Benchmarking Heterogeneous Treatment Effect Models through the Lens of XAI

Mihaela van der Schaar
John Humphrey Plummer Professor of Machine Learning, Artificial Intelligence and Medicine, University of Cambridge // Director, Cambridge Center for AI in Medicine // Turing Faculty Fellow, The Alan Turing Institute

mv472@cam.ac.uk
@MihaelaVDS
linkedin.com/in/mihaela-van-der-schaar/
Meet our fantastic PhD Students!

Alan Jeffares
Alex Chan
Alicia Curth
Boris van Breugel
Hao Sun
Jeroen Berrevoets
Jonathan Crabbé

Julianna Piskorz
Kasia Kobalcyk
Krzysztof Kacprzyk
Max Ruiz Luyten
Nabeel Seedat
Nicholas Huyn
Nicolás Astorga

Paulius Rauba
Sam Holt
Tennison Liu
Qiyao Wei
Yangming Li
Yuchao Qin

vanderschaar-lab.com
Enormous leaps in AI
And yet….

- Transform healthcare systems?
- Alleviate traffic control?
- Solve the energy crisis?
- Prevent climate catastrophes?
- Avoid the next financial crisis?
- Transform education?
And yet….

- Transform healthcare systems?
- Alleviate traffic control?
- Solve the energy crisis?
- Prevent climate catastrophes?
- Avoid the next financial crisis?
- Transform education?

Not just a few breakthroughs away…..
Because of the complexity of the real-world – a different AI/ML paradigm
A complex world – our lab’s work for the past 20 years

Solving these complex human-centric problems is our biggest task as AI researchers!

vanderschaar-lab.com

university_of_cambridge
A fundamentally new paradigm is needed for AI

• Reality-centric AI aims to reorientate AI towards the complexities of the real world

https://www.vanderschaar-lab.com/the-case-for-reality-centric-ai/
What is Reality-centric AI?

- AI which aims to solve real-world problems
- AI which operates effectively and accountably given the inherent and unavoidable complexities of the real world
- AI which empowers, and does not marginalize humans
Interpretability & Reality-Centric AI

We need to go beyond interpretability of static prediction models
What do clinicians want from an explanation?

www.vanderschaar-lab.com/making-machine-learning-interpretable-a-dialog-with-clinicians/
## 5 classes of explanation methods

<table>
<thead>
<tr>
<th>Explanation class</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature-based</td>
<td>Provides the importance of each feature to model predictions</td>
</tr>
<tr>
<td>Example-based</td>
<td>Explains model predictions with reference to other examples</td>
</tr>
<tr>
<td>Concept-based</td>
<td>Explains model predictions with reference to a human-defined concept</td>
</tr>
<tr>
<td>Model-based</td>
<td>Explains model predictions via an auxiliary meta-model</td>
</tr>
<tr>
<td>Counterfactual</td>
<td>Explains model predictions by generating synthetic example(s) that are similar but with a different prediction</td>
</tr>
</tbody>
</table>
Multiple stakeholders drive diverse interpretability requirements for machine learning in healthcare

Received: 31 January 2023
Accepted: 28 June 2023

Fergus Imrie, Robert Davis & Mihaela van der Schaar

van_der_Schaar
LAB
vanderschaar-lab.com

UNIVERSITY OF CAMBRIDGE
Multiple stakeholders drive diverse interpretability requirements for machine learning in healthcare.
Our lab: Four types of interpretability

<table>
<thead>
<tr>
<th>Explanation class</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature-based</td>
<td>Provides the importance of each feature to model predictions</td>
</tr>
<tr>
<td>Example-based</td>
<td>Explains model predictions with reference to other examples</td>
</tr>
<tr>
<td>Concept-based</td>
<td>Explains model predictions with reference to a human-defined concept</td>
</tr>
<tr>
<td>Model-based</td>
<td>Explains model predictions via an auxiliary meta-model</td>
</tr>
<tr>
<td>Counterfactual</td>
<td>Explains model predictions by generating synthetic example(s) that are similar but with a different prediction</td>
</tr>
</tbody>
</table>
Overview of our lab’s work related to interpretability

Interpretable Resources

Interpretable machine learning

Peering into the ultimate black box

It should be noted that numerous risk factors can be applied to either interpretability study areas. This includes factors such as biases and limitations in the data and models used. Furthermore, it is important to consider the potential for overfitting and the possibility of adversarial attacks. The goal is to develop models that are robust and able to generalize well to unseen data. This requires careful consideration of the trade-offs between model complexity and interpretability. In addition, it is important to carefully evaluate the performance of models in real-world scenarios, where data may be scarce and the models must be able to handle high-dimensional data. Finally, it is important to consider the ethical implications of using interpretable models, particularly in applications where decisions have significant consequences for individuals.

vanderschaar-lab.com/ → Research pillars → Interpretable ML

vanderschaar-lab.com
## Interpretability Resources

### Explainers

Different model architectures can require different interpretability models, or "Explainers". Below are all the explainers included in this repository, with links to their source code and the papers that introduced them. SimplEx, Dynamask, shap, and Symbolic Pursuit have a common python interface implemented for them for ease of implementation (see Interface above and Implementation and Notebooks below). But any of the other methods can also be implemented by using the code in the GitHub column of the table below.

<table>
<thead>
<tr>
<th>Explainers</th>
<th>Affiliation</th>
<th>GitHub</th>
<th>Paper</th>
<th>Date of Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept Activation Regions (CARS)</td>
<td>van der Schaar Lab</td>
<td>CARS source Code</td>
<td>CARS Paper</td>
<td>2022</td>
</tr>
<tr>
<td>ITERpretability</td>
<td>van der Schaar Lab</td>
<td>ITERpretability Source Code</td>
<td>ITERpretability Paper</td>
<td>2022</td>
</tr>
<tr>
<td>Label-Free XAI</td>
<td>van der Schaar Lab</td>
<td>Label-Free XAI Source Code</td>
<td>Label-Free XAI Paper</td>
<td>2022</td>
</tr>
<tr>
<td>SimplEx</td>
<td>van der Schaar Lab</td>
<td>SimplEx Source Code</td>
<td>SimplEx Paper</td>
<td>2021</td>
</tr>
<tr>
<td>Dynamask</td>
<td>van der Schaar Lab</td>
<td>Dynamask Source Code</td>
<td>Dynamask Paper</td>
<td>2021</td>
</tr>
<tr>
<td>INVASE</td>
<td>van der Schaar Lab</td>
<td>INVASE Source Code</td>
<td>INVASE Paper</td>
<td>2019</td>
</tr>
<tr>
<td>SHAP</td>
<td>University of Washington</td>
<td>SHAP Source Code (pytorch implementation: Captum GradientShap)</td>
<td>SHAP Paper</td>
<td>2017</td>
</tr>
</tbody>
</table>

### Open Source Code

[github.com/vanderschaarlab/Interpretability](https://github.com/vanderschaarlab/Interpretability)

---

## Selecting an Interpretability Method

Figure 3 shows a flowchart to help with the process of selecting the method that is most appropriate for your project.

![Interpretability Method selection flowchart](image)

### Implementation and Notebooks

This repository includes a common python interface for the following interpretability methods: SimplEx, Dynamask, shap, and Symbolic Pursuit. The Interface provides the same methods for each of the methods such that you can use the same python methods in your scripts to set up an explainer for each interpretability method. The methods that are:

- `init`: Instantiate the class of explainer of your choice.
- `fit`: Performs and training for the explainer (This is not required for Shap explainers).
- `explain`: Provide the explanation of the data provided.
- `summary_plot`: Visualize the explanation.

There are also Notebooks in this GitHub repository to demonstrate how each create the explainer object. These explainers can be saved and uploaded into the interpretability Suite user interface.
SimplEx

SimplEx is a case-based interpretability method. It can work with either tabular or time series data. You can read more about it in the [paper](#). For clinically focussed examples go to the bespoke SimplEx Demonstrator. And for further information, [here](#) is a video demonstration of the clinical SimplEx app.

### Examples

**Data type:**
- Tabular

**Dataset:**
- ids

**Model:**
- MLP

**Test record:**
- 0

### Test record:

<table>
<thead>
<tr>
<th>sepal length (cm)</th>
<th>sepal width (cm)</th>
<th>petal length (cm)</th>
<th>petal width (cm)</th>
<th>Test Prediction</th>
<th>Test Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.70000</td>
<td>3.00000</td>
<td>6.10000</td>
<td>2.30000</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

### Corpus:

<table>
<thead>
<tr>
<th>Corpus member 0</th>
<th>sepal length (cm)</th>
<th>sepal width (cm)</th>
<th>petal length (cm)</th>
<th>petal width (cm)</th>
<th>Examine Importance</th>
<th>Corpus Prediction</th>
<th>Corpus Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.60000</td>
<td>3.00000</td>
<td>6.00000</td>
<td>2.50000</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.70000</td>
<td>2.80000</td>
<td>6.30000</td>
<td>2.60000</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.70000</td>
<td>2.50000</td>
<td>6.50000</td>
<td>2.30000</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.70000</td>
<td>3.60000</td>
<td>6.70000</td>
<td>2.20000</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.00000</td>
<td>3.60000</td>
<td>6.20000</td>
<td>2.40000</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[github.com/vanderschaarlab/Interpretability](#)
Our Resources to go Further

Our Papers
vanderschaar-lab.com/interpretable-machine-learning/

Our Code
github.com/vanderschaarlab/Interpretability
Goal in our work

Estimating the effect of a treatment on an individual patient

Will a given treatment work for an individual patient?

Learning from observational data

Patient characteristics
Treatment status
Observed outcomes

vanderschaar-lab.com
Treatment Effect Estimation using Potential Outcomes

Potential Outcomes (POs)  Fundamental problem of causal inference  CATE: Conditional Average Treatment Effect

We only observe one PO

→ Either $Y_i^{(0)}$ or $Y_i^{(1)}$

$\tau(x) = \mathbb{E}[Y_i^{(1)} - Y_i^{(0)} \mid X_i = x]$

Label $Y_i^{(1)} - Y_i^{(0)}$ is never observed

→ Challenging estimation and model selection!
The rapidly expanding ML CATE/ITE Estimator toolbox*
*under ignorability assumptions

- **Model-specific strategies:** Adapting specific ML methods for CATE/ITE estimation
  - Random forest, Xgboost, Gaussian Processes, Deep Learning – GANs, MLPs, etc.

- **Model-agnostic/meta-learner strategies:** constructed *recipes* for estimating CATE/ITE using *any arbitrary* ML method
  - Different learning strategies (e.g. S-, T-, DR-, R-learner)
  - Can be implemented using different ML methods

**Key question:** Given multiple CATE/ITE estimates based on (i) different meta-learner strategies and (ii) different ML methods, how do I pick the best one for use in practice??
**ITErpretability:**

Benchmarking Heterogeneous Treatment Effect Models through the Lens of Interpretability

[Crabbe, Curth, Bica, vdS, NeurIPS 2022]

**Desiderata:** Interpretability methods that *discover true drivers of effect heterogeneity* (and lead to *correct interpretations of ITE methods*)

- Investigate how to use *feature importance methods* to interpret black-box ML estimators and use this to evaluate them on their ability to discover drivers of heterogeneity
- Propose a benchmarking environment – *ITErpretability benchmark*
- Provide new insights into ITE methods
How to interpret heterogeneous treatment effect estimators?

Two *types of important covariates*:

1. **Prognostic covariates**, which influence outcomes regardless of treatment (i.e. common risk factors such as e.g. age)
2. **Predictive covariates**, which determine differential responses to treatment (i.e. drivers of heterogeneity e.g. hormone receptor status)

→ *Predictive covariates* are of most interest when interpreting the learned treatment effect function \( \hat{r}(x) \)
How to interpret heterogeneous treatment effect estimators?

- **Our approach**: We use a feature importance method $a_i(\hat{\tau}, x) \in \mathbb{R}$ to assess whether $\hat{\tau}(x)$ has learned that feature $i$ is predictive (importance increases with $|a_i(\hat{\tau}, x)|$).

- "Do estimators pay attention to the correct covariates?"
  1. $\text{Attr}_{\text{pred}}$: proportion of importance correctly attributed to predictive cov. 😊
  2. $\text{Attr}_{\text{prog}}$: proportion of importance incorrectly attributed to prognostic cov. 😞
The ITErpretability Benchmark

• We propose a *semi-synthetic* benchmarking environment: using real covariates but simulated outcomes and treatment assignments
  → Ensures that covariate classification is *known by design*
The ITerpretability Benchmark

Covariates Dataset
- Patient Covariates $x \in \mathbb{R}^d$

Data Generating Process

(1) Feed
(2) Generate
(3) Generate

Predictive Covariates $\mathcal{P}_{\text{pred}} \subset \{1, \ldots, d\}$

Semi-Synthetic Dataset $\mathcal{D}$

- Patient Covariates $x \in \mathbb{R}^d$
- Simulated Assignment $w \in \{0,1\}$
- Simulated Outcome $y \in \mathbb{R}$

(4) Fit
(5) Interpret

Black-Box CATE Estimator $\hat{\tau}$

Feature Importance Scores $\alpha_i(\hat{\tau}, x)$ for each covariate $x_i$

(6) Compare

van_derschaar_LAB
vanderschaar-lab.com
The ITErpretability Benchmark

• We propose a *semi-synthetic* benchmarking environment: using real covariates but simulated outcomes and treatment assignments
  → Ensures that covariate classification is *known by design*

• Flexible DGP allows to study the effect of interesting *experimental knobs*
  • Strength of predictive vs prognostic effects
  • Type of treatment selection
  • Etc.
Turning the lights on

Different types of black-box estimators in multiple experimental setups
Turning the lights on

✓ S-Learners appear to struggle most to make correct attribution
✓ Using attribution metrics leads to interesting new insights relative to considering only PEHE
✓ S-learner does best in terms of PEHE when predictive strength is low, while the T-learner does worst – yet this does not seem to translate into better discoveries!

• We additionally investigate the effects of
  ✓ The degree of nonlinearity
  ✓ The type and degree of confounding (treatment selection bias)
Take-aways

1. Aim to solve complex Reality-Centric Problems

2. Reality-Centric AI requires Empowering Humans!

3. XAI is an important part of the solution

4. XAI goes beyond static predictions — causality, time-series, RL/IRL etc.

5. Multiple stakeholders in XAI – Engage with them, explore the multiple facets of interpretability!
Our Resources to go Further

Our Code
github.com/vanderschaarlab/Interpretability

Our Papers
vanderschaar-lab.com/interpretable-machine-learning/
Engagement sessions: Inspiration Exchange

https://www.vanderschaar-lab.com/
→ Engagement sessions
→ Inspiration Exchange

Nov 6, 2023
4pm UK time