Interactive sequential analysis of a model improves the performance of human decision-making



Hubert Baniecki, Dariusz Parzych, Przemysław Biecek

MI^2.AI, University of Warsaw & Warsaw University of Technology, Poland

ML in PL, Warsaw, Poland

November 5, 2022

Hi!

PhD student in Computer Science
at the University of Warsaw

Working on explainable machine learning

And ML applications in biomedicine

Co-authors

- Przemyslaw Biecek
  Warsaw University of Technology
- Bastian Pfeifer
  Medical University of Graz
- Anna Saranti
  Human-Centered AI, University o…
- Andreas Holzinger
  Human-Centered AI Lab, Univer…
- Mateusz Krzyziński
  Warsaw University of Technology
- Mikołaj Spytek
  Warsaw University of Technology
- Brandon Michael Henry, M.D.
  Cincinnati Children's Hospital Me…

HOMER

MI RESEARCH

measures: faithfulness
software: Quantus
benchmarks: XAI-Bench
leaderboards: OpenXAI

(1) Evaluating model explanations is challenging.

HCI & user studies with humans

HOMER

MI RESEARCH

Explainable machine learning: from credit scoring to precision diagnostics in bio-medicine

The Washington Post
Democracy Dies in Darkness

This article was published more than **3 years ago**

BUSINESS

2019

Apple Card algorithm sparks gender bias allegations against Goldman Sachs

Entrepreneur David Heinemeier Hansson says his credit limit was 20 times that of his wife, even th
higher credit score

By Taylor Telford
November 11, 2019 at 10:44 a.m. EST

Many articles have been published in 2020 describing new machine learning-based models for [detection and prognostication of COVID-19], but it is unclear which are of potential clinical utility. [...] Our review finds that none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases.

nature machine intelligence

Explore content ⌄ About the journal ⌄ Publish with us ⌄

2021

nature › nature machine intelligence › analyses › article

Analysis | Open Access | Published: 15 March 2021

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, AIX-COVNET, James H. F. Rudd, Evis Sala & Carola-Bibiane Schönlieb

Nature Machine Intelligence **3**, 199–217 (2021) | Cite this article

74k Accesses | **237** Citations | **1159** Altmetric | Metrics

HOMER

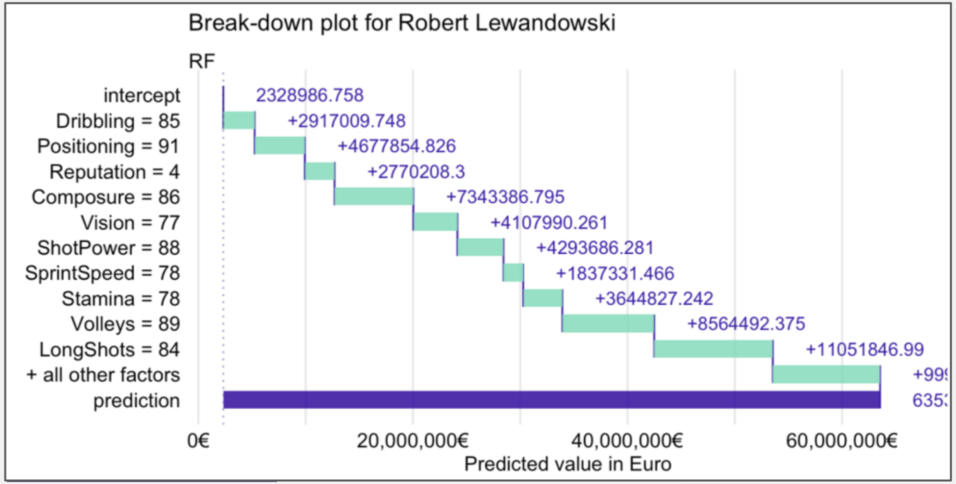
We explain black-box machine learning for...

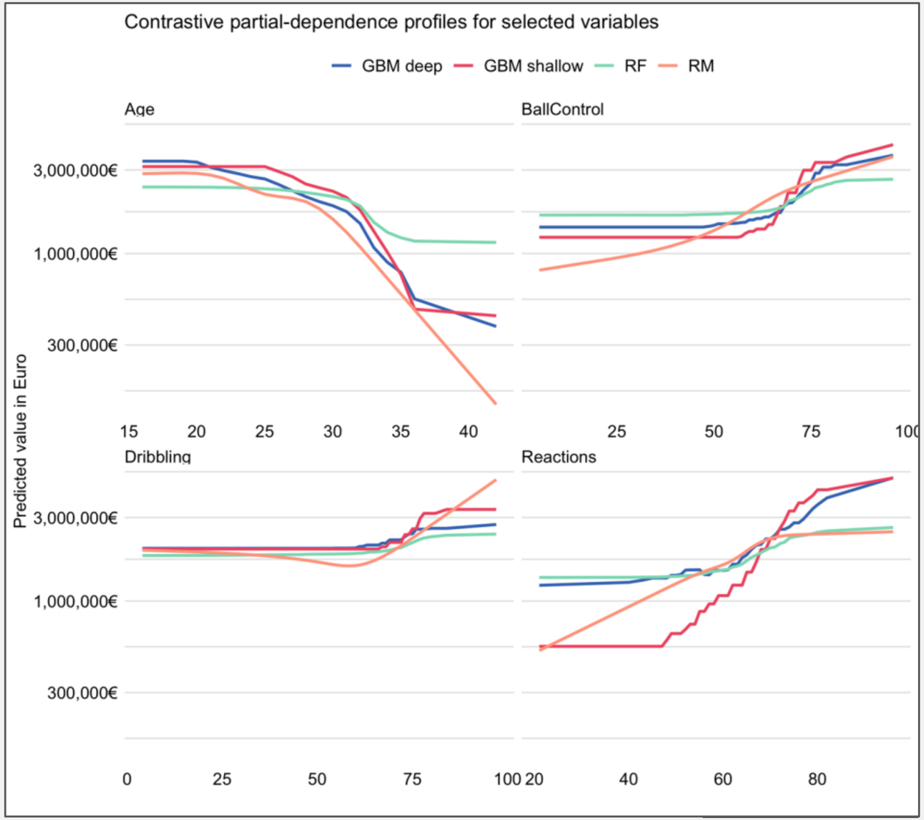
(1) Validation & debugging

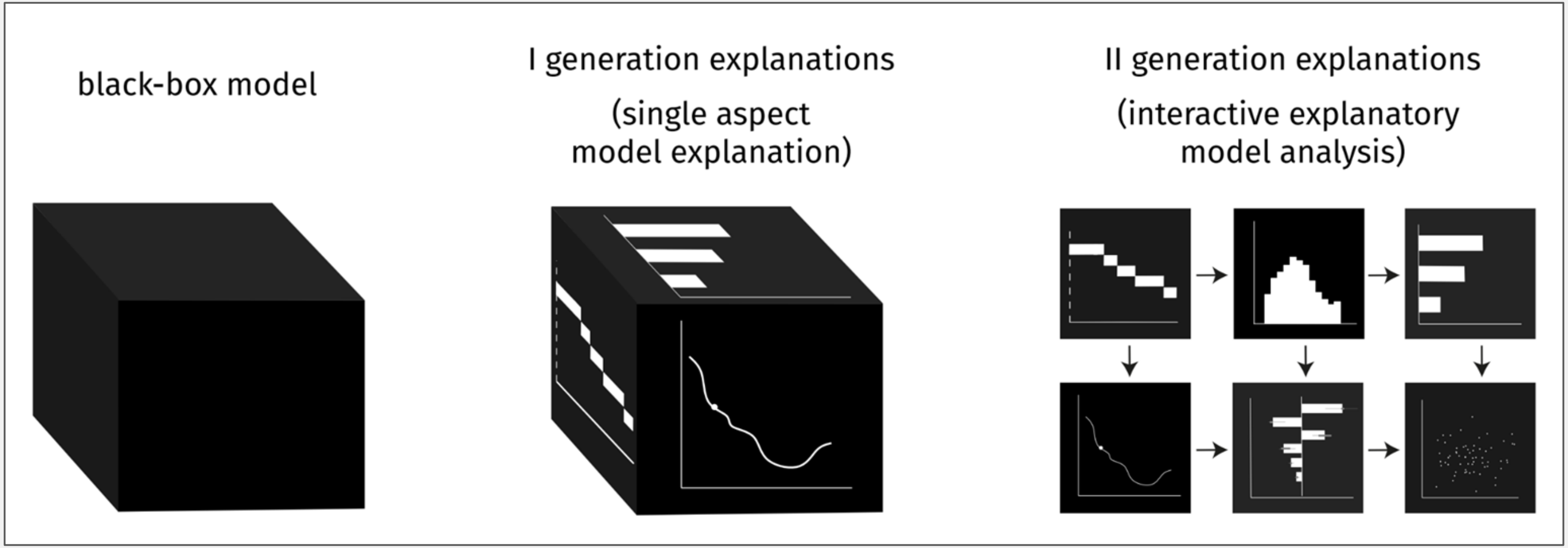
(2) Scientific insights

(3) Model improvement

ema.drwhy.ai



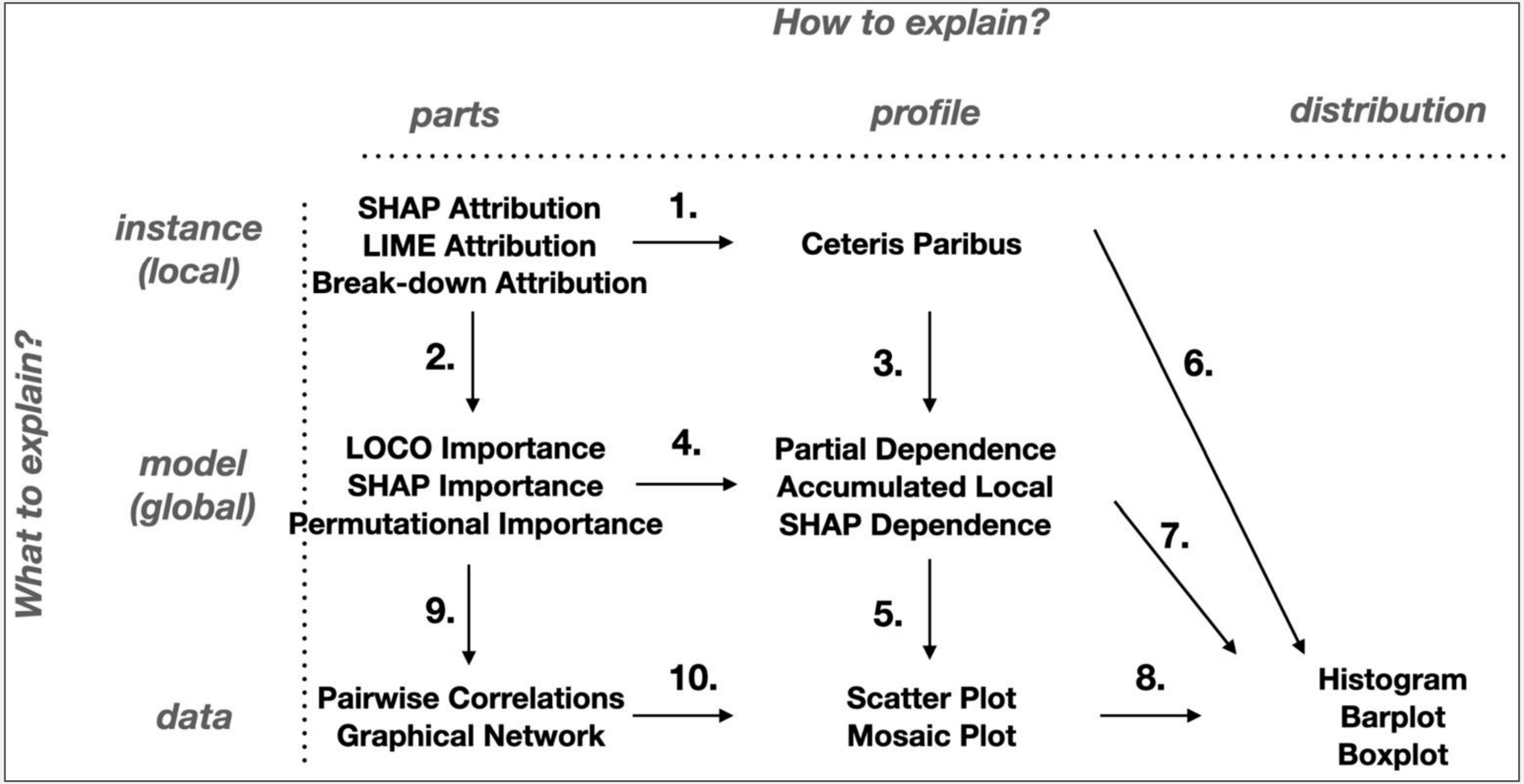


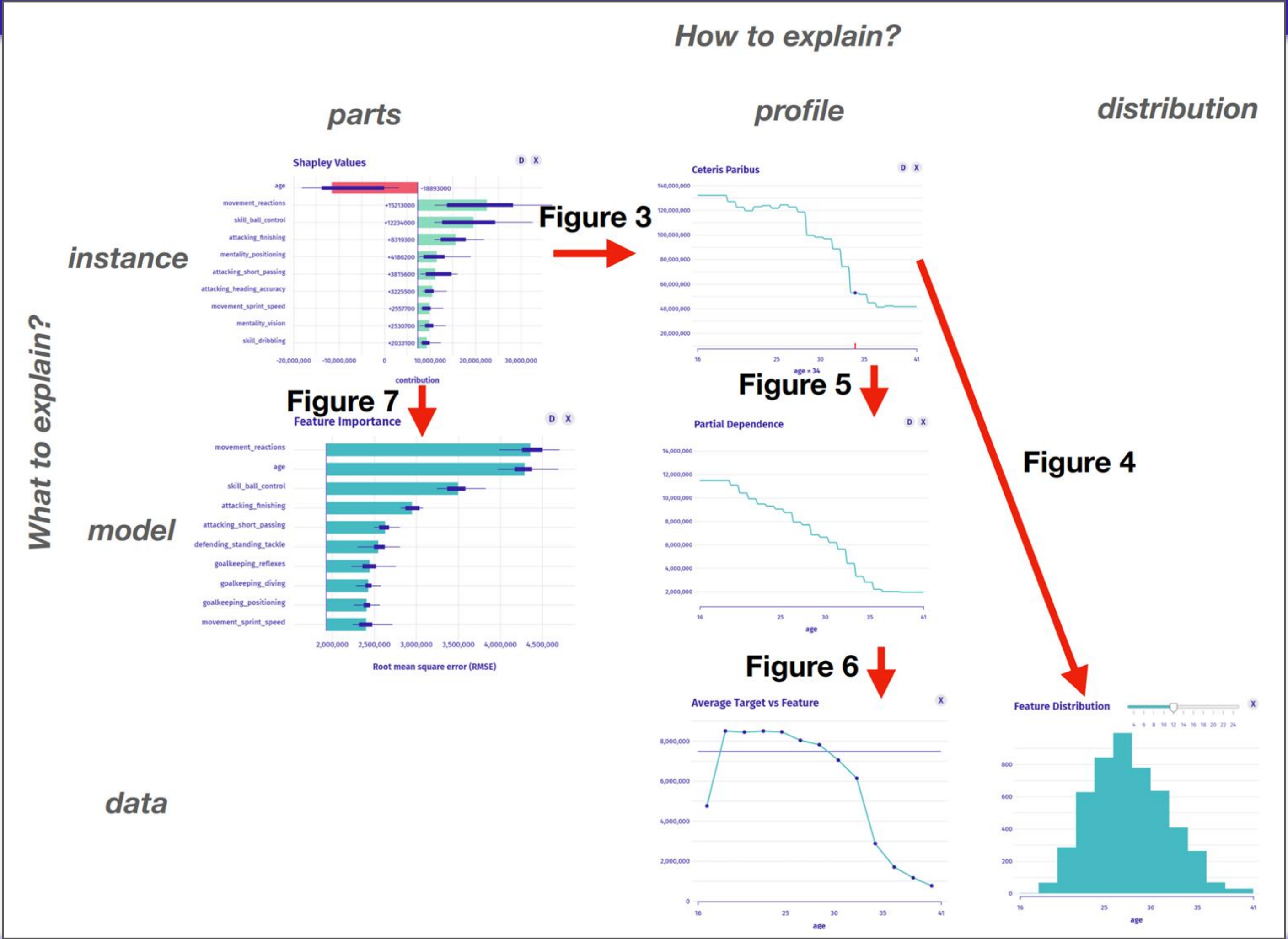


H. Baniecki, D. Parzych, P. Biecek. **The Grammar of Interactive Explanatory Model Analysis.** arXiv preprint, 2022.

HOMER

MI RESEARCH


How to explain?
parts
profile
distribution
What to explain?
instance
(local)
model
(global)
data
SHAP Attribution
LIME Attribution
Break-down Attribution
1.
Ceteris Paribus
2.
3.
6.
LOCO Importance
SHAP Importance
Permutational Importance
4.
Partial Dependence
Accumulated Local
SHAP Dependence
7.
9.
5.
Pairwise Correlations
Graphical Network
10.
Scatter Plot
Mosaic Plot
8.
Histogram
Barplot
Boxplot

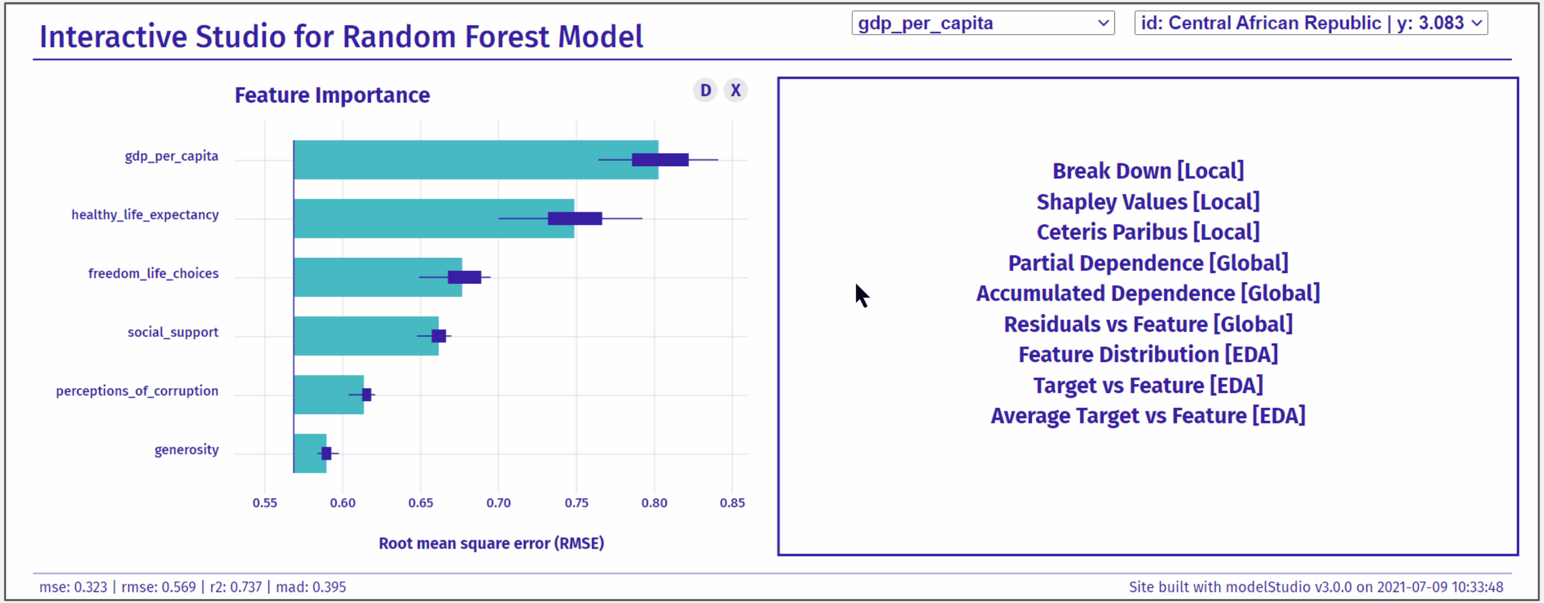


sequence

interactivity

process

MI RESEARCH



modelStudio

Star 274

CRAN 3.1.0

downloads 32K

H. Baniecki, P. Biecek. **modelStudio: Interactive Studio with Explanations for ML Predictive Models.** JOSS, 2019.

ML in PL 2019

HOMER

MI RESEARCH

User study: a 45-minute questionnaire

Goal: see if an interactive and sequential analysis of a model **brings value** to explaining black-box machine learning

Research question: Do juxtaposing complementary explanations increase the **usefulness** of explanations?

Usefulness: accuracy and confidence of human decision-making

Target group: model developers, not domain experts

HOMER

MI RESEARCH

A case study of Acute Kidney Injury (AKI) prediction

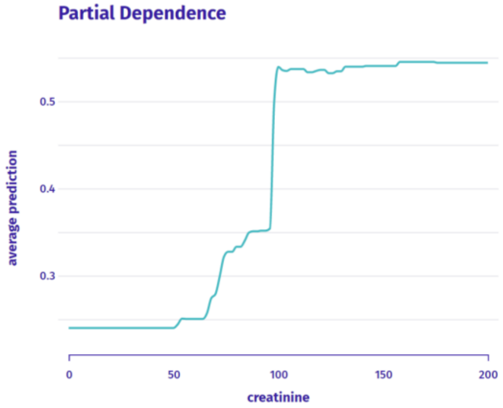
The model predicts a probability of AKI occurance during the patient's hospitalization due to COVID-19

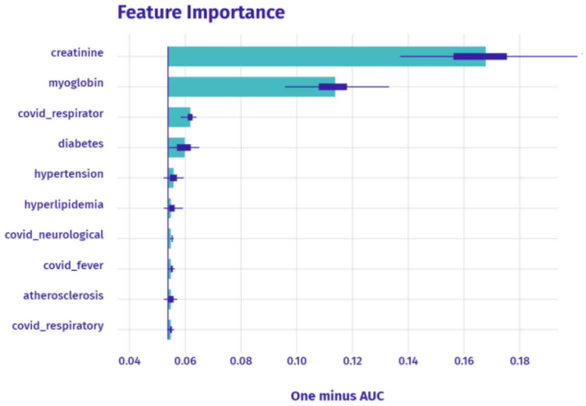
Classification performance measures:

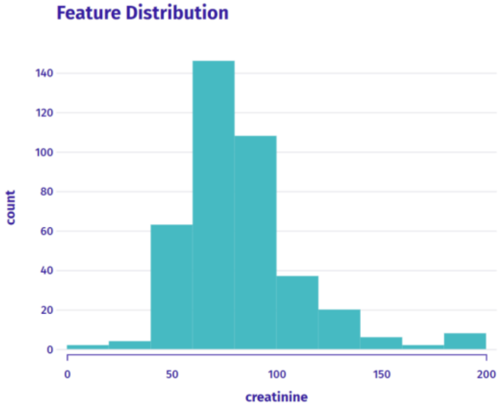
recall: 0.866 | precision: 0.644 | f1: 0.739 | accuracy: 0.896 | auc: 0.946

Classification threshold: 0.5

Frequency of AKI among patients: about 18%
(the fraction of class 1)





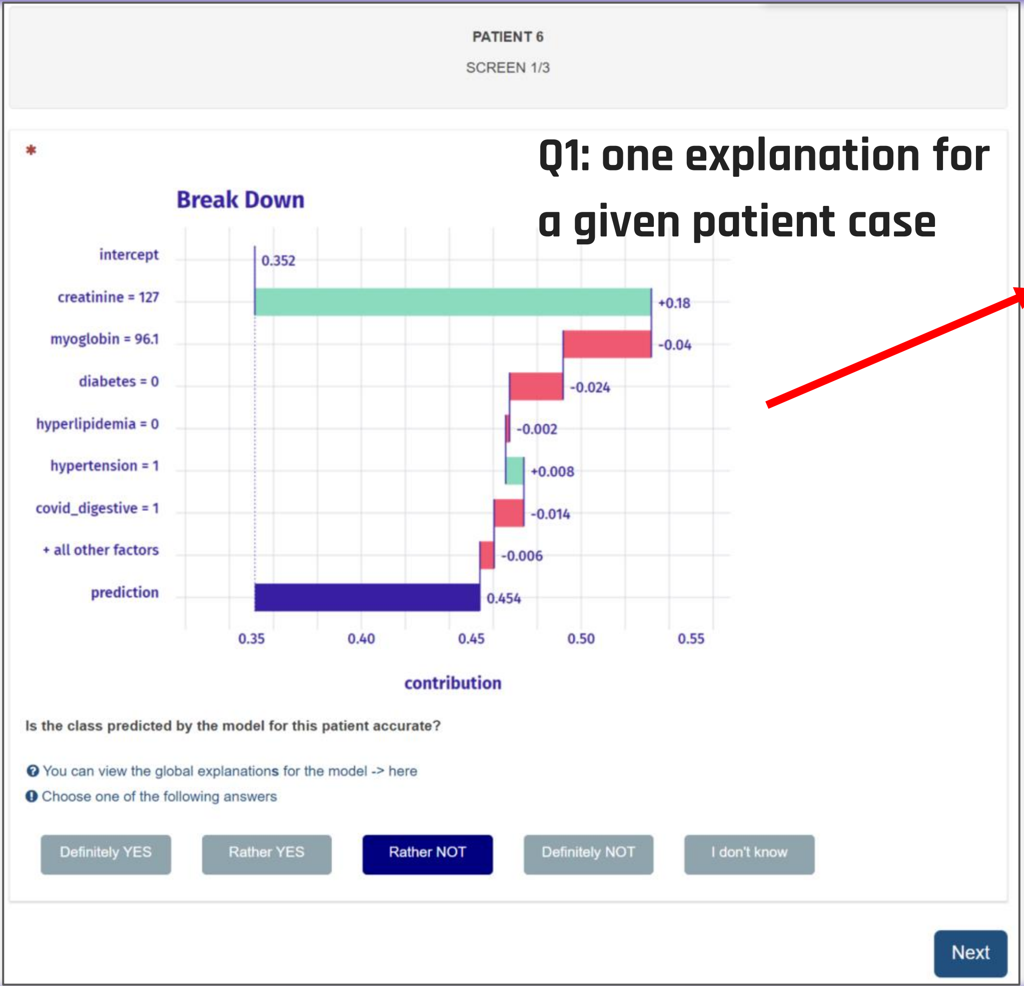


hospital

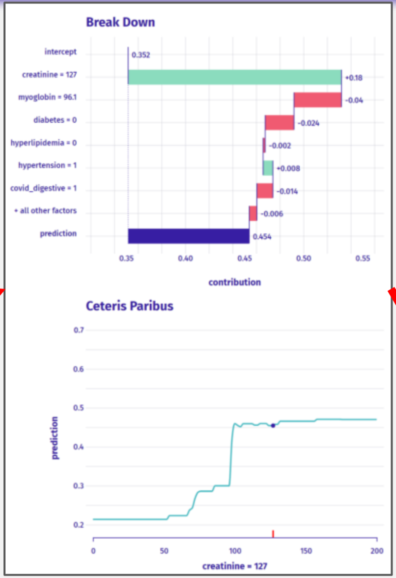
patient

decision

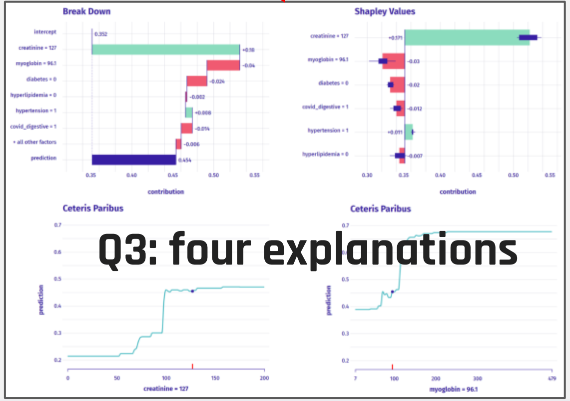
Q1: one explanation for a given patient case

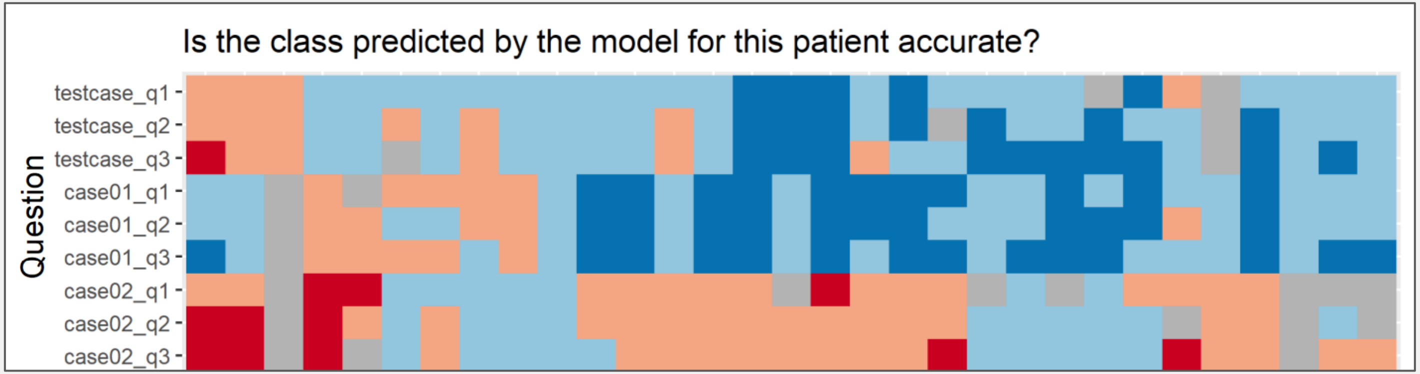


Q2: two explanations for the same patient

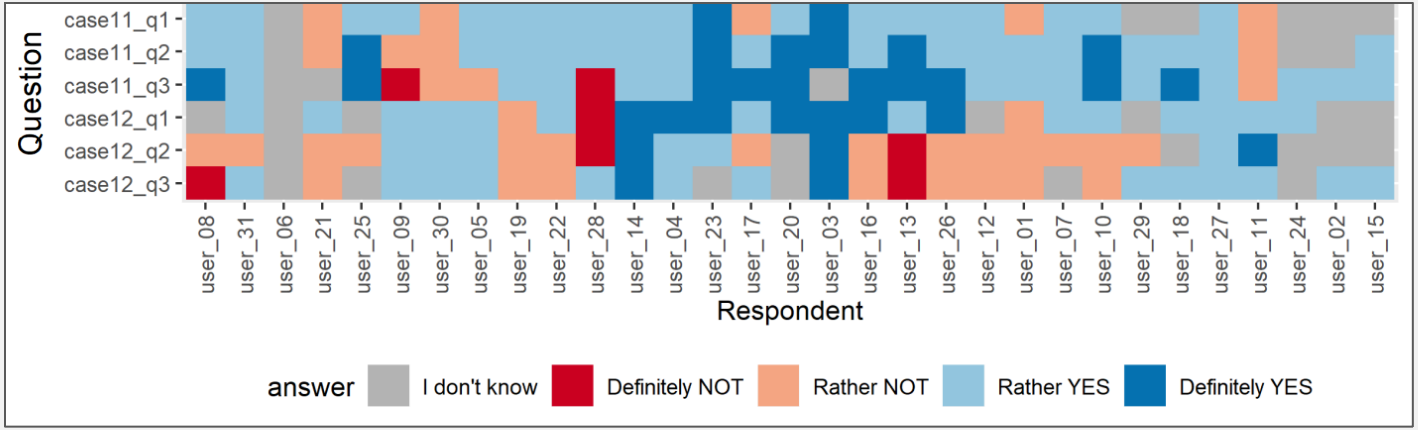


Q3: four explanations

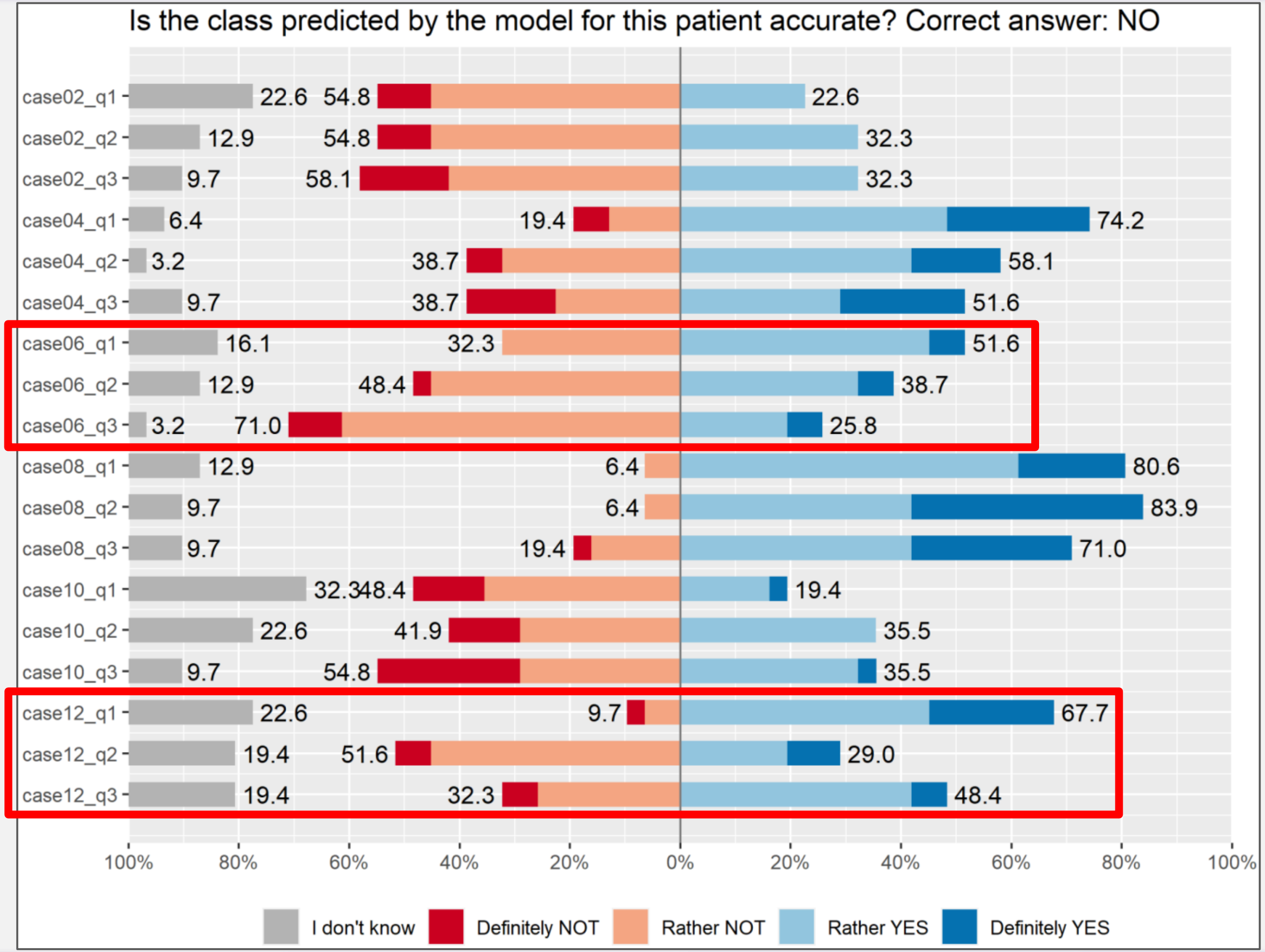




...8 more patient cases...



46 respondents -> 31 full responses



Analogous results for “Correct answer: YES”

Answers aggregated over 30 respondents

How to draw conclusions?

Accuracy: frequency of proper answers given by **30** respondents

Its variance: accuracy aggregated over **12** patient cases

Hypothesis (number of cases = 12)	Q_1	Q_3	$\Delta Q_3 Q_1$	P-values
Accuracy increases between Q_3 and Q_1	$52.2_{\pm 29.3}$	$65.8_{\pm 24.2}$	$13.6_{\pm 11.4}$	0.002; 0.004
Confidence increases between Q_3 and Q_1	$23.1_{\pm 13.7}$	$35.3_{\pm 15.6}$	$12.2_{\pm 11.8}$	0.004; 0.018
"I don't know" *decreases* between Q_3 and Q_1	$12.8_{\pm 9.8}$	$5.2_{\pm 5.0}$	$-7.5_{\pm 7.8}$	0.007; 0.007

Table 4 Aggregated results from the user study validate our hypotheses. We report $mean_{\pm sd}$ across the participants' performance in 12 patient cases, and measure their difference between Q_3 and Q_1 marked as $\Delta Q_3 Q_1$. We validate each hypothesis with the t-test and Wilcoxon signed-rank test, hence two p-values. There is a significant increase in accuracy and confidence between the sequential questions. Additionally, the frequency of ambiguous answers decreases.

Q_4: Which of the following aspects had the greatest impact on your decision making in the presented patient case?	
Answer	**Frequency**
Break-down explanation (1st screen)	16.7%
Ceteris Paribus "What-if?" explanation (2nd screen)	27.5%
Shapley Values explanation or/and an additional Ceteris Paribus "What-if?" explanation (3rd screen)	35.3%
Comparison of the local explanations with the global explanations	19.2%
My answer was random, I ran out of information to make a decision	0.5%
Other (three descriptive answers in total: a Permutational Importance explanation, both Ceteris Paribus explanations, a high residual value)	0.8%

Table 5 Frequency of answers for Q_4 averaged across 12 cases times 30 participants.

HOMER

MI RESEARCH

5.2 Qualitative analysis

At the end of the user study, we asked our participants to share their thoughts on the user study. In the first question, we asked if they saw any positive aspects of presenting a greater number of explanations to the model. This optional question was answered by 19 participants, who most often pointed to the following positive aspects: the greater number of the presented explanations, the more information they obtain (n = 18; 95%), which allows a better understanding of the model (n = 13; 68%), and ultimately increases the certainty of the right decision making (n = 8; 42%) as well as minimizes the risk of making a mistake (n = 2; 11%). Additionally, we asked if the participants identified any potential problems, limitations, threats related to presenting additional model explanations? In 21 people answering this question, the most frequently given answers were: too many explanations require more analysis, which generates the risk of cognitive load (n = 15; 71%), and which may, in consequence, distract the focus on the most important factors (n = 7; 33%). Therefore, some participants highlighted the number of additional explanations as a potential limitation (n = 10; 48%). Moreover, the participants noticed that the explanations must be accompanied by clear instructions for a better understanding of the presented data, because otherwise they do not fulfill their function (n = 6; 29%), and may even introduce additional uncertainty to the assessment of the model (n = 4; 19%).

(1) Evaluating explanations with human subjects is **challenging**.

(2) Our user study indicates that an **interactive sequential analysis of a model** has a potential to increase the **accuracy** and **confidence** of human decision making.

Details? See the paper on arXiv! Questions?

Paper: arXiv:2005.00497

Contact:

www.hbaniecki.com

h.baniecki@uw.edu.pl

Call for postdocs ;-)
www.mi2.ai

This work was financially supported by the NCN OPUS grant no. 2017/27/B/ST6/01307 and SONATA BIS grant no. 2019/34/E/ST6/00052.

HOMER

NATIONAL SCIENCE CENTRE
POLAND

MI RESEARCH