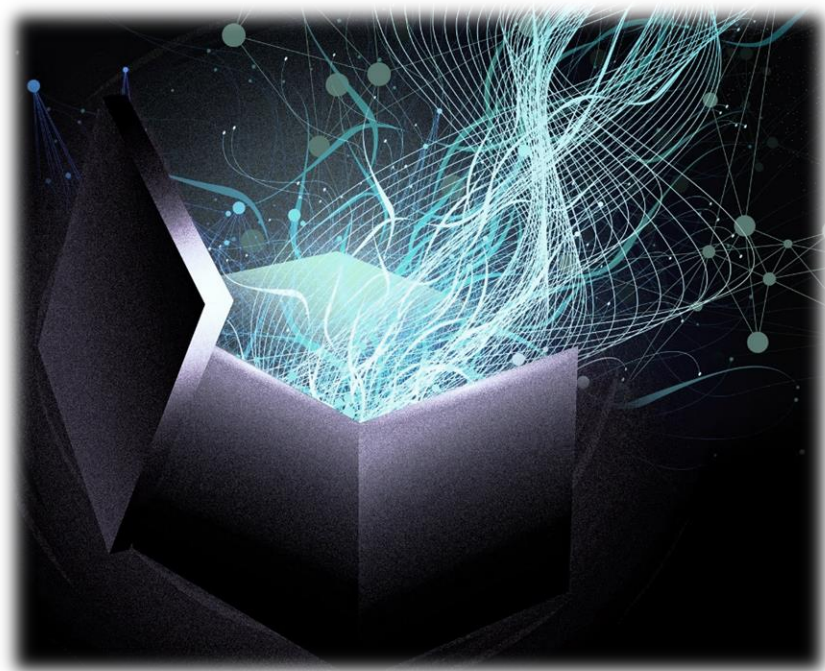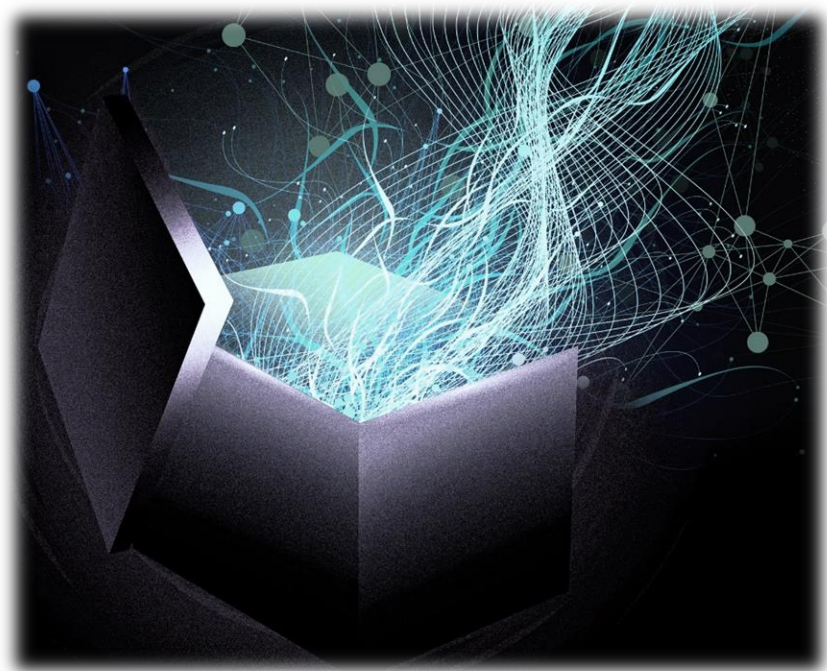
Motivation

Explainable AI

- Gradient-based (SmoothGrad, Integrated Gradients, Deep Taylor Decomposition (DTD))

- Layer-wise propagation

- Perturbation-based (LIME, Shap)

Figure taken from https://news.mit.edu/2022/machine-learning-explainability-0505

Australian e-Health
Research Centre

Explainable AI

- Gradient-based (SmoothGrad, Integrated Gradients, Deep Taylor Decomposition (DTD))

- Layer-wise propagation

- Perturbation-based (LIME, Shap)

Australian e-Health
Research Centre

Methods: Data

Readmission within next 30 days (RA30)

Identifying patients at risk of unplanned re-hospitalisation using statewide electronic health records

Vital signs (VS)

Explainable machine learning for real-time deterioration alert prediction to guide pre-emptive treatment

Data used: 1 Jan 2015 – 31 Dec 2018

Data used: 1 Jan 2016 – 31 Dec 2018
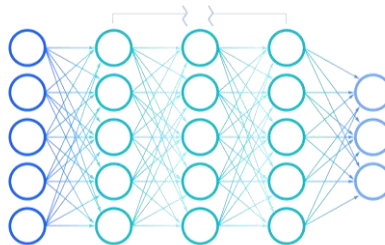
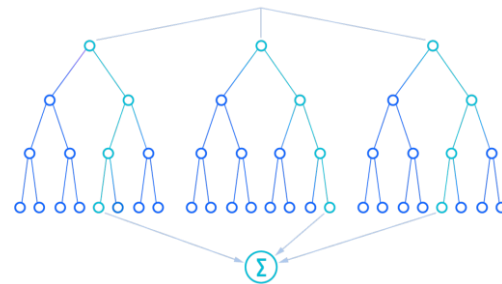# Methods: Models and Explainable AIs

Models

Logistic regression with l1
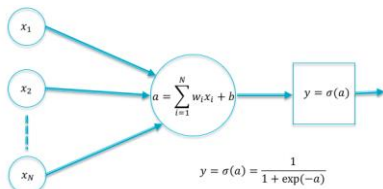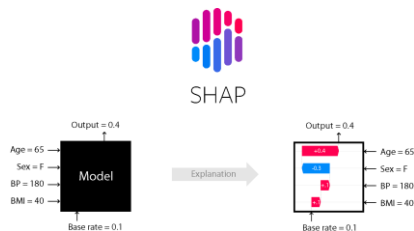
DNN with convolutional layer
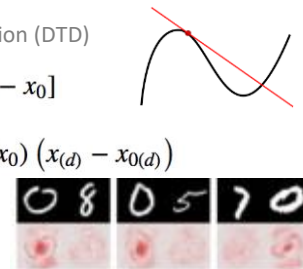
XGB

XAI

Perturbation-based

Gradient-based

SHAP

Deep Taylor Decomposition (DTD)

$$f(x) \approx f(x_0) + Df(x_0)[x - x_0]$$

$$= f(x_0) + \sum_{d=1}^{V} \frac{\partial f}{\partial x_{(d)}}(x_0)\left(x_{(d)} - x_{0(d)}\right)$$

Australian e-Health
Research Centre

CSIRO

Results

Data	Rank	Shap	L1 coeff	DTD
RA30	1	Prev. inpat. stay count	Adm. source: Boarder	Prev. inpat. stay count
	2	Patho tests count	Care type: Other	Ed presentations stay counts
	3	Elect. status: Not assigned	Elect. status: Not assigned	age
	4	LOS	Care type: Acute	Prev. inpat. stay count^2
	5	Planned the same day	Adm. source: Emergency	Elect. status: Emergency
VS	1	LOS	LOS	AVPU
	2	SpO2	SpO2	O2 Flow rate
	3	SBP	SBP count	*NMesuredEvents*
	4	*NRecordedVS*	Resp. Rate count	*NRecordedVS*
	5	min SpO2	DBP	SpO2

Table 1 Top 5 features obtained with considered explainers for VS and RA30 datasets

Australian e-Health Research Centre

Results
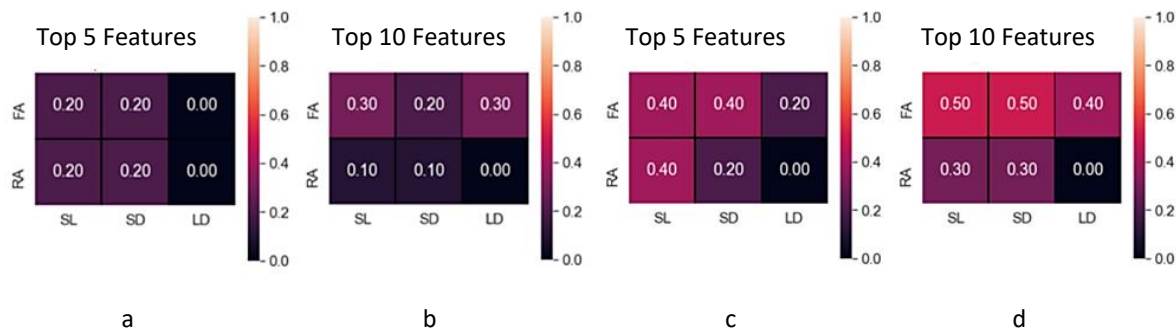
Figure 1. Feature Agreement (FA) and Rank Agreement (RA) obtained for Shap - L1 coefficient (SL), Shap-Deep Taylor Decomposition (SD) and L1 coefficient - Deep Taylor Decomposition (LD) model explanations using RA30 (a, b) and VA (c, d) datasets i.e., the outcomes of interest.

Discussion and Remarks

- Identified causes of the disagreement as the first step in establishing the criteria for trustworthiness

  - Clinical
    i. Missing features which if available, would explain more of the variation and causation of the outputs
    ii. Dependence between factors (different groups of features could have the same implications for clinicians)
    iii. Errors, missing data that could be observed or contradictory information

  - Model-related
    i. Causality might not be discovered and hence post-hoc not explainable
    ii. Different methods operate on different principles, different features might be identified as the most relevant

- Rigorous evaluation for robustness needed

Thank you

Health and Biosecurity
Dr Aida Brankovic
Research Scientist

aida.brankovic@csiro.au
aehrc.csiro.au

Australia's National Science Agency

Health Intelligence @ CSIRO AEHRC

Health System Productivity & Efficiency
Operational & Clinical Decision Support
Evidence-driven Policy & Healthcare Delivery