Physikalisch-Technische Bundesanstalt
Braunschweig und Berlin
Nationales Metrologieinstitut

Quality in Artificial Intelligence
QAI LABS

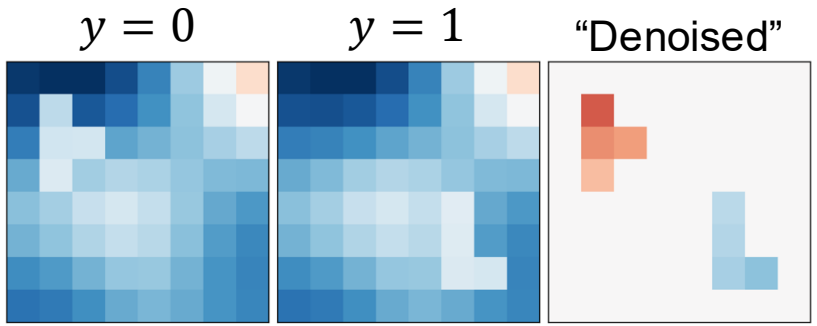
Correcting misinterpretations of additive models

Benedict Clark, *Rick Wilming,*
Hjalmar Schulz, Rustam Zhumagambetov,
Danny Panknin, Stefan Haufe

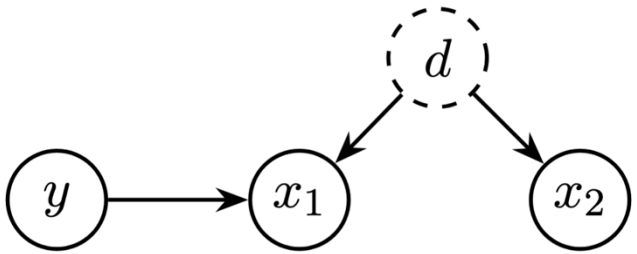
NEURAL INFORMATION PROCESSING SYSTEMS 2025

The illusion of "intrinsic interpretability"

- **Suppressors**: variables uninformative about the target y but used by model (i.e., for denoising)
- Optimal models **must** put non-zero weights on suppressors (Haufe et al., 2014; Wilming et al., 2023)
- However, weights are traditionally interpreted as statistical association
- The **linear activation pattern** $a_i = Cov[x_i, \hat{y}]Var[\hat{y}]^{-1}$ correctly assigns $w_2 = 0$ (Haufe et al., 2014)



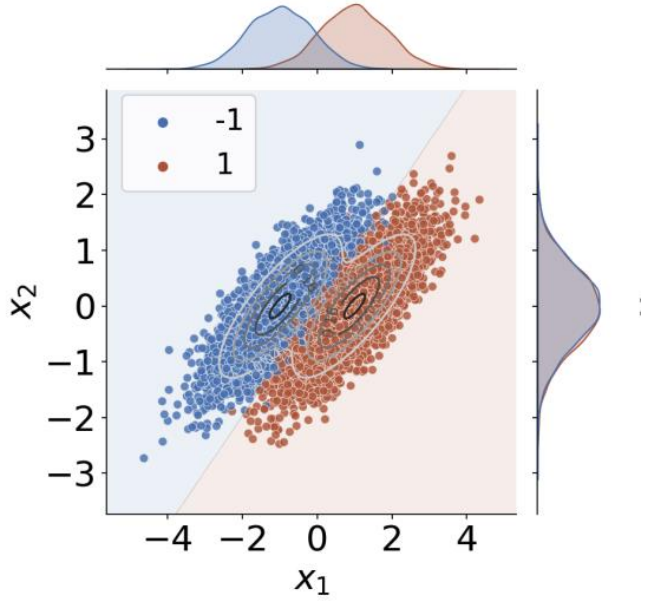
(a)



(b)

$$y \sim \mathcal{N}(\mu_y, \sigma_y^2)$$
$$d \sim \mathcal{N}(\mu_d, \sigma_d^2)$$
$$x_1 := y + d$$
$$x_2 := d$$
$$\hat{y} := x_1 - x_2 = y$$
$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$$

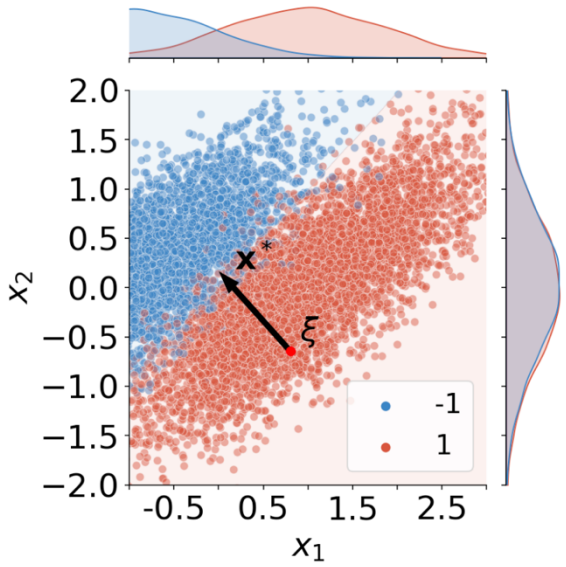
(c)



(d)

XAI methods attribute significant importance to suppressors

Counterfactual Explanation

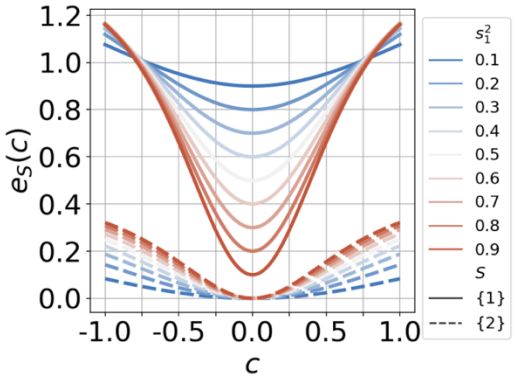


Integrated Gradient

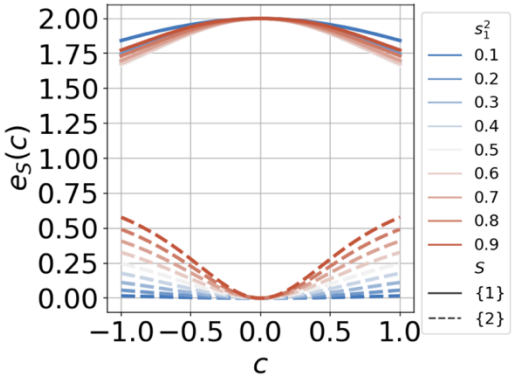
$$e_j(\mathbf{x}) \coloneqq (x_j - x'_j) \int_{[0,1]} \frac{\partial f}{\partial x_j}(\gamma(t))\,\mathrm{d}t.$$

$$e_{\{1\}}(\mathbf{x}) = \alpha(x_1 - x'_1)$$

$$e_{\{2\}}(\mathbf{x}) = -\frac{\alpha c s_1}{2 s_2}(x_2 - x'_2)$$



(a) Faithfulness through pixel flipping



(b) Permutation feature importance

(a)
Non-zero attribution to suppressor feature x_2

(b)
Faithfulness and permutation-based metrics are also susceptible

Wilming, Rick, et al., Theoretical Behavior of XAI Methods in the Presence of Suppressor Variables, *ICML, 2023.*

Why this means XAI underserves its goals

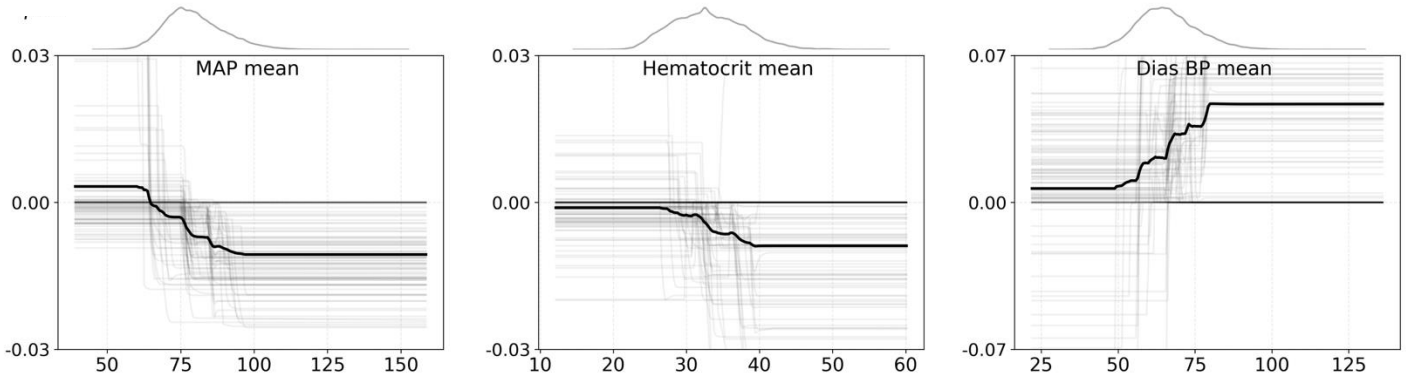
Model (in)validation	What did the model learn?	Unfair model rejection
Data (in)validation	Does the data have any issues?	XAI cannot distinguish suppressors from other types of features
Intervention	Can we change the outcome?	Manipulating a suppressor will not change the real-world outcome
Scientific discovery	Can we discover novel features and interactions?	Hypotheses based on suppressors are a waste of time

Haufe, Stefan, et al., Explainable AI needs formalization*, under review at npj Artifical Intelligence, presented at the NeurIPS 2024 workshop on Interpretable AI: Past, Present, and Future, 2024.*

Generalised Additive Models (GAMs)

- Each f_i^{GAM}, f_{jk}^{GAM} can be an arbitrary non-linear function
- We can also fit bivariate functions f_{jk}^{GAM} for pairs of features x_j, x_k
- **Shape functions can represent suppressor effects**
- Non-flat shape functions do **not** imply statistical association to the prediction target

$$g(E[y|\boldsymbol{x}]) = \sum_{i}^{D} f_i^{GAM}(x_i) + \sum_{j<k}^{D} f_{jk}^{GAM}(x_j, x_k)$$



MIMIC-IV
(Johnson et al., 2023)

Activation pattern analogues for GAMs

Statistical Association Property (SAP) – Features should only be attributed significant non-zero importance if they have a statistical association to the prediction target (Wilming et al., 2023)

For most downstream purposes of XAI, this is a prerequisite condition (Haufe et al., 2024)

$$a_i = Cov[x_i, \hat{y}]Var[\hat{y}]^{-1}$$

(a) Linear Activation Pattern

$$a_i^{PGAM} = Cov\left[f_i^{GAM}, g(\hat{y})\right]Var[g(\hat{y})]^{-1}$$

(b) GAM Activation Pattern

Activation pattern analogues for GAMs

Statistical Association Property (SAP) – Features should only be attributed significant non-zero importance if they have a statistical association to the prediction target (Wilming et al., 2023)

For most downstream purposes of XAI, this is a prerequisite condition (Haufe et al., 2024)

$$a_i = Cov[x_i, \hat{y}]Var[\hat{y}]^{-1}$$

(a) Linear Activation Pattern

$$a_i^{PGAM} = Cov\left[f_i^{GAM}, g(\hat{y})\right]Var[g(\hat{y})]^{-1}$$

(b) GAM Activation Pattern

PatternGAM: SAP-compliant shape functions

- Raw shape functions f_i^{GAM} can possess arbitrary scale
  - This makes it difficult to interpret each a_i^{PGAM}

$$a_i^{PGAM} = Cov[f_i, g(\hat{y})]Var[g(\hat{y})]^{-1}$$
$$\equiv a_i^{PGAM}\hat{y}_i + c_i = x_i$$

- Univariate Linear Logistic Regression (LLR) fits f_i^{PGAM} (and their coefficients b_i) avoid this issue, and possess the SAP

$$f_i^{PGAM}(x_i) = b_i f_i^{GAM}(x_i) + d_i$$

- We observe other SAP-compliant quantities:

$$SD(f_i^{PGAM}) \qquad DISCR(f_i^{GAM})$$

PatternGAM: SAP-compliant shape functions

- Raw shape functions f_i^{GAM} can possess arbitrary scale
  - This makes it difficult to interpret each a_i^{PGAM}

$$a_i^{PGAM} = Cov[f_i, g(\hat{y})]Var[g(\hat{y})]^{-1}$$
$$\equiv a_i^{PGAM}\hat{y}_i + c_i = x_i$$

- Univariate Linear Logistic Regression (LLR) fits f_i^{PGAM} (and their coefficients b_i) avoid this issue, and possess the SAP

$$f_i^{PGAM}(x_i) = b_i f_i^{GAM}(x_i) + d_i$$

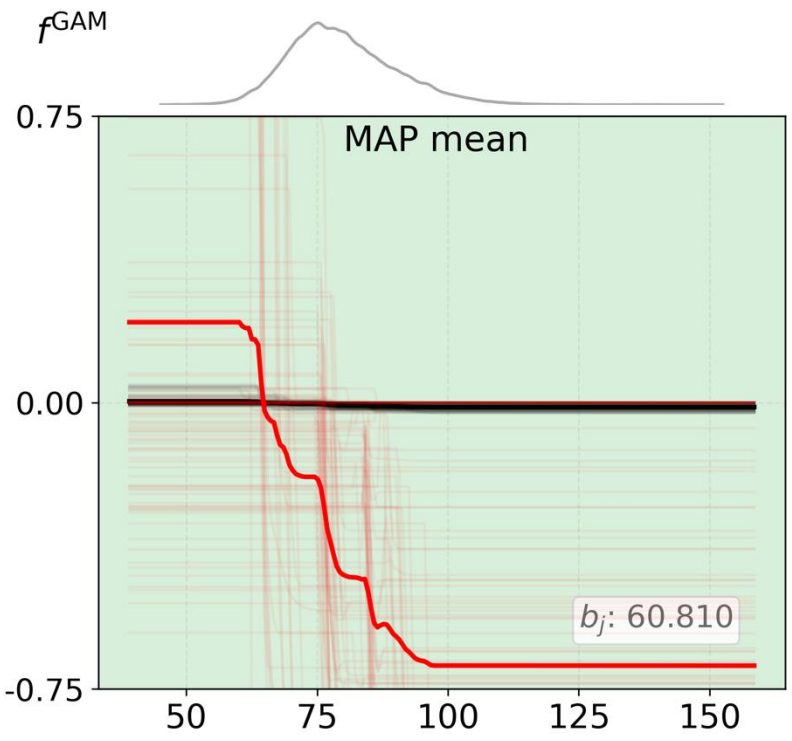
- We observe other SAP-compliant quantities:

$SD(f_i^{PGAM})$ $DISCR(f_i^{GAM})$

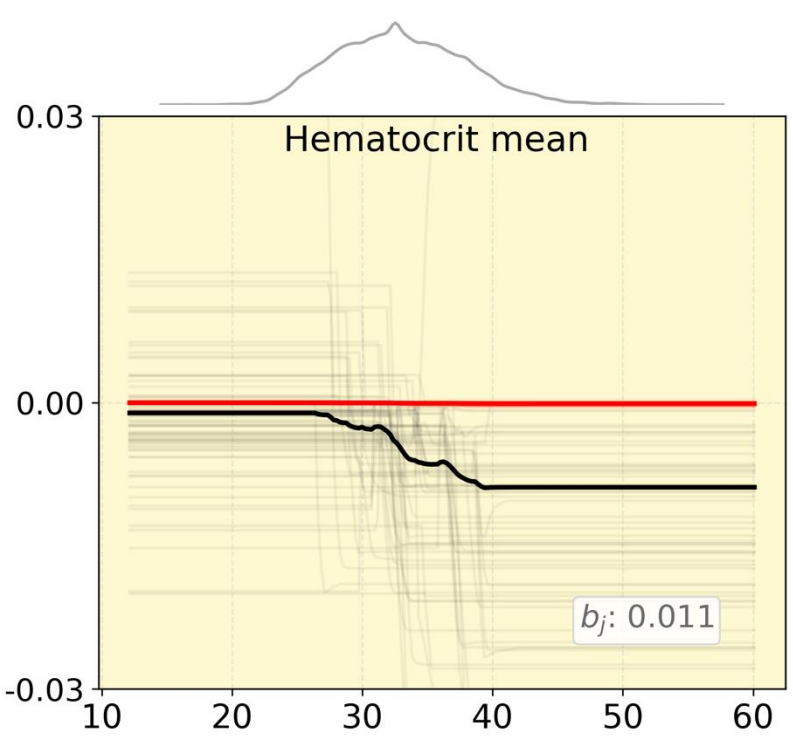
- But not:

f_i^{GAM} $SD(f_i^{GAM})$

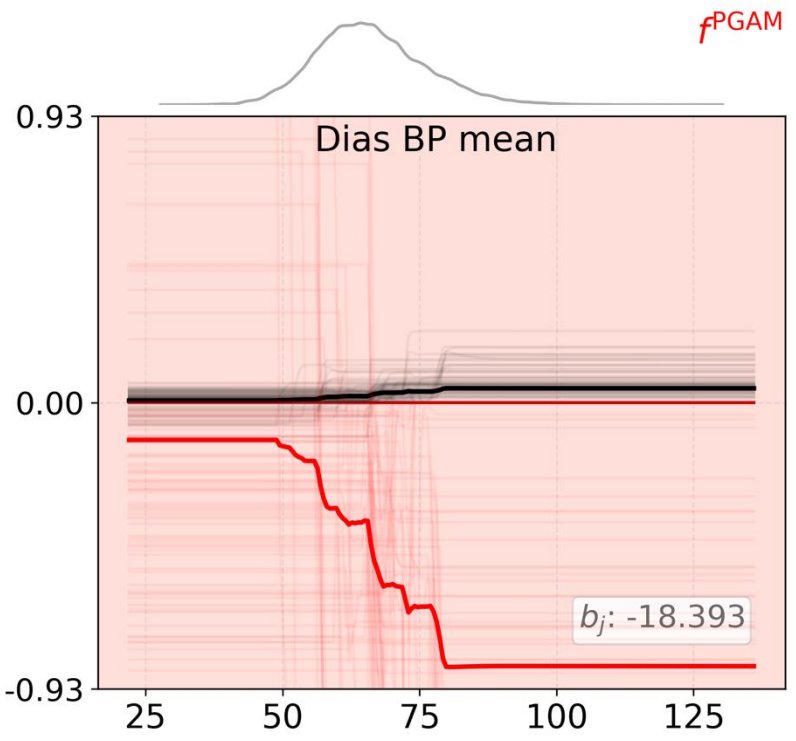
MIMIC-IV mortality prediction



Indirect Cause
Not used multivariately, but univariately predictive

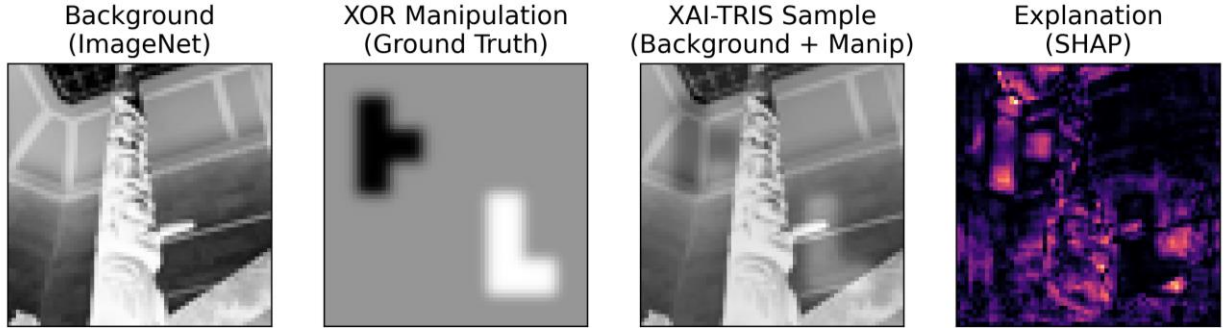


Classical Suppressor
Not informative for mortality prediction



Negative Suppressor
Somewhat informative but used by the model in the opposite direction

Empirical results: XAI-TRIS (Clark et al., 2024)



Method	MULT WHITE	MULT CORR	MULT DIST WHITE	MULT DIST CORR
PAT(f^{GAM})†	0.94 ± 0.01	0.80 ± 0.04	0.94 ± 0.01	0.81 ± 0.04
PAT(f^{QLR})†	0.89 ± 0.00	0.73 ± 0.00	0.89 ± 0.00	0.72 ± 0.00
SD(f^{PGAM})†	0.98 ± 0.01	0.96 ± 0.02	0.98 ± 0.01	0.96 ± 0.02
SD(f^{PQLR})†	0.96 ± 0.00	0.76 ± 0.00	0.96 ± 0.00	0.76 ± 0.00
DISCR(f^{GAM})†	0.91 ± 0.01	0.88 ± 0.03	0.91 ± 0.01	0.88 ± 0.03
DISCR(f^{QLR})†	0.90 ± 0.00	0.76 ± 0.00	0.90 ± 0.00	0.76 ± 0.00
PROD($f^{P/GAM}$)†	$\mathbf{1.00 \pm 0.01}$	$\mathbf{0.99 \pm 0.01}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.99 \pm 0.01}$
PROD($f^{P/QLR}$)†	$\mathbf{1.00 \pm 0.00}$	0.82 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.83 ± 0.00
SD(f^{GAM})	0.97 ± 0.01	0.89 ± 0.02	0.97 ± 0.02	0.89 ± 0.01
SD(f^{QLR})	0.90 ± 0.00	0.75 ± 0.00	0.90 ± 0.00	0.75 ± 0.00
EBM	0.96 ± 0.00	0.90 ± 0.00	0.96 ± 0.00	0.89 ± 0.00
Kernel Pattern	0.68 ± 0.02	0.63 ± 0.05	0.68 ± 0.02	0.62 ± 0.05
PatternNet	0.76 ± 0.04	0.69 ± 0.01	0.77 ± 0.04	0.70 ± 0.02
PatternAttribution	0.95 ± 0.05	0.82 ± 0.05	0.95 ± 0.03	0.83 ± 0.05
SHAP	0.88 ± 0.03	0.82 ± 0.04	0.88 ± 0.03	0.82 ± 0.04
Int. Grads.	0.98 ± 0.01	0.87 ± 0.00	0.98 ± 0.01	0.88 ± 0.00

Further insights are in the paper

- Principled methods for handling bivariate **feature interactions**

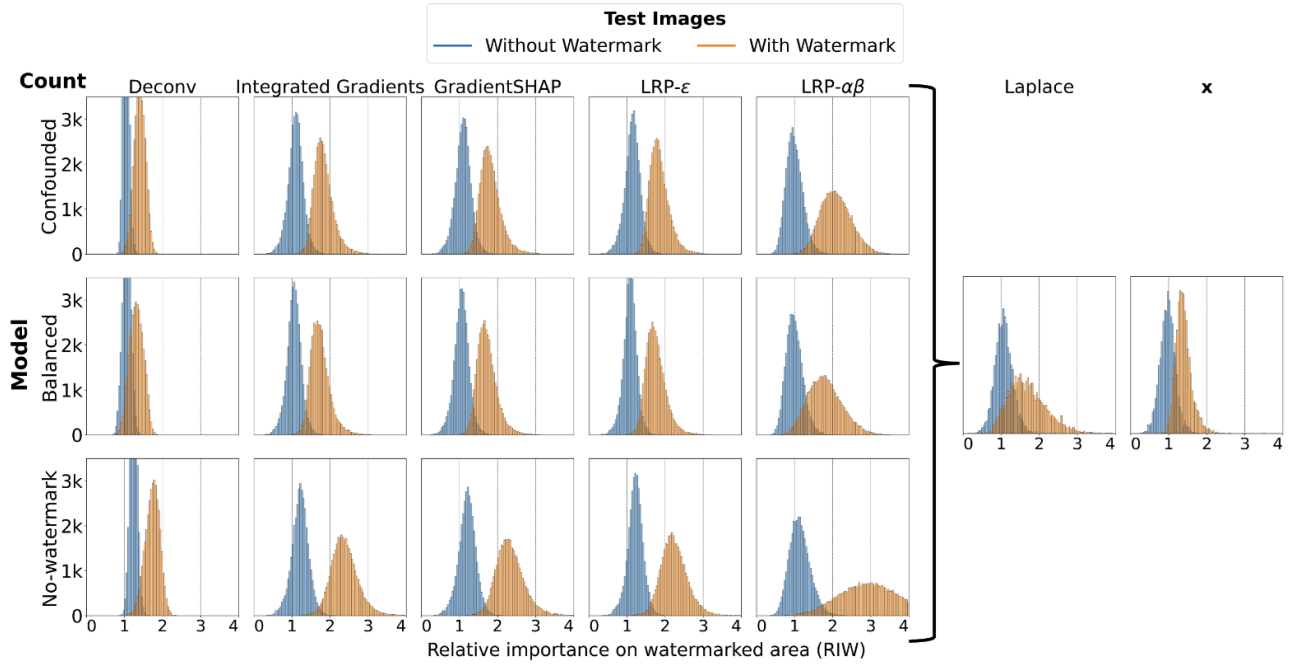
- Extensions to Wasserstein distance-based measures of **explanation correctness**

- Further empirical results and analysis of real-world data

Feature attribution is driven by saliency, not task-informativeness

- We construct datasets with varying presences of watermarks
  - **Confounded**: 80% Dog, 20% Cat
  - **Balanced**: 50% Dog, 50% Cat
  - **No-watermark**: 0% Dog, 0% Cat
- Regardless of training dataset, watermarked regions are attributed significant importance when watermarks are present in test data
- Further implications for XAI benchmarks and downstream usages

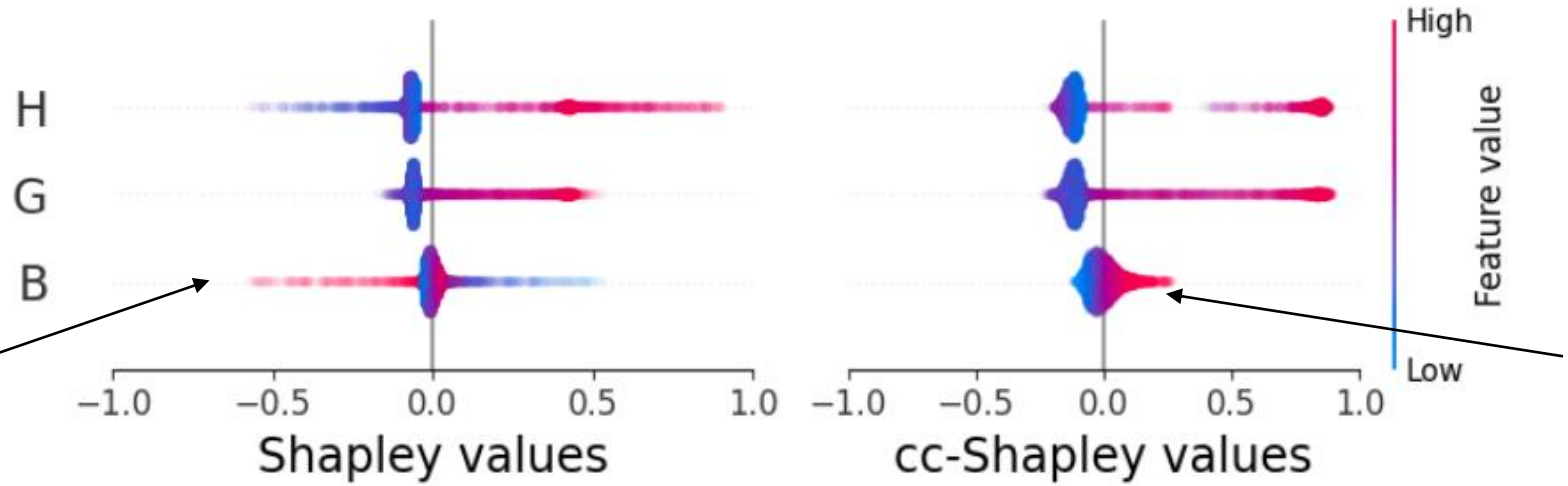
Model	AUROC × 100 on Test Data		
	Confounded	Balanced	No-watermark
Confounded	**92.32** ± 0.006	78.98 ± 0.027	86.32 ± 0.013
Balanced	88.68 ± 0.015	88.59 ± 0.007	**89.24** ± 0.007
No-watermark	84.73 ± 0.033	87.09 ± 0.010	**89.77** ± 0.010



Clark et al., Feature salience - not task-informativeness - drives machine learning model explanations, under review at *Nature Machine Intelligence*, 2026.

Causal context fixes spurious attribution

Spurious:
*„high BMI **lowers** diabetes risk“*



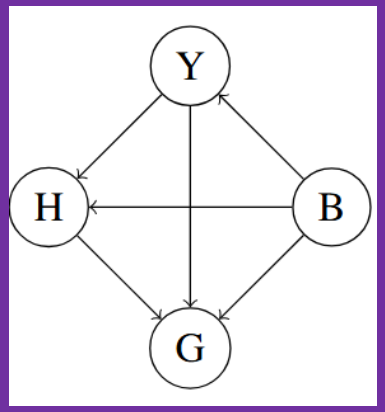
Fixed:
*„high BMI **increases** diabetes risk“*

Shapley values with **c**ausal **c**ontext (cc-Shapley)

$$\phi_{cc}(X_j) = \sum_S \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \cdot \left(E\left[Y \middle| X_j, \boldsymbol{do}(S)\right] - E[Y|\boldsymbol{do}(S)]\right)$$

Causal context

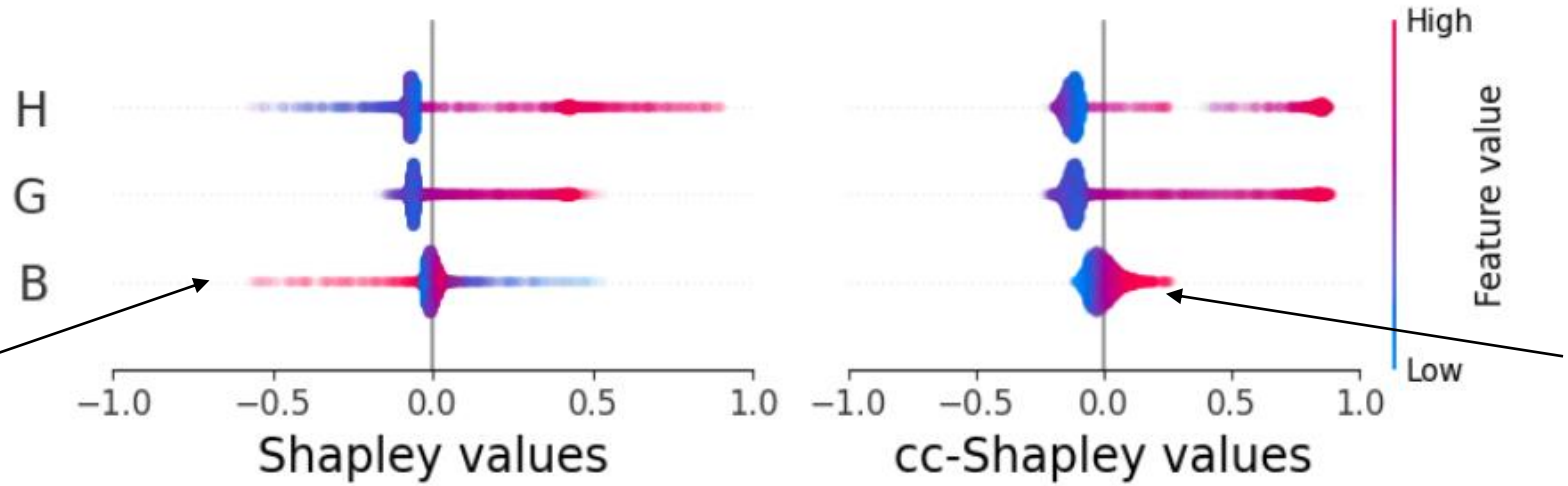
Key ingredient: *causal* relation between features



*Martin, Jörg and Haufe, Stefan. cc-*Shapley: Measuring Multivariate Feature Importance Needs Causal Context. *Preprint arXiv:2602.20396*

Causal context fixes spurious attribution

Spurious:
*„high BMI **lowers** diabetes risk“*

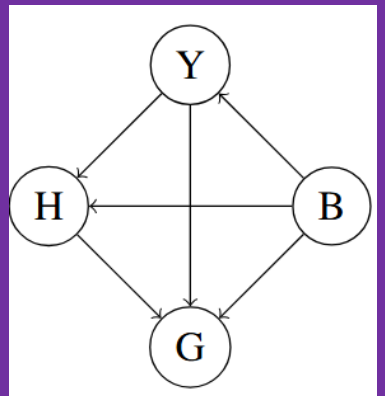


Fixed:
*„high BMI **increases** diabetes risk“*

Shapley values with **c**ausal **c**ontext (cc-Shapley)

$$\phi_{cc}(X_j) = \sum_S \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \cdot \underbrace{\left(E\left[Y \middle| X_j, \boldsymbol{do}(S)\right] - E\left[Y \middle| \boldsymbol{do}(S)\right]\right)}_{\text{Like a univariate fit}}$$

Key ingredient: *causal* relation between features



*Martin, Jörg and Haufe, Stefan. cc-*Shapley: Measuring Multivariate Feature Importance Needs Causal Context. *Preprint arXiv:2602.20396*

Takeaways

- **An SAP-grounded solution**: PatternGAM actively enforces statistical association in the interpretation of the widely popular family of additive models

- **Further risk**: Current XAI methods often measure low-level visual salience rather than true statistical informativeness
  - This limits their reliability for safe model debugging
  - Presents challenges for existing benchmarks

- **Beyond statistical association**: Incorporating causal knowledge of the underlying data generation process and feature interactions is key for correct and consistent explainability

- **A need for formalisation**: Quality assurance of AI through explainability requires moving away from subjective evaluations, and towards adopting rigorous, statistically correct measurement methods

Get in touch!

Me:
Benedict.clark@ptb.de

Stefan Haufe:
haufe@tu-berlin.de

References

Clark, B., Wilming, R., and Haufe, S. XAI-TRIS: non-linear image benchmarks to quantify false positive post-hoc attribution of feature importance. Springer Machine Learning / ECML, 113(9):6871–6910, 2024.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., and Bießmann, F. On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage, 87:96–110, 2014.

Haufe, S., Wilming, R., Clark, B., Zhumagambetov, R., Panknin, D., and Boubekki, A. Position: Xai needs formal notions of explanation correctness. In the NeurIPS Workshop on Interpretable AI: Past, Present and Future, 2024.

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., et al. Mimic-iv, a freely accessible electronic health record dataset. Scientific data, 10(1):1, 2023.

Wilming, R., Kieslich, L., Clark, B., and Haufe, S. Theoretical behavior of XAI methods in the presence of suppressor variables. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pp. 37091–37107. PMLR, 2023.