Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond

Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, Marina M.-C. Höhne

Neural Information Processing Systems (NeurIPS), 2023

@anna_hedstroem
@TUBerlin_UMI

JMLR

VOL. 24

Journal of Machine Learning Research 24 (2023) 1-11 Submitted 2/22; Revised 11/22; Published 1/23

Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond

Anna Hedström[1,†] ANNA.HEDSTROEM@TU-BERLIN.DE
Leander Weber[3] LEANDER.WEBER@HHI.FRAUNHOFER.DE
Dilyara Bareeva[1] DILYARA.BAREEVA@CAMPUS.TU-BERLIN.DE
Daniel Krakowczyk[4] DANIEL.KRAKOWCZYK@UNI-POTSDAM.DE
Franz Motzkus[3] FRANZ.MOTZKUS@HHI.FRAUNHOFER.DE
Wojciech Samek[2,3,5] WOJCIECH.SAMEK@HHI.FRAUNHOFER.DE
Sebastian Lapuschkin[3,†] SEBASTIAN.LAPUSCHKIN@HHI.FRAUNHOFER.DE
Marina M.-C. Höhne[1,5,†] MARINA.HOEHNE@TU-BERLIN.DE

[1] *Understandable Machine Intelligence Lab, TU Berlin, 10587 Berlin, Germany*
[2] *Department of Electrical Engineering and Computer Science, TU Berlin, 10587 Berlin, Germany*
[3] *Department of Artificial Intelligence, Fraunhofer Heinrich-Hertz-Institute, 10587 Berlin, Germany*
[4] *Department of Computer Science, University of Potsdam, 14476 Potsdam, Germany*
[5] *BIFOLD – Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany*
[†] *corresponding authors*

Editor: Joaquin Vanschoren

Abstract

The evaluation of explanation methods is a research topic that has not yet been explored deeply, however, since explainability is supposed to strengthen trust in artificial intelligence, it is necessary to systematically review and compare explanation methods in order to confirm

The Quantus Team

*ANNA HEDSTRÖM | LEANDER WEBER | DANIEL KRAKOWCZYK | DILYARA BAREEVA




FRANZ MOTZKUS | WOJCIECH SAMEK | SEBASTIAN LAPUSCHKIN | MARINA M.-C. HÖHNE

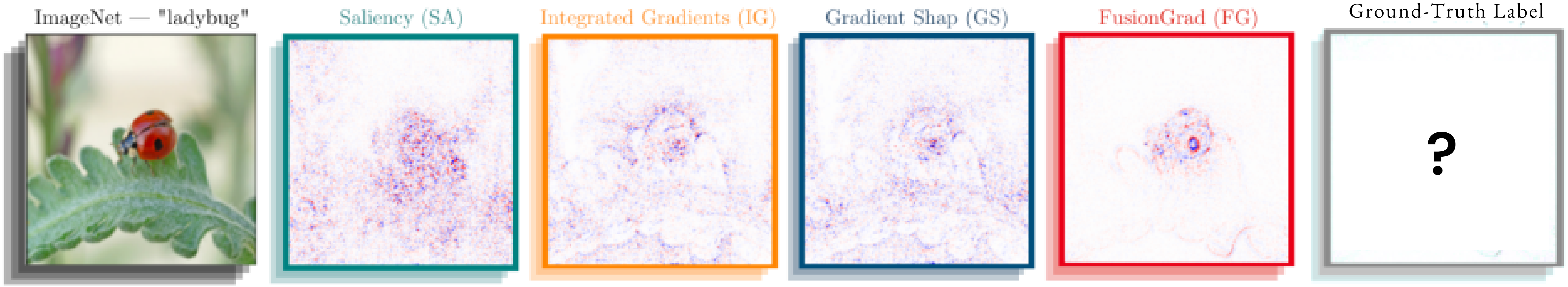



+ FANTASTIC CONTRIBUTORS

1. Problem — Evaluating Explainability

The Challenge of Explanation Method Selection

- Without access to ground truth explanation labels, difficult in determining the quality of explanations

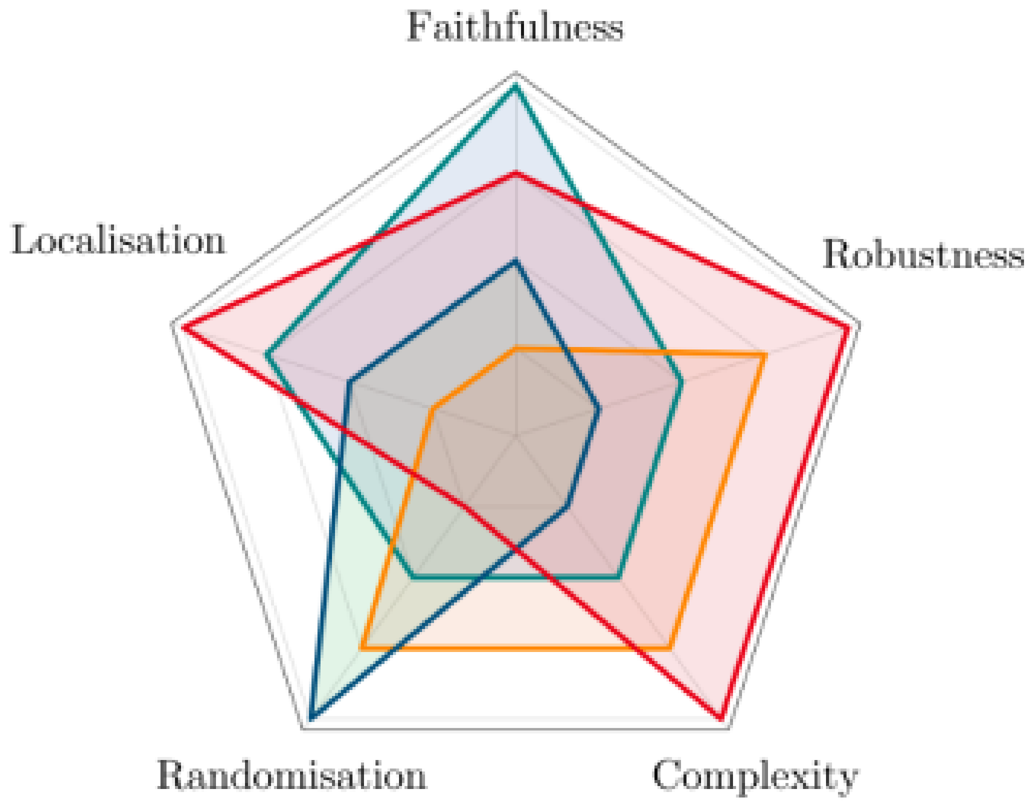


- Complete lack of open-source tools for XAI evaluation

2. Objectives — Automate Evaluation

Enable XAI Quantification for Researchers at Large Scale

- Enable automation and large-scale experimentation, across a diverse set of evaluation properties, models and datasets
- Provide the XAI and ML communities with an efficient, easy-to-use open-sourced API to perform XAI evaluation
- Give a quantitative snapshot of the explanation quality



3. Related Works

From little to booming interest

- No single evaluation-centric XAI library, at the time of development

Table 1: Comparison of four XAI libraries — (`AIX360` [2], `captum` [29], `TorchRay` [30] and `Quantus`) in terms of the number of XAI evaluation methods for six different evaluation categories, as implemented in each library.

Library	Faithfulness	Robustness	Localisation	Complexity	Axiomatic	Randomisation
`Captum` (2)	1	1	0	0	0	0
`AIX360` (2)	2	0	0	0	0	0
`TorchRay` (1)	0	0	1	0	0	0
`Quantus` (27)	**9**	**4**	**6**	**3**	**3**	**2**

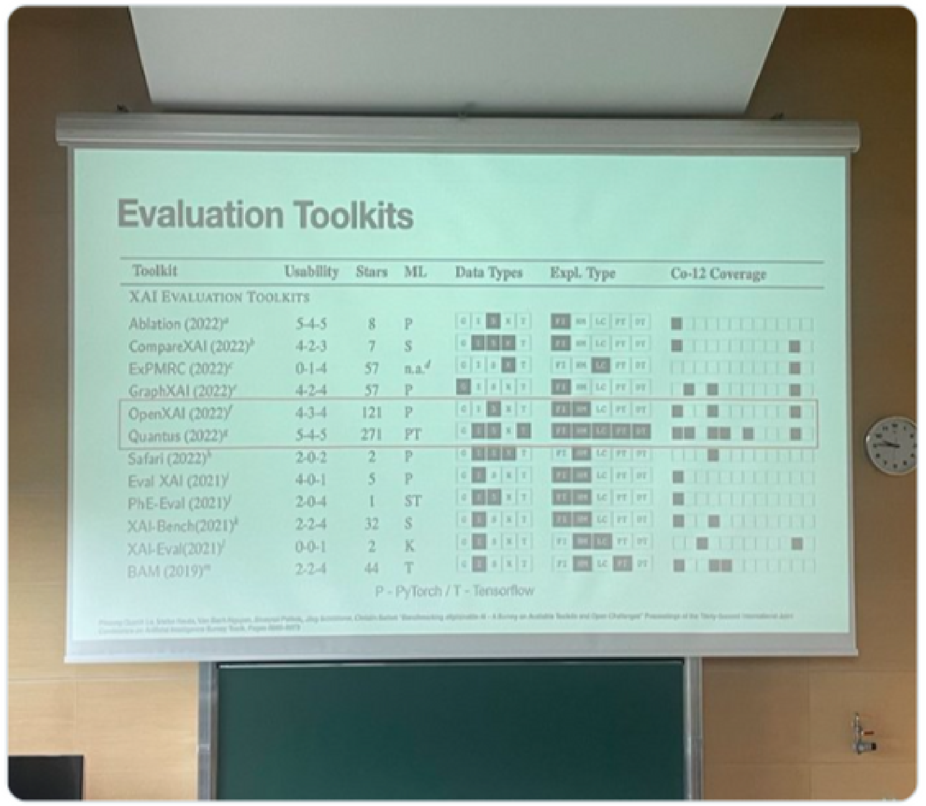
3. Related Works

From little to booming interest

- In recent years, XAI evaluation has boomed

Understandable Machine Intelligence Lab
@UMI_Lab_AI · **Follow**

In @SeifertChristin's talk at XAI^3 worship at #ECAI2023, our Quantus library was named one of the top XAI evaluation libraries! Great job @anna_hedstroem and co-authors!



9:52 AM · Oct 1, 2023

10 Reply Copy link

Read more on X

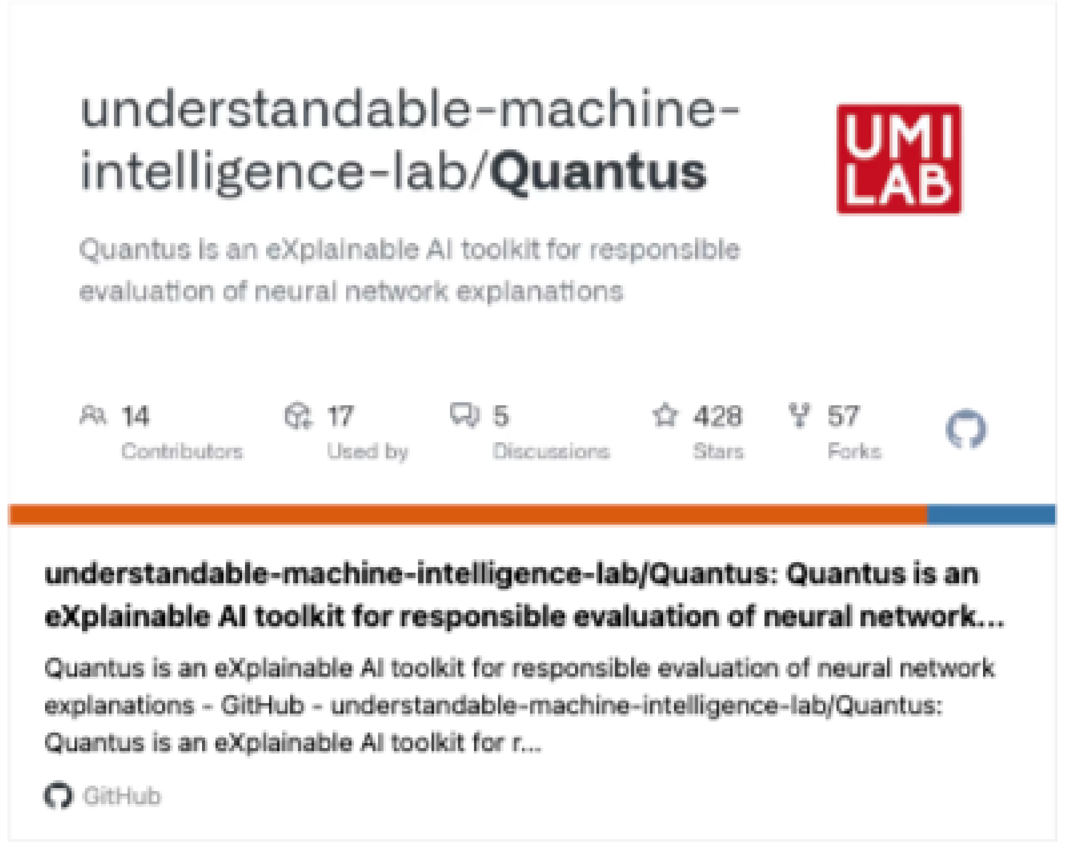
4. Library Content — Metrics

Evaluate Explanations from PyTorch and Tensorflow Models

- Metrics. 30+ metrics in 6 categories for XAI evaluation with tutorials and API reference

- Data and model types. Support (image, time-series, tabular, NLP in progress!) datasets for PyTorch and Tensorflow ML models

- Feature-importance methods. E.g., gradient-, back-propagation-, model-agnostic, local surrogate-, attention-, prototype-based explanations



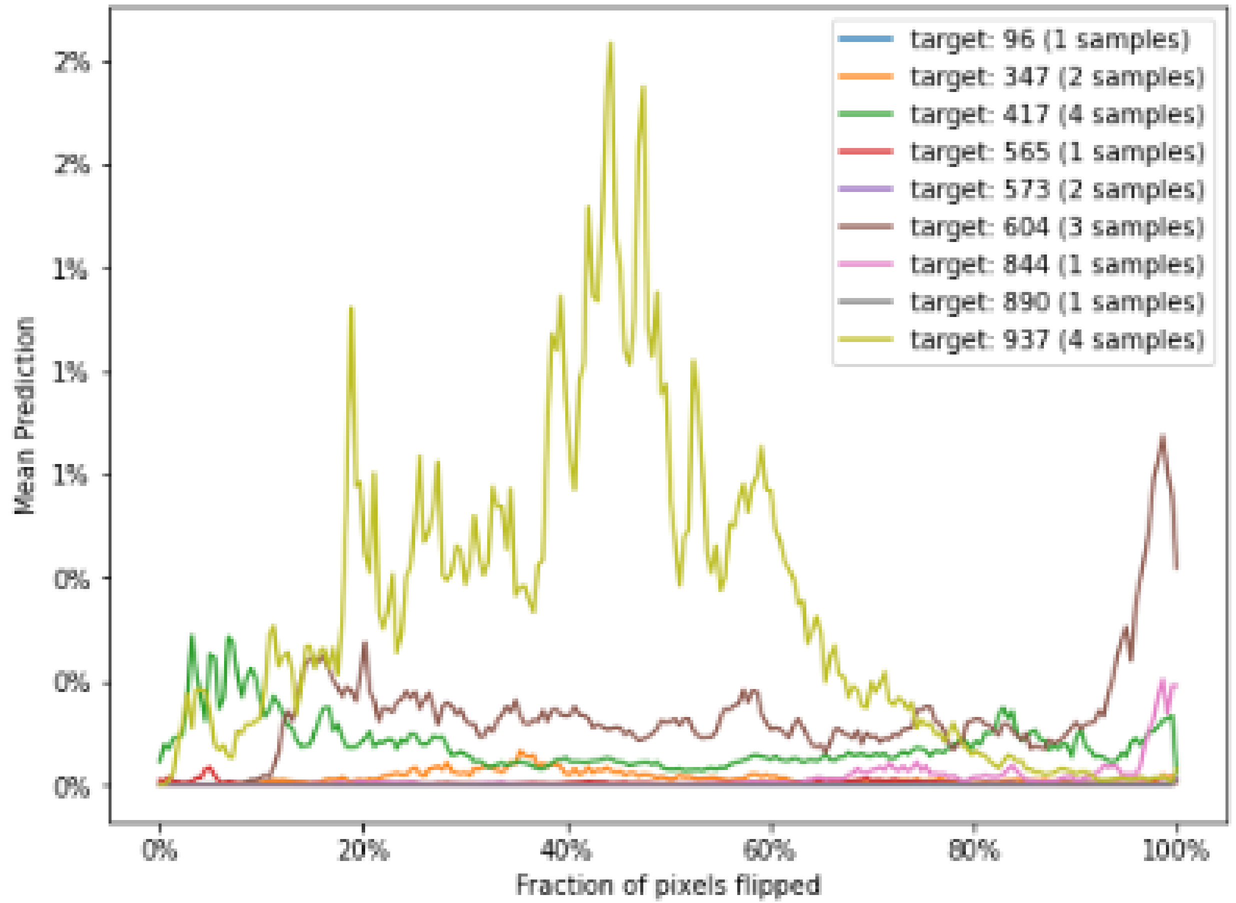
5. Library Syntax

Evaluation in an one-liner or with **quantus.evaluate()**

```
# Create the pixel-flipping experiment.
pixel_flipping = quantus.PixelFlipping(
    features_in_step=224,
    perturb_baseline="black",
    perturb_func=quantus.baseline_replacement_by_indices,
)

# Call the metric instance to produce scores.
scores = pixel_flipping(model=model,
                        x_batch=x_batch,
                        y_batch=y_batch,
                        a_batch=a_batch,
                        device=device,)

# Plot example!
pixel_flipping.plot(y_batch=y_batch, scores=scores)
```



__init__ the metric in one go

plot() to visualise some results

score xAI methods using **__call__**

6. Applications — Highlights

Diverse Applications Across Fields

Climate science [1, 2]

Healthcare [3, 4, 5, 6, 7]

Object Detection [12]

Remote sensing [14]

Image Classification [8, 9]

Security [15]

Meta-evaluation [10, 11]

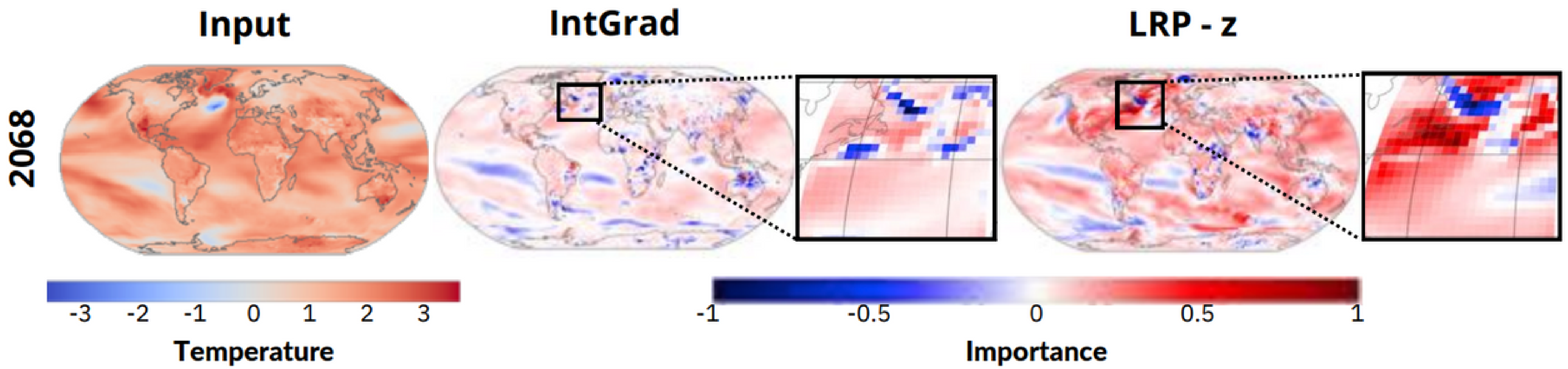
Network Canonization [13]

....

6. Applications — Highlights

Diverse Applications Across Fields

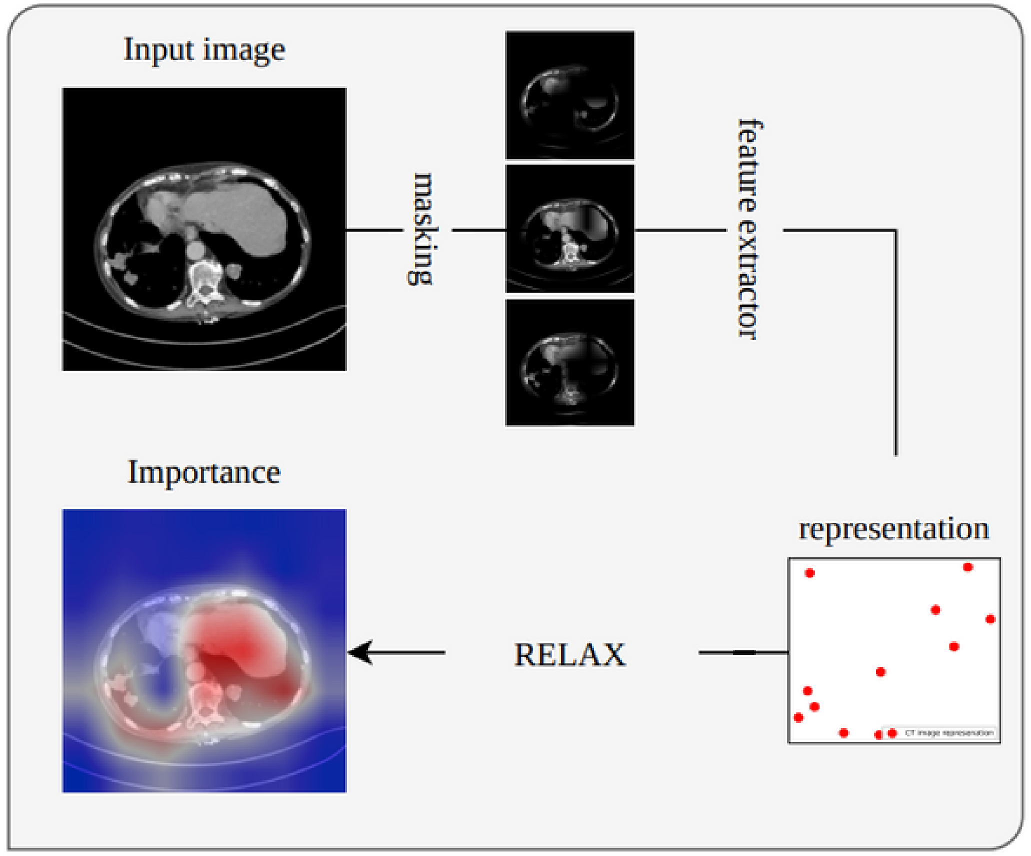
Climate science [1, 2] — Evaluate explanations of temperature prediction models



Bommer, Philine, et al. "Finding the right XAI method--A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science." arXiv preprint arXiv:2303.00652 (2023).

6. Applications — Highlights

Diverse Applications Across Fields

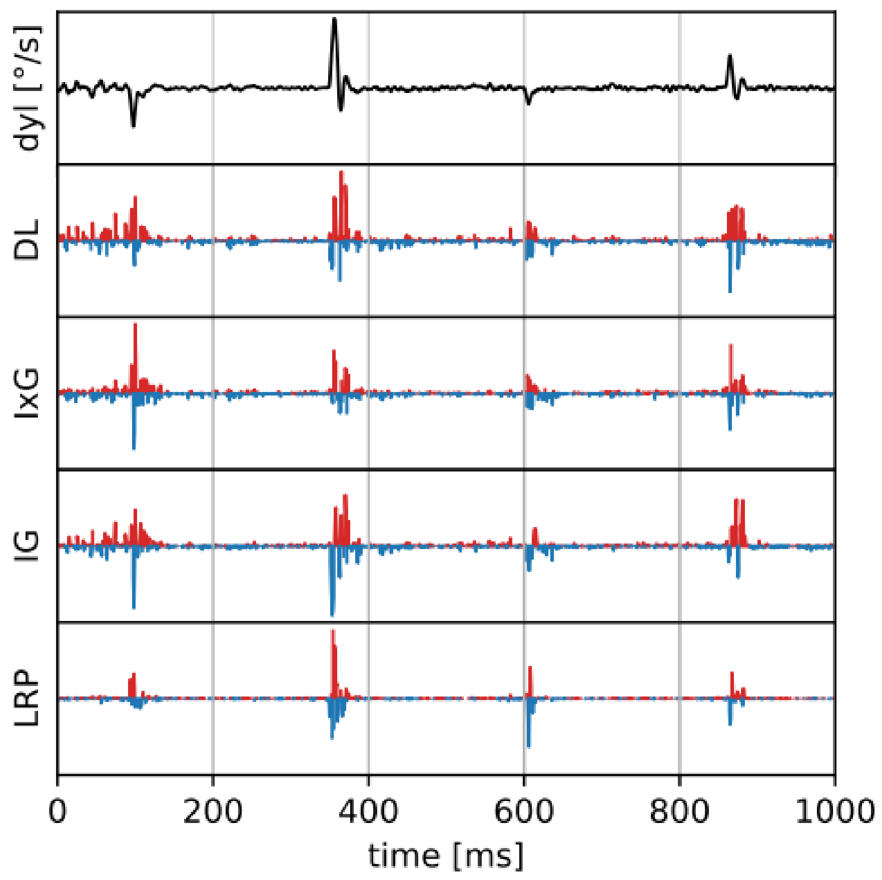


Healthcare [3, 4, 5, 6, 7] — Evaluate explanations of liver disease models

Wickstrøm, Kristoffer Knutsen, et al. "A clinically motivated self-supervised approach for content-based image retrieval of CT liver images." Computerized Medical Imaging and Graphics 107 (2023): 102239.

6. Applications — Highlights

Diverse Applications Across Fields



(c) pitch velocities of left eye

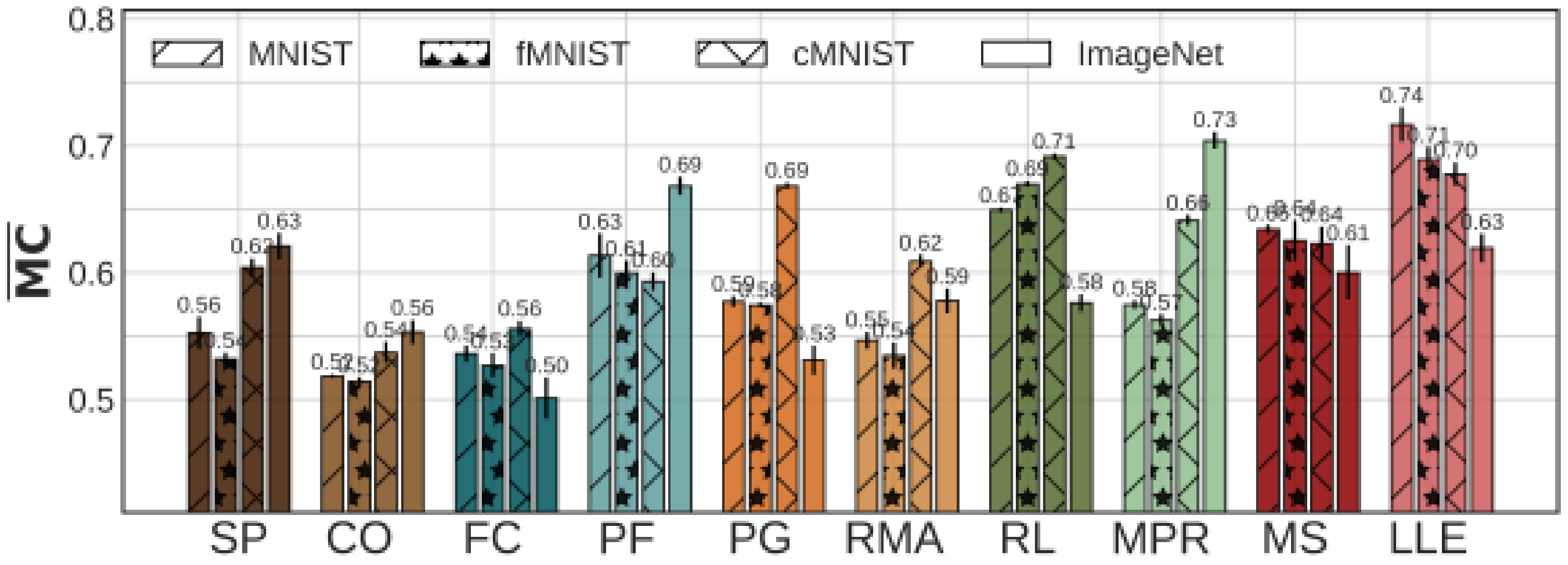
Healthcare [3, 4, 5, 6, 7] — Evaluate explanations of biometric eye-tracking models

Krakowczyk, Daniel G., et al. "Bridging the Gap: Gaze Events as Interpretable Concepts to Explain Deep Neural Sequence Models." Proceedings of the 2023 Symposium on Eye Tracking Research and Applications. 2023.

6. Applications — Highlights

Diverse Applications Across Fields

Meta-Evaluation [10, 11] — Evaluate the "evaluation methods" themselves



Hedström, Anna, et al. "The Meta-Evaluation Problem in Explainable AI: Identifying Reliable Estimators with MetaQuantus." arXiv preprint arXiv:2302.07265 (2023).

Post-script

Thank you

- **Learn more?** Read the paper and check out the API documentation
- **Get started?** Check out the repository with tutorials.
- **Contribute?** Check out our current issues.
- **Contact?** Write to hedstroem.anna@gmail.com

@anna_hedstroem
@TUBerlin_UMI

Thank you

References

[1] Bommer, Philine, et al. "Finding the right XAI method--A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science." arXiv preprint arXiv:2303.00652 (2023).

[2] Bommer, Philine, et al. "Tutorial: Quantus x Climate - Applying explainable AI evaluation in climate science (Tutorials Track)" ICLR https://www.climatechange.ai/papers/iclr2023/1 (2023)

[3] Wickstrøm, Kristoffer Knutsen, et al. "A clinically motivated self-supervised approach for content-based image retrieval of CT liver images." Computerized Medical Imaging and Graphics 107 (2023): 102239.

[4] You, Suhang, Roland Wiest, and Mauricio Reyes. "SaRF: Saliency regularized feature learning improves MRI sequence classification." Computer methods and programs in biomedicine 243 (2024): 107867.

[5] Komorowski, Piotr, Hubert Baniecki, and Przemyslaw Biecek. "Towards Evaluating Explanations of Vision Transformers for Medical Imaging." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[6] P. Kang, J. Li, S. Jiang and P. B. Shull, "Reduce System Redundancy and Optimize Sensor Disposition for EMG-IMU Multimodal Fusion Human-Machine Interfaces With XAI," in IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1-9, 2023, Art no. 2500209, doi: 10.1109/TIM.2022.3232159.

[7] Krakowczyk, Daniel G., et al. "Bridging the Gap: Gaze Events as Interpretable Concepts to Explain Deep Neural Sequence Models." Proceedings of the 2023 Symposium on Eye Tracking Research and Applications. 2023.

References

[8] Bykov, Kirill, et al. "Noisegrad—enhancing explanations by introducing stochasticity to model weights." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 6. 2022.

[9] Ahmad, Ola, et al. "Causal Analysis for Robust Interpretability of Neural Networks." arXiv preprint arXiv:2305.08950 (2023).

[10] Hedström, Anna, et al. "The Meta-Evaluation Problem in Explainable AI: Identifying Reliable Estimators with MetaQuantus." arXiv preprint arXiv:2302.07265 (2023).

[11] Stassin, Sédrick, et al. "An Experimental Investigation into the Evaluation of Explainability Methods." arXiv preprint arXiv:2305.16361 (2023).

[12] Dreyer, Maximilian, et al. "Revealing Hidden Context Bias in Segmentation and Object Detection through Concept-specific Explanations." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[13] Pahde, Frederik, et al. "Optimizing Explanations by Network Canonization and Hyperparameter Search." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[14] Mohan, Akshatha, and Joshua Peeples. "Quantitative Analysis of Primary Attribution Explainable Artificial Intelligence Methods for Remote Sensing Image Classification." arXiv preprint arXiv:2306.04037 (2023).

[15] Bhusal, Dipkamal, et al. "SoK: Modeling Explainability in Security Analytics for Interpretability, Trustworthiness, and Usability." Proceedings of the 18th International Conference on Availability, Reliability and Security. 2023.