List of Publications

Journal Articles

1. Heström A, Bommer P L, Burns T F, Lapuschkin S, Samek W and Höhne M-C M (2025).

"Evaluating Interpretable Methods via Geometric Alignment of Functional Distortions".

In: Transactions on Machine Learning Research 2835–8856.

https://github.com/annahedstroem/GEF/ | TMLR Survey Certification

2. Bley F, Lapuschkin S, Samek W and Montavon G (2025).

"Explaining Predictive Uncertainty by Exposing Second-Order Effects".

In: Pattern Recognition 160:111171.

https://github.com/florianbley/XAI-2ndOrderUncertainty

3. Vielhaben J, Lapuschkin S, Montavon G and Samek W (2024).

"Explainable AI for Time Series via Virtual Inspection Layers".

In: Pattern Recognition 150:110309.

https://github.com/jvielhaben/DFT-LRP

4. Becker S, Vielhaben J, Ackermann M, Müller K-R, Lapuschkin S and Samek W (2024).

"AudioMNIST: Exploring Explainable Artificial Intelligence for Audio Analysis on a Simple Benchmark". In: *Journal of the Franklin Institute* 361(1):418–428.

https://github.com/soerenab/AudioMNIST

5. Achtibat R, Dreyer M, Eisenbraun I, Bosse S, Wiegand T, Samek W and Lapuschkin S (2023).

"From attribution maps to human-understandable explanations through Concept Relevance Propagation".

In: *Nature Machine Intelligence* 5(9):1006–1019.

https://github.com/rachtibat/zennit-crp | https://github.com/maxdreyer/crp-human-study

6. Hedström A, Bommer P, Wickstrøm K K, Samek W, Lapuschkin S and Höhne M-C M (2023).

"The Meta-Evaluation Problem in Explainable AI: Identifying Reliable Estimators with MetaQuantus". In: Transactions on Machine Learning Research 2835–8856.

https://github.com/annahedstroem/MetaQuantus

7. Weber L, **Lapuschkin S**, Binder A and Samek W (2023).

"Beyond Explaining: Opportunities and Challenges of XAI-Based Model Improvement".

In: Information Fusion 92:154-176

8. Hedström A, Weber L, Krakowczyk D G, Bareeva D, Motzkus F, Samek W, **Lapuschkin S** and Höhne M-C M (2023).

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

In: Journal of Machine Learning Research 24(34):1–11.

https://github.com/understandable-machine-intelligence-lab/quantus

9. Hofmann S M, Beyer F, **Lapuschkin S**, Golterman O, Loeffler M, Müller K-R, Villringer A, Samek W and Witte A V (2022).

"Towards the Interpretability of Deep Learning Models for Multi-modal Neuroimaging: Finding Structural Changes of the Ageing Brain".

In: NeuroImage 261:119504

10. Ma J, Schneider L, **Lapuschkin S**, Achtibat R, Durchrau M, Krois J, Schwendicke F and Samek W (2022). "Towards Trustworthy AI in Dentistry".

In: Journal of Dental Research 00220345221106086

11. Rieckmann A, Dworzynski P, Arras L, **Lapuschkin S**, Samek W, Onyebuchi A A, Rod N H, Ekstrøm C T (2022).

"Causes of Outcome Learning: A Causal Inference-inspired Machine Learning Approach to Disentangling Common Combinations of Potential Causes of a Health Outcome".

In: International Journal of Epidemiology dyac078.

 $https://github.com/ekstroem/cool \mid https://www.causes of outcome learning.org$

12. Slijepcevic D, Horst F, **Lapuschkin S**, Horsak B, Raberger A-M, Kranzl A, Samek W, Breiteneder C, Schöllhorn W I and Zeppelzauer M (2022).

"Explaining Machine Learning Models for Clinical Gait Analysis".

In: ACM Transactions on Computing for Healthcare 3(2):14:1–27.

https://github.com/sebastian-lapuschkin/explaining-deep-clinical-gait-classification

13. Anders C J, Weber L, Neumann D, Samek W, Müller K-R and **Lapuschkin S** (2022). "Finding and Removing Clever Hans: Using Explanation Methods to Debug and Improve Deep Models". In: *Information Fusion* 77:261–295

14. Sun J, Lapuschkin S, Samek W and Binder A (2022).

"Explain and Improve: LRP-inference Fine-tuning for Image Captioning Models".

In: Information Fusion 77:233-246

15. Samek W, Montavon G, **Lapuschkin S**, Anders C J, and Müller K-R (2021).

"Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications".

In: *Proceedings of the IEEE* 109(3):247–278

16. Yeom S-K, Seegerer P, Lapuschkin S, Binder A, Wiedemann S, Müller K-R and Samek W (2021).

"Pruning by Explaining: A Novel Criterion for Deep Neural Network Pruning".

In: Pattern Recognition 115:107899.

https://github.com/seulkiyeom/LRP_pruning | https://github.com/seulkiyeom/LRP_Pruning_toy_example

17. Aeles J, Horst F, Lapuschkin S, Lacourpaille L, and Hug F (2021).

"Revealing the Unique Features of Each Individual's Muscle Activation Signatures".

In: Journal of the Royal Society Interface 18(174):20200770.

https://github.com/sebastian-lapuschkin/interpretable-emg-signatures

18. Horst F, Slijepcevic D, Zeppelzauer M, Raberger AM, **Lapuschkin S**, Samek W, Schöllhorn WI, Breiteneder C, and Horsak B (2020).

"Explaining Automated Gender Classification of Human Gait".

In: Gait & Posture 81(S1):159-160

19. Hägele M, Seegerer P, **Lapuschkin S**, Bockmayr M, Samek W, Klauschen F, Müller K-R and Binder A (2020).

"Resolving Challenges in Deep Learning-based Analyses of Histopathological Images using Explanation Methods".

In: Scientific Reports 10:6423

20. Alber M, **Lapuschkin S**, Seegerer P, Hägele M, Schütt K T, Montavon G, Samek W, Müller K-R, Dähne S and Kindermans P-J (2019).

"iNNvestigate Neural Networks!".

In: Journal of Machine Learning Research 20(93):1–8.

https://github.com/albermax/innvestigate

21. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W and Müller K-R (2019).

"Unmasking Clever Hans Predictors and Assessing what Machines Really Learn".

In: Nature Communications 10:1069

22. Horst F, Lapuschkin S, Samek W, Müller K-R and Schöllhorn W I (2019).

"Explaining the Unique Nature of Individual Gait Patterns with Deep Learning".

In: Scientific Reports 9:2391.

https://github.com/sebastian-lapuschkin/interpretable-deep-gait

23. Montavon G, Lapuschkin S, Binder A, Samek W and Müller K-R (2017).

"Explaining NonLinear Classification Decisions with Deep Taylor Decomposition".

In: Pattern Recognition 65:211–222.

Pattern Recognition Best Paper Award and Pattern Recognition Medal winner

24. Samek W, Binder A, Montavon G, Lapuschkin S, and Müller K-R (2017).

"Evaluating the Visualization of what a Deep Neural Network has Learned".

In: IEEE Transactions of Neural Networks and Learning Systems

25. Sturm I, Lapuschkin S, Samek W and Müller K-R (2016).

"Interpretable Deep Neural Networks for Single-Trial EEG Classification".

In: Journal of Neuroscience Methods 274:141–145

26. Lapuschkin S, Binder A, Montavon G, Müller K-R and Samek W (2016).

"The Layer-wise Relevance Propagation Toolbox for Artificial Neural Networks".

In: Journal of Machine Learning Research 17(114):1–5.

https://github.com/sebastian-lapuschkin/lrp_toolbox

27. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R and Samek W (2015).

"On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation". In: *PLoS ONE* 10(7):e0130140

Contributions to Conference Proceedings and Workshops

1. Bareeva D, Yolcu GÜ, Hedström A, Wiegand T, Samek W Lapuschkin S (2024).

"Quanda: An Interpretability Toolkit for Training Data Attribution Evaluation and Beyond".

 $In: \textit{NeuRIPS 2024 Workshop on Attributing Model Behavior at Scale (ATTRIB 2024)} \;.$

https://github.com/dilyabareeva/quanda

2. Naujoks J R, Krasowski A, Weckbecker M, Wiegand T, **Lapuschkin S**, Samek W and Klausen R P (2024). "PINNfluence: Influence Functions for Physics-Informed Neural Networks".

In: NeuRIPS 2024 Workshop on Machine Learning and the Physical Sciences (ML4PS).

https://github.com/aleks-krasowski/PINNfluence

Reproducibility Badge Winner

3. Kopf L, Bommer P L, Hedström A, Lapuschkin S, Höhne M M-C and Bykov K (2024).

"CoSy: Evaluating Textual Explanations of Neurons".

In: Advances in Neural Information Processing Systems (NeuRIPS) TBA. (OpenReview)

https://github.com/lkopf/cosy

4. Nobis G, Springenberg M, Aversa M, Detzel M, Daems R, Murray-Smith R, Nakajima S, **Lapuschkin S**, Ermon S, Birdal T, Opper M, Knochenhauer C, Oala L and Samek W (2024).

"Generative Fractional Diffusion Models".

In: Advances in Neural Information Processing Systems (NeuRIPS) TBA. (OpenReview)

https://github.com/GabrielNobis/gfdm

5. Mekala R R, Pahde F, Baur S, Chandrashekar S, Diep M, Wenzel M A, Wisotzky E L, Yolcu G Ü, **Lapuschkin S**, Ma J, Eisert P, Lindvall M, Porter A and Samek W (2024).

"Synthetic Generation of Dermatoscopic Images with GAN and Closed-Form Factorization".

In: ECCV 2024 Workshop on Synthetic Data for Computer Vision (SyntheticData4CV) TBA. (Green Open Access)

6. Achtibat R, Hatefi S M V, Dreyer M, Jain A, Wiegand T, Lapuschkin S, Samek W (2024).

"AttnLRP: Attention-Aware Layer-wise Relevance Propagation for Transformers".

In: Proceedings of the 41st International Conference on Machine Learning (ICML) 135–168.

https://github.com/rachtibat/LRP-for-Transformers

7. Hatefi S M V, Dreyer M, Achtibat R, Wiegand T, Samek W and Lapuschkin S (2024).

"Pruning By Explaining Revisited: Optimizing Attribution Methods to Prune CNNs and Transformers". In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* TBA.

https://github.com/erfanhatefi/Pruning-by-eXplaining-in-PyTorch

8. Hedström A, Weber L, Lapuschkin S, Höhne M M-C (2024).

"A Fresh Look at Sanity Checks for Saliency Maps".

In: Proceedings of the 2nd XAI World Conference 403–420. (Green Open Access)

https://github.com/annahedstroem/sanity-checks-revisited

9. Tinauer C, Damulina A, Sackl M, Soellradl M, Achtibat R, Dreyer M, Pahde F, **Lapuschkin S**, Schmidt R, Ropele S, Samek W, Langkammer C (2024).

"Explainable Concept Mappings of MRI: Revealing the Mechanisms Underlying Deep Learning-based Brain Disease Classification".

In: Proceedings of the 2nd XAI World Conference 202–216. (Green Open Access)

10. Dreyer M, Purelku E, Vielhaben J, Samek W, Lapuschkin S (2024).

"PURE: Turning Polysemantic Neurons Into Pure Features by Identifying Relevant Circuits".

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops 8212–8217.

https://github.com/maxdreyer/PURE | Spotlight Paper

11. Bareeva D, Dreyer M, Pahde F, Samek W and Lapuschkin S (2024).

"Reactive Model Correction: Mitigating Harm to Task-Relevant Features via Conditional Bias Suppression".

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops 3532–3541.

https://github.com/dilyabareeva/reactive correction

12. Dreyer M, Achtibat R, Samek W and Lapuschkin S (2024).

"Understanding the (Extra-)Ordinary: Validating Deep Model Decisions with Prototypical Concept-based Explanations".

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops 3491–3501.

https://github.com/maxdreyer/pcx

- 13. Dreyer M, Pahde F, Anders C J, Samek W and Lapuschkin S (2024).
 - "From Hope to Safety: Unlearning Biases of Deep Models via Gradient Penalization in Latent Space". In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* 38(19):21046–21054. https://github.com/frederikpahde/rrclarc
- 14. Dawoud K, Samek W, Eisert P, Lapuschkin S and Bosse S (2023).
 - "Human-Centered Evaluation of XAI Methods".
 - In: Proceedings of the IEEE International Conference on Data Mining (ICDM) 912–921. (Green Open Access)
- 15. Frommholz A, Seipel F, **Lapuschkin S**, Samek W and Vielhaben J (2023). "XAI-based Comparison of Audio Event Classifiers with different Input Representations".
 - In: Proceedings of the International Conference on Content-based Multimedia Indexing (CBMI) 126–132
- 16. Hedström A, Weber L, **Lapuschkin S** and Höhne M M-C (2023).
 - "Sanity Checks Revisited: An Exploration to Repair the Model Parameter Randomisation Test". In: NeuRIPS 2023 Workshop on XAIX (XAI in Action: Past, Present, and Future Applications) (vVpefYmnsG)
- 17. Pahde F, Dreyer M, Samek W and Lapuschkin S (2023).
 - "Reveal to Revise: An Explainable AI Life Cycle for Iterative Bias Correction of Deep Models".
 - In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention 596–606. (Green Open Access)
 - https://github.com/maxdreyer/reveal2revise
- 18. Binder A, Weber L, Lapuschkin S, Montavon G, Müller K-R and Samek W (2023).
 - "Shortcomings of Top-Down Randomization-Based Sanity Checks for Evaluations of Deep Neural Network Explanations".
 - In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 16143–16152
- 19. Dreyer M, Achtibat R, Wiegand T, Samek W and Lapuschkin S (2023).
 - "Revealing Hidden Context Bias in Segmentation and Object Detection through Concept-specific Explanations".
 - In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops 3828–3838
- 20. Pahde F, Yolcu GÜ, Binder A, Samek W and Lapuschkin S (2023).
 - "Optimizing Explanations by Network Canonization and Hyperparameter Search".
 - In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops 3818–3827
- 21. Krakowczyk D G, Prasse P, Reich D R, Lapuschkin S, Scheffer T, Jäger L A (2023).
 - "Bridging the Gap: Gaze Events as Interpretable Concepts to Explain Deep Neural Sequence Models". In: *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA)* 1–8. *Best Short Paper Award Winner*
- 22. Krakowczyk D G, Reich D R, Prasse P, Lapuschkin S, Jäger L A and Scheffer T (2022).
 - "Selection of XAI Methods Matters: Evaluation of Feature Attribution Methods for Oculomotoric Biometric Identification".
 - In: NeuRIPS 2022 Workshop on Gaze Meets ML (GOLdDAP2AtI)
- 23. Motzkus F, Weber L and Lapuschkin S (2022).
 - "Measurably Stronger Explanation Reliability via Model Canonization".
 - In: Proceedings of the International Conference on Image Processing (ICIP) 516–520
- 24. Ede S, Baghdadlian S, Weber L, Nguyen A, Zanca D, Samek W and Lapuschkin S (2022).
 - "Explain to Not Forget: Defending Against Catastrophic Forgetting with XAI".
 - In: Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE) 1–18. (Gold Open Access link)
- 25. Sun J, **Lapuschkin S**, Samek W, Zhao Y, Cheung N-M and Binder A (2021).
 - "Explanation-Guided Training for Cross-Domain Few-Shot Classification".
 - In: Proceedings of the 25th International Conference on Pattern Recognition (ICPR) 7609–7616
- 26. Goh G S W, Lapuschkin S, Weber L, Samek W and Binder A (2021).
 - "Understanding Integrated Gradients with SmoothTaylor for Deep Neural Network Attribution". In: *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)* 4949–4956
- 27. Kohlbrenner M, Bauer A, Nakajima S, Binder A, Samek W, and Lapuschkin S (2020).
 - "Towards Best Practice in Explaining Neural Network Decisions with LRP".
 - In: Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN) 1-7
- 28. Sun J, Lapuschkin S, Samek W and Binder A (2020).
 - "Understanding Image Captioning Models beyond Visualizing Attention".
 - In: XXAI: Extending Explainable AI Beyond Deep Models and Classifiers. ICML Workshop

29. Anders C J, Neumann D, Marinč T, Samek W, Müller K-R and Lapuschkin S (2020).

"XAI for Analyzing and Unlearning Spurious Correlations in ImageNet".

In: XXAI: Extending Explainable AI Beyond Deep Models and Classifiers. ICML Workshop

30. Sun J, Lapuschkin S, Samek W, Zhao Y, Cheung N-M and Binder A (2020).

"Explain and Improve: Cross-Domain-Few-Shot-Learning Using Explanations".

In: XXAI: Extending Explainable AI Beyond Deep Models and Classifiers. ICML Workshop

31. Alber M, **Lapuschkin S**, Seegerer P, Hägele M, Schütt K T, Montavon G, Samek W, Müller K-R, Dähne S and Kindermans P-J (2018).

"How to iNNvestigate Neural Networks' Predictors!".

In: Machine Learning Open Source Software: Sustainable Communities. NIPS Workshop

32. Lapuschkin S, Binder A, Müller K-R and Samek W (2017).

"Understanding and Comparing Deep Neural Networks for Age and Gender Classification".

In: Proceedings of the ICCV'17 Workshop on Analysis and Modeling of Faces and Gestures (AMFG) 2017:1629-1638

33. Srinivasan V, **Lapuschkin S**, Hellge C, Müller K-R and Samek W (2017).

"Interpretable Action Recognition in Compressed Domain".

In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2017:1692-1696

34. Bach S, Binder A, Müller K-R and Samek W (2016).

"Controlling Explanatory Heatmap Resolution and Semantics via Decomposition Depth".

In: Proceedings of the IEEE International Conference of Image Processing (ICIP) 2016:2271-2275

35. Binder A, Samek W, Montavon G, Bach S, and Müller K-R (2016).

"Analyzing and Validating Neural Network Predictions".

In: Proceedings of the ICML'16 Workshop on Visualization for Deep Learning.

Best Paper Award Winner

36. **Lapuschkin S**, Binder A, Montavon G, Müller K-R and Samek W (2016).

"Analyzing Classifiers: Fisher Vectors and Deep Neural Networks".

In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016:2912-2920

37. Montavon G, Bach S, Binder A, Samek W and Müller K-R (2016).

"Deep Taylor Decomposition of Neural Networks".

In: Proceedings of the ICML'16 Workshop on Visualization for Deep Learning 2016:1-3

38. Samek W, Montavon G, Binder A, Lapuschkin S and Müller K-R (2016).

"Interpreting the Predictions of Complex ML Models by Layer-wise Relevance Propagation".

In: Proceedings of the Interpretable ML for Complex Systems NIPS'16 Workshop

Books

1. Longo L, **Lapuschkin S** and Seifert C, editors (2024).

"Explainable Artificial Intelligence (Second World Conference, xAI 2024, Valletta, Malta, July 17–19, 2024, Proceedings, Part I-IV)".

Springer (Cham), Part I ISBN: 978-3-031-63787-2. Part II ISBN: 978-3-031-63797-1.

Part III ISBN: 978-3-031-63800-8. Part IV ISBN: 978-3-031-63803-9

Book Chapters

1. Becking D, Dreyer M, Samek W, Müller K and Lapuschkin S (2022).

"ECQx: Explainability-Driven Quantization for Low-Bit and Sparse DNNs".

In: xxAI – Beyond Explainable AI 271-296. Springer, Cham

2. Montavon G, Binder A, Lapuschkin S, Samek W and Müller K-R (2019).

"Layer-wise relevance propagation: An Overview".

In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning 193-209. Springer, Cham

3. Binder A, Bach S, Montavon G, Müller K-R and Samek W (2016).

"Layer-wise Relevance Propagation for Deep Neural Network Architectures".

In: *Information Science and Applications (ICISA) 2016. Lecture Notes in Electrical Engineering* 276:913-922. Springer, Singapore

4. Binder A, Montavon G, **Lapuschkin S**, Müller K-R and Samek W (2016).

"Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers".

In: Lecture Notes in Computer Science 9887:63-71. Springer, Berlin/Heidelberg

Preprints

1. Pahde F, Wiegand T, Lapuschkin S and Samek W (2025).

"Ensuring Medical AI Safety: Explainable AI-Driven Detection and Mitigation of Spurious Model Behavior and Associated Data".

In: CoRR abs/2501.13818.

https://github.com/frederikpahde/medical-ai-safety

2. Dreyer M, Berend J, Labarta T, Vielhaben J, Wiegand T, Lapuschkin S and Samek W (2025).

"Mechanistic understanding and validation of large AI models with SemanticLens".

In: CoRR abs/2501.05398.

https://github.com/jim-berend/semanticlens | Demo: https://semanticlens.hhi-research-insights.eu/umap-view

3. Yolcu G Ü, Wiegand T, Samek W and Lapuschkin S (2024).

"DualView: Data Attribution from the Dual Perspective".

In: CoRR abs/2402.12118.

https://github.com/gumityolcu/DualView

4. Weber L, Berend J, Weckbecker M, Binder A, Wiegand T, Samek W and **Lapuschkin S** (2023). "Efficient and Flexible Neural Network Training through Layer-wise Feedback Propagation". In: *CoRR abs*/2308.12053

5. Gerstenberger M, Lapuschkin S, Eisert P and Bosse S (2022).

"But That's Not Why: Inference Adjustment by Interactive Prototype Deselection". In: *CoRR abs*/2203.10087

6. Pahde F, Dreyer M, Weber L, Weckbecker M, Anders C J, Wiegand T, Samek W and **Lapuschkin S** (2022). "Navigating Neural Space: Revisiting Concept Activation Vectors to Overcome Directional Divergence". In: *CoRR abs*/2202.03482.

Accepted for publication at ICLR 2025

7. Anders C J, Neumann D, Samek W, Müller K-R and Lapuschkin S (2021).

"Software for Dataset-wide XAI: From Local Explanations to Global Insights with Zennit, CoRelAy, and ViRelAy".

In: CoRR abs/2106.13200. https://github.com/chr5tphr/zennit | https://github.com/virelay/corelay | https://github.com/virelay/virelay

8. Schwenk G and **Bach S** (2014).

"Detecting Behavioural and Structural Anomalies in Media-Cloud Applications".

In: CoRR abs/1409.8035