Dr. rer. nat. Sebastian Lapuschkin

* December 16, 1986 in Würzburg

Fraunhofer Institut für Nachrichtentechnik, Heinrich-Hertz-Institut, HHI
Einsteinufer 37, 10587 Berlin
http://www.hhi.fraunhofer.de
sebastian.lapuschkin@hhi.fraunhofer.de
+49 (30) 31002-371 • +49 (177) 483-2754

github.com/sebastian-lapuschkin • linkedin.com/in/sebastian-lapuschkin
scholar.google.com/citations?user=wpLQuroAAAAJ

Short Bio

Sebastian Lapuschkin is the Head of the Explainable Artificial Intelligence research group at Fraunhofer Heinrich Hertz Institute (HHI) in Berlin.

He received his Ph.D. degree with distinction from the Technische Universität Berlin in 2018 for his pioneering contributions to the field of Explainable Artificial Intelligence (XAI) and interpretable machine learning. From 2007 to 2013 he studied computer science (B. Sc. and M. Sc.) at the Technische Universität Berlin, with a focus on software engineering and machine learning.

Sebastian is the recipient of multiple awards, including the Hugo-Geiger-Prize for outstanding doctoral achievement and the 2020 Pattern Recognition Best Paper Award.

His research has shaped the field of XAI from the very beginning, with contributions to the first wave of XAI such as the popular and widely-used Layer-wise Relevance Propagation method, as well as timely works influencing the second wave of XAI with additions to the sub-fields of Mechanistic Interpretability, Data Attribution and XAI-based model- and data improvement.

Sebastian is an avid advocate for Open Science, demonstrated by numerous Free Open Source Software toolboxes published with the intent to warrant and facilitate reproducibility in AI research.

Since 2024 he is co-organizing The World Conference on eXplainable Artificial Intelligence and serves as a Topic Editor on "Opportunities and Challenges in Explainable Artificial Intelligence" for the MDPI Open Access Journals.

Further research interests include efficient machine learning and data analysis, as well as data and algorithm visualization.

Professional Experience

Technological University Dublin — Dublin, Ireland

External Scholar *2025 -*
at the Centre of eXplainable Artificial Intelligence. The Centre is the first of its kind in the Republic of Ireland and it aims to increase further and synergise cross-centres, college and external research collaboration.

Multidisciplinary Digital Publishing Institute (MDPI)

Topic Editor *2024 - 2026*
for "Opportunities and Challenges in Explainable Artificial Intelligence".
Submission pre-screening, review management and decision handling.

XAI4Science

Organizer *2024 -*
of the 1st workshop "XAI4Science: From Understanding Model Behavior to Discovering New Scientific Knowledge (2025)", co-located with ICLR 2025 at Singapore EXPO, Singapore.
TBA: 2nd of the workshop at AAAI 2026, Singapore.

World Conference on eXplainable Artificial Intelligence

Steering Committee Member *2024 -*
for the 3rd XAI World Conference (2025) in Istanbul, Turkey, as well as future undisclosed instances of the conference.
Panelist, Conference and Special Track co-organization.

Programme Committee Chair *2023 - 2024*
for the 2nd XAI World Conference (2024) in Valetta, Malta.
Conference and Special Track co-organization.

Fraunhofer Heinrich-Hertz-Institute — Berlin, Germany

Contact Person *2025 -*
for the Erasmus Mundus Joint Master in Intelligent Field Robotic Systems (IFRoS) associate partnership of Fraunhofer HHI.

Ethics Committee Member *2023 -*
Founding member of the first ethics committee at Fraunhofer HHI.
Responsibilities: Fulfilling the task and demand of assessing ethical aspects of research on and with humans, as well as general ethical issues.

Head of Explainable Artificial Intelligence *2021 -*
Research Group Leadership and direction of XAI research & applications.
(current number of staff: 1 PostDocs, 20 PhD researchers, 30+ student research assistants & 3 Technical Staff).

Research: Work towards the next generation of local-global Explainable AI approaches and XAI-based model improvement by, e.g., increasing efficiency (see also) and debugging model training, reasoning and datasets (see also). Provision of powerful modified backprop XAI for Pytorch models, and tools for reproducible XAI evaluations to the community, e.g., via Zennit, LXT and Quantus.

Further responsibilities: Project management and (funding) acquisition. Recruitment and guidance of research personnel. Founding member of the Ethics Committee at Fraunhofer HHI.

Tenured Researcher *2019 - 2020*
PostDoc research position in the Machine Learning Group at Fraunhofer HHI.
Research: Development of Spectral Relevance Analysis, automating the detection of "Clever Hans" moments in machine learning. Measurably increasing the explanation quality of local XAI. Provision of modified backprop XAI in Keras/Tensorflow via iNNvestigate.

Further responsibilities: Project (funding) acquisition. Recruitment and guidance of PhD students and student research assistants.

Research Associate *2014 - 2018*
Founding member of the Machine Learning Group at Fraunhofer HHI.

Research: Furthering XAI research with the development and evaluation of corresponding methods, as well as applications in various expert domains, resulting in several highly cited publications, open source software tools and repositories, and the first recorded encounter of the "Clever Hans" effect in machine learning via XAI.

Other contributions: Contributions to the h.266 (VVC) video codec via learnable intra-frame prediction filters. Planning and conceptualization of an HPC cluster with modern GPU hardware implemented at Fraunhofer HHI. Development and showcasing multiple XAI demos at international events.

Technische Universität Berlin BERLIN, GERMANY
Research Associate *2013 - 2014*
Research: Formalization and development of the "Layer-wise Relevance Propagation" (LRP) method of Explainable AI for explaining individual predictions of nonlinear machine learning models.

Supervision by Prof. Dr. Klaus-Robert Müller.

Student Research- & Teaching Assistant *2011 - 2013*
Research: Structure and cell type detection in large histopathology images using Bag of Words features and SVM classifiers. Development of XAI for the pipeline.

Teaching: Preparation and lecturing (of exercise sessions) in the courses "Machine Learning 1" and "Machine Learning 2 – Theory and Application" and associated academic courses. Visualization and animation of data and learning algorithms discussed throughout the course work.

Research & Teaching assistant to Prof. Dr. Klaus-Robert Müller.

Student Teaching Assistant *2009 - 2011*
Course instruction for algorithmic and practical foundations of computer science (B.Sc.): Basic and advanced Java development, software engineering and OOP concepts, algorithms on image and graph data, among others.

Teaching assistant to Prof. Dr. Marc Alexa, Prof. Dr. Odej Kao and Prof. Dr. Oliver Brock.

Education

Technische Universität Berlin BERLIN, GERMANY
PhD in Machine Learning (*summa cum laude*) *2018*
Date of oral defense: December 19th, 2018.
Dean's signature on Doctorate Certificate dated January 23rd, 2019.
Research and application of methods of *Explainable AI* (*XAI*): Layer-wise Relevance Propagation, Deep Taylor Decomposition and Spectral Relevance Analysis.
Thesis: "Opening the machine learning black box with Layer-wise Relevance Propagation"
Supervision headed by Prof. Dr. Klaus-Robert Müller.

Master of Science in Computer Science *2013*
Focus on machine learning, computer vision and large scale data analysis.
Thesis: "On Pixel-wise Predictions from Image-wise Bag of Words Classification"
Thesis supervision headed by Prof. Dr. Alexander Binder.

Bachelor of Science in Computer Science *2010*
Focus on algorithms and software development
Thesis: "Keyword-Based Image Browsing of Large Image Databases"
Thesis supervision headed by Prof. Dr. Kristian Hildebrand.

Deutschhaus-Gymnasium Würzburg, Germany
Abitur (pre-university secondary education) *2007*

Teaching

See section "Talks & Lectures / Invited Lectures" below for a list of additional invited and individual lectures held.

WS 24/25 Machine Learning Seminar.
[**Universitat de Girona / IFRoS.** Guest Lecturer. Interactive Block Seminar "An Introduction to Explainable AI", 1 full week.]

WS 23/24 Machine Learning Seminar.
[**Universitat de Girona / IFRoS.** Guest Lecturer. Interactive Block Seminar "Explainable AI", 1 full week.]

WS 23/24 Responsible Artificial Intelligence 1.
[**Technische Universität Berlin.** Co-Teaching, Lecture Design, Interactive Coding Sessions.]

WS 21/22 Machine Learning Seminar.
[**Universitat de Girona.** Guest Lecturer. Interactive Block Seminar "Explainable AI", 1 full week.]

SS 17 Seminar Cognitive Algorithms (block seminar).
[**Technische Universität Berlin.** 1:1 Student Guidance and Co-Supervision, Grading.]

WS 13/14 Python Programming for Machine Learning (block seminar).
[**Technische Universität Berlin.** Co-Teaching, Grading, Exercise Design.]
Matlab Programming for Machine Learning and Data Analysis (block beminar).
[**Technische Universität Berlin.** Co-Teaching, Grading, Exercise Design.]

SS13 Integrated Lecture Machine Learning II.
[**Technische Universität Berlin.** Teaching (Exercise sessions), Grading, Exercise Design.]
Python Programming for Machine Learning (block seminar).
[**Technische Universität Berlin.** Co-Teaching, Grading, Exercise Design.]

WS 12/13 Integrated Lecture Machine Learning I.
[**Technische Universität Berlin.** Teaching (Exercise sessions), Grading, Exercise Design.]
Matlab Programming for Machine Learning and Data Analysis (block beminar).
[**Technische Universität Berlin.** Co-Teaching, Grading, Exercise Design.]

SS12 Machine Learning II – Theory and Application.
[**Technische Universität Berlin.** Teaching (Exercise sessions), Grading, Exercise Design.]
Matlab Programming for Machine Learning and Data Analysis (block beminar).
[**Technische Universität Berlin.** Co-Teaching, Grading, Exercise Design.]

WS 11/12 Machine Learning I.
[**Technische Universität Berlin.** Teaching (Exercise sessions), Grading, Exercise Design.]
Matlab Programming for Machine Learning and Data Analysis (block beminar).
[**Technische Universität Berlin.** Co-Teaching, Grading, Exercise Design.]

SS 11 Methodisch-praktische Grundlagen der Informatik 2 ("Algorithms and Data Structures").
[**Technische Universität Berlin.** Teaching (Exercise sessions), Grading, Exercise Design.]

WS 10/11 Methodisch-praktische Grundlagen der Informatik 4 ("Advanced Algorithms").
[**Technische Universität Berlin.** Teaching (Exercise sessions), Grading, Exercise Design.]

SS 10 Methodisch-praktische Grundlagen der Informatik 2 ("Algorithms and Data Structures").
[**Technische Universität Berlin.** Teaching (Exercise sessions), Grading, Exercise Design.]

WS 09/10 Methodisch-praktische Grundlagen der Informatik 4 ("Advanced Algorithms").
[**Technische Universität Berlin.** Teaching (Exercise sessions), Grading, Exercise Design.]

Talks & Lectures

Talks *excludes internal/confidential events*

1. "Explainability in the Era of LLMs: New Challenges and Pathways to Actionable Insights" (2025-08-14).
   2025 Workshop on Self-Supervised Learning for Signal Decoding, Aalborg, Denmark, (*invited talk*)
2. "Decoding AI: How Explainability Unlocks Actionable Insights" (2025-05-23).
   dida conference 2025, Berlin, Germany, (*invited talk*)
3. "Artificial Intelligence We Can Trust – From Explainable to Actionable and Regenerative AI" (2024-02-02).
   MPNE Consensus 2024 Workshop, Berlin, Germany, (*invited talk*)
4. "From Concepts to Prototypes – Towards Minimal Effort Post-Hoc Interpretability" (2024-01-12).
   2nd Machine Teaching for XAI Workshop (*MT4XAI*), *Valencia, Spain,* (*invited talk*)
5. "Explaining AI with Concept Relevance Propagation" (2023-10-06).
   4th Japanese-American-German Frontiers of Science (*JAGFOS*) *Symposium, Dresden, Germany,* (*flash talk & poster, invited*)
6. "Model-Assisted Data Analysis via XAI" (2023-07-05).
   19th Machine Learning in Healthcare Meetup Berlin, Berlin Institute of Health, (*invited talk*)
7. "Accessing the Hidden Space with Explainable Artificial Intelligence" (2023-06-27).
   Informatik-Kolloquium, Universität Bremen, (*invited talk*)
8. "Explainable AI and Beyond with Concept Relevance Propagation" (2023-05-24).
   Data Professional Days / Data4Business Days Köln, (*keynote*)
9. "Beyond Heatmaps – Explaining with Concepts" (2022-10-21).
   BIFOLD Graduate School Welcome Days, (*invited talk*)
10. "Explain to Not Forget: Defending Against Catastrophic Forgetting with XAI " (2022-08-24).
    CD-MAKE 2022, (*paper presentation*)
11. "Zukünftige Trends in der KI und Einsatzmöglichkeiten im Bauwesen" (2022-06-24).
    BIMKIT Jahresveranstaltung 2022, (*keynote*)
12. "Beyond Explaining" (2021-06-03).
    Melanoma Patient Network Europe Meet-up – MPNE meets AI , (*invited talk*)
13. "Beyond Explaining: Explainable AI for Model Improvement" (2021-05-05).
    Sensor and Measurement Science International 2021, (*invited talk*)
14. "Efficient and Effective Neural Network Pruning with Layer-wise Relevance Propagation" (2020-11-12).
    Machine Learning Seminar at Fraunhofer HHI / Technische Universität Berlin
15. "Towards Best Practice in Explaining Neural Network Decisions with LRP" (2020-07-21).
    IEEE World Congress on Computational Intelligence 2020 / IJCNN 2020
16. "XAI for Analyzing and Unlearning Spurious Correlations in ImageNet" (2020-07-18).
    XXAI: Extending Explainable AI Beyond Deep Models and Classifiers, (*ICML 2020 Workshop*)
17. "XAI via LRP and SpRAy" (2020-07-02).
    Ada Day at Ada Lovelace Center / Fraunhofer IIS, (*invited talk*)
18. "Interpretable Machine Learning through Layer-wise Relevance Propagation" (2020-02-18).
    Fraunhofer Symposium Netzwert 2020
19. "Interpretable Machine Learning through Layer-wise Relevance Propagation" (2019-12-12).
    Gesellschaft von Freunden des HHI e.V.
20. "Explainable Artificial Intelligence — Opening the Machine Learning Black Box with Layer-wise Relevance Propagation" (2019-09-26).
    AMA Wissenschaftsrat 2019, (*invited talk*)
21. "Finding Clever Hans" (2019-07-16).
    Universität Bamberg, (*invited talk & press interview*)
22. "AI – Opening the Black Box" (2019-02-25).
    Robert Koch Institut, (*invited talk*)
23. "AI – Opening the Black Box" (2019-02-22).
    Technology Innovation Day – 91 Years HHI

24. "Understanding and Comparing Deep Neural Networks for Age and Gender Classification" (2017-10-27).
ICCV'17 Workshop on Analysis and Modeling of Faces and Gestures
25. "Layer-wise Relevance Propagation" (2014-09-10).
IDA Retreat'14

Invited Lectures *Individual Lectures as Parts of Seminars and Workshops*

1. "XAI as a Tool Beyond Model Understanding – From Heatmaps to Concepts and XAI Automation" (2024-11-27).
CBS CoCoNUT | Max Planck Institute for Human Cognitive and Brain Sciences | Leipzig
2. "Human-Understandable Explanations through Concept Relevance Propagation" (2023-01-12).
Machine Teaching for Humans Workshop, Madeira | University of Bergen, (invited, keynote)
3. "Towards Human-understandable Explanations with XAI 2.0" (2022-10-24).
AI4Good webinar series of the International Telecommunication Union (ITU), (streaming link)
4. "Towards Actionable XAI" (2022-09-27).
International Artificial Intelligence Doctoral Academy, (link to slides and video)
5. "Recent Advances in Explainable AI" (2022-09-08).
BB-KI-Chips Summer School Potsdam | Universität Potsdam
6. "XAI BEYOND EXPLAINING: Using Explainability for Improving Deep Machine Learning Models" (2021-08-27).
2nd Summer School on Machine Learning in Bioinformatics | Higher School of Economics Moscow, (link to video)
7. "Neuronale Netze mit LRP (richtig) erklären" (2020-08).
KI-Campus | Die Lernplattform für Künstliche Intelligenz
8. "Explainable Artificial Intelligence — Opening the Machine Learning Black Box with Layer-wise Relevance Propagation" (2019-08-16).
SIMULA Summer School on Smart cities for a Sustainable Energy Future - From Design to Practice

Supervision & Guidance

Collaboration with and (co-)supervision of 3 (Senior) Technical Staff, 2 PostDocs, 32 PhD Students, 40+ Master's Students, 3 Bachelor's Students and 9 Guest Researchers since 2017

Third-Party Funded Research Projects

xJuRAG *2025/10 - 2028/09*
Explainable AI für juristische RAG-Anwendungen. leading role.

EPIBMI *2025/08 - 2026/10*
Explainable & Physics-Informed-based Model Improvements. leading role.

ACHILLES *2024/12 - 2028/11*
Human-Centred Machine Learning: Lighter, Clearer, Safer. leading role.

TEMA *2022/12 - 2026/11*
Trusted Extremely Precise Mapping and Prediction for Emergency Management. leading role.

DAKI-FWS *2021/12 - 2024/22*
Data- and AI-supported Early Warning System

iToBoS *2021/04 - 2025/03*
Intelligent Total Body Scanner. leading role.

BerDiBa *2021/01 - 2024/07*
Berlin Digital Rail Operations

Patho234 *2020/01 - 2022/12*
Machine Learning-driven Multidimensional Imaging Analysis of Reactive and Neoplastic Lymph Nodes

TraMeExCo *2018/09 - 2021/08*
Transparent Medical Expert Companion

Honors & Awards

Machine Learning and the Physical Sciences Reproducibility Badge *2024*
For the paper "PINNfluence: Influence Functions for Physics-Informed Neural Networks"

Stanford Top 2% Scientist Worldwide* *2022 - 2025*
Among the 2% most impactful researchers of 2022$^{(\text{ranked } 195{,}784^{\text{th}})}$ – 2025$^{(\text{ranked } 97{,}340^{\text{th}})}$
*) which is to be taken with a grain of salt.

Best Short Paper Award @ ETRA *2023*
The ACM Symposium for Eye Tracking Research and Applications

Pattern Recognition Best Paper Award and Pattern Recognition Medal *2020*
For the paper "Explaining NonLinear Classification Decisions with Deep Taylor Decomposition"

Hugo-Geiger-Prize (1st place) *2019*
Förderpreis für herausragende Promotionsleistungen

Freunde des HHI Nachwuchspreis *2019*
Förderpreis für exzellente wissenschaftliche Arbeiten am HHI

ERCIM Cor van Baayen Award (finalist) *2019*
Cor Baayen Young Researcher Award

Best Paper Award @ ICMLW *2016*
ICML'16 Workshop on Visualization for Deep Learning

Patents

Attention Head Functionalities in Machine Learning Models *2025*

TBA "Attention Head Functionalities in Machine Learning Models" published 2025-05-15

Relevance Score Assignment dealing with an Attention Module and Applications thereof *2024*

WO 2025163210 A1 "Relevance Score Assignment dealing with an Attention Module and Applications thereof" published 2025-02-03

A Concept Representation of a Machine Learning Model *2024*

TBA "A concept representation of a machine learning model" published 2024-12-11

Analyzing an Inference of a Machine Learning Predictor *2023*

WO 2023237560 A1 "Analyzing an Inference of a Machine Learning Predictor" published 2023-12-14

Method and System for Simulating an Optical Image of a Photonic and/or Electronic Device *2022*

EP 4001902 A1 "Method and System for Simulating an Optical Image of a Photonic and/or Electronic Device" published 2022-05-25

Pruning and/or Quantizing Machine Learning Predictors *2020*

EP 3991102 A1 "Pruning and/or Quantizing Machine Learning Predictors" published 2022-05-04

US 2022/0114455 A1 "Pruning and/or Quantizing Machine Learning Predictors" published 2022-04-14

WO 2020/260656 A1 "Pruning and/or Quantizing Machine Learning Predictors" published 2020-12-30

Relevance Score Assignment for Artificial Neural Networks *2016*

US 20180018553 "Relevance Score Assignment for Artificial Neural Networks " granted 2024-10-04

CN 107636693 "Relevance Score Assignment for Artificial Neural Networks" granted 2022-01-11

EP 3271863 "Relevance Score Assignment for Artificial Neural Network" granted 2021-07-28

JP 6725547 "Relevance Score Assignment for Artificial Neural Networks" granted 2020-07-22

KR 102130162 "Assignment of Relevance Scores for Artificial Neural Networks" granted 2020-07-06

CA 2979579 "Relevance Score Assignment for Artificial Neural Networks" granted 2020-02-18

RU 2703343 "Relevancy Assessment for Artificial Neural Networks" granted 2019-10-16

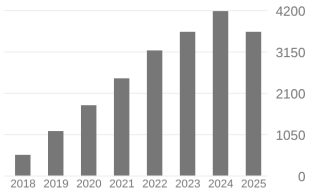
BR 112017019821 "Relevance Score Assignment for Artificial Neural Networks " published 2018-05-15

WO 2016150472 A1 "Relevance score assignment for artificial neural network" published 2016-09-29

Publications

Summary of Scientific Impact

	All	Since 2020
# Publications	100	80
# Citations	21211	19021
h-index	38	37
i10-index	63	62



per Google Scholar, retreived on November 18th, 2025.

List of Publications

Journal Articles

1. 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: *Nature Machine Intelligence* 1–14. *https://github.com/jim-berend/semanticlens | Demo: https://semanticlens.hhi-research-insights.eu*
2. 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: *Machine Learning* 114(9):206. *https://github.com/frederikpahde/medical-ai-safety*
3. Ma J, Weicken E, Pahde F, Weitz K, **Lapuschkin S**, Samek W and Wiegand T (2025). "Künstliche Intelligenz auf dem Prüfstand: Anforderungen, Qualitätskriterien und Prüfwerkzeuge für medizinische Anwendungen [*Artificial intelligence under scrutiny: requirements, quality criteria, and testing tools for medical applications*]". In: *Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz* 68:915-923
4. Storås A M, Dreyer M, Pahde F, **Lapuschkin S**, Samek W, Halvorsen P, de Lange T, Mori Y, Hann A, Berzin T M, Parasa S and Riegler M A (2025). "Exploring the Clinical Value of Concept-based AI Explanations in Gastrointestinal Disease Detection". In: *Scientific Reports* 15(1):28860. *https://github.com/AndreaStoraas/conceptXAI-GItract*
5. Weber L, Berend J, Weckbecker M, Binder A, Wiegand T, Samek W and **Lapuschkin S** (2025). "Efficient and Flexible Neural Network Training through Layer-wise Feedback Propagation". In: *Transactions on Machine Learning Research* 9oToxYVOSW. *https://github.com/leanderweber/layerwise-feedback-propagation*
6. Hedström 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* ukLxqA8zXj. *https://github.com/annahedstroem/GEF/ | TMLR Survey Certification*
7. 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*
8. 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*
9. 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*
10. 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*
11. 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* j3FK00HyfU.
https://github.com/annahedstroem/MetaQuantus

12. 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
13. 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
14. 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
15. Ma J, Schneider L, **Lapuschkin S**, Achtibat R, Duchrau M, Krois J, Schwendicke F and Samek W (2022).
"Towards Trustworthy AI in Dentistry".
In: *Journal of Dental Research* 00220345221106086
16. 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 | https://www.causesofoutcomelearning.org
17. 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
18. 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
19. 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
20. 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
21. 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
22. 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
23. 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
24. 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
25. 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

26. **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
27. 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
28. 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
29. 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*
30. 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
31. **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
32. **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. Labarta T, Hoang N, Weitz K, Samek W, **Lapuschkin S** and Weber L (2025).
"See What I Mean? CUE: A Cognitive Model of Understanding Explanations".
In: *Proceedings of the IJCAI Workshops 2025: XAI Workshop* .
https://arxiv.org/abs/2506.14775
2. Puri B, Jain A, Golimblevskaia E, Kahardipraja P, Wiegand T, Samek W and **Lapuschkin S** (2025).
"FADE: Why Bad Descriptions Happen to Good Features".
In: *Findings of the Association for Computational Linguistics* (*ACL*) 17138–17160.
https://github.com/brunibrun/FADE
3. Naujoks J, Krasowski A, Weckbecker M, Yolcu G Ü, Wiegand T, **Lapuschkin S**, Samek W and Klausen R P (2025).
"Leveraging Influence Functions for Resampling Data in Physics-Informed Neural Networks".
In: *Proceedings of the 3rd XAI World Conference* TBA.
https://github.com/aleks-krasowski/PINNfluence_resampling
4. Erogullari E, **Lapuschkin S**, Samek W and Pahde F (2025).
"Post-Hoc Concept Disentanglement: From Correlated to Isolated Concept Representations".
In: *Proceedings of the 3rd XAI World Conference* TBA.
https://github.com/erenerogullari/cav-disentanglement
5. Joseph S, Suresh P, Hufe L, Stevinson E, Graham R, Vadi Y, Bzdok D, **Lapuschkin S**, Sharkey L and Richards A (2025).
"Prisma: An Open Source Toolkit for Mechanistic Interpretability in Vision and Video".
In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*) *Workshops: MIV Workshop* TBA.
https://arxiv.org/abs/2504.19475 | https://github.com/Prisma-Multimodal/ViT-Prisma
6. Pahde F, Dreyer M, Weckbecker M, Weber L, Anders C J, Wiegand T, Samek W and **Lapuschkin S** (2025).
"Navigating Neural Space: Revisiting Concept Activation Vectors to Overcome Directional Divergence".
In: *Proceedings of the International Conference on Learning Representations* (*ICLR*) .
https://github.com/frederikpahde/pattern-cav
7. Bareeva D, Yolcu GÜ, Hedström A, Wiegand T, Samek W and **Lapuschkin S** (2024).
"Quanda: An Interpretability Toolkit for Training Data Attribution Evaluation and Beyond".
In: *NeuRIPS 2024 Workshop on Attributing Model Behavior at Scale* (*ATTRIB 2024*) .
https://github.com/dilyabareeva/quanda

8. 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
9. 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*) 34656–34685. (*OpenReview*)
https://github.com/lkopf/cosy
10. 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*) 25469–25509. (*OpenReview*)
https://github.com/GabrielNobis/gfdm
11. 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*) 15642:368–384. (*Green Open Access*)
12. 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
13. 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* 152–169. (*Green Open Access*)
https://github.com/erfanhatefi/Pruning-by-eXplaining-in-PyTorch
14. 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
15. 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*)
16. 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
17. 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
18. 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
19. 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
20. 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*)

21. 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
22. 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)
23. 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* (*MICCAI*) 596–606. (*Green Open Access*)
https://github.com/maxdreyer/reveal2revise
24. 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
25. 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
26. 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
27. 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
28. 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)
29. 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
30. 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*)
31. 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
32. 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
33. 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
34. 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*
35. 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*
36. 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*
37. 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*

38. **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
39. 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
40. **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
41. 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
42. **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
43. 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
44. 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).
"ECQ^{x}: 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. Puri B, Berend J, **Lapuschkin S** and Samek W (2025).
"Atlas-Alignment: Making Interpretability Transferable Across Language Models".
In: *CoRR abs/2510.27413*
2. Klausen R P, Timofeev I, Frank J, Naujoks J, Wiegand T, **Lapuschkin S** and Samek W (2025).
"LieSolver: A PDE-constrained solver for IBVPs using Lie symmetries".
In: *CoRR abs/2510.25731*
3. Golimblevskaia E, Jain A, Puri B, Ibrahim A, Samek W and **Lapuschkin S** (2025).
"Circuit Insights: Towards Interpretability Beyond Activations".
In: *CoRR abs/2510.14936*.
https://github.com/egolimblevskaia/WeightLens | https://github.com/egolimblevskaia/CircuitLens

4. Komorowski P, Golimblevskaia E, Achtibat R, Wiegand T, **Lapuschkin S** and Samek W (2025).
"Attribution-Guided Decoding".
In: *CoRR abs/2509.26307*.
https://github.com/piotr-komorowski/attribution-guided-decoding
5. Panfilov A, Kortukov E, Nikolić K, Bethge M, **Lapuschkin S**, Samek W, Prabhu A, Andriushchenko M, Geiping J (2025).
"Strategic Dishonesty Can Undermine AI Safety Evaluations of Frontier LLMs".
In: *CoRR abs/2509.18058*
6. Sandmann E, **Lapuschkin S** and Samek W (2025).
"Iterative Inference in a Chess-Playing Neural Network".
In: *CoRR abs/2508.21380*.
https://github.com/hartigel/leela-logit-lens
Accepted for publication at NeurIPS 2025 Workshops
7. Hufe L, Venhoff C, Dreyer QM, **Lapuschkin S** and Samek W (2025).
"Towards Mechanistic Defenses Against Typographic Attacks in CLIP".
In: *CoRR abs/2508.20570*
8. Hatefi S M V, Dreyer M, Achtibat R, Kahardipraja P, Wiegand T, Samek W and **Lapuschkin S** (2025).
"Attribution-guided Pruning for Compression, Circuit Discovery, and Targeted Correction in LLMs".
In: *CoRR abs/2506.13727*.
https://github.com/erfanhatefi/SparC3
9. Cantú E D, Wittmann R K, Abdeen O, Wagner P, Samek W, Baier M and **Lapuschkin S** (2025).
"Deep Learning-based Multi Project InP Wafer Simulation for Unsupervised Surface Defect Detection".
In: *CoRR abs/2506.10713*
10. Gururaj S, Grüne L, Samek W, **Lapuschkin S** and Weber L (2025).
"Relevance-driven Input Dropout: an Explanation-guided Regularization Technique".
In: *CoRR abs/2505.21595*.
https://github.com/Shreyas-Gururaj/LRP_Relevance_Dropout
11. Dreyer M, Hufe L, Berend J, Wiegand T, **Lapuschkin S** and Samek W (2025).
"From What to How: Attributing CLIP's Latent Components Reveals Unexpected Semantic Reliance".
In: *CoRR abs/2505.20229*.
https://github.com/maxdreyer/attributing-clip
12. Kahardipraja P, Achtibat R, Wiegand T, Samek W and **Lapuschkin S** (2025).
"The Atlas of In-Context Learning: How Attention Heads Shape In-Context Retrieval Augmentation".
In: *CoRR abs/2505.15807*.
https://github.com/pkhdipraja/in-context-atlas
Accepted for publication at NeurIPS 2025
13. Bareeva D, Höhne M M C, Warnecke A, Pirch L, Müller K-R, Rieck K, **Lapuschkin S** and Bykov K (2025).
"Manipulating Feature Visualizations with Gradient Slingshots".
In: *CoRR abs/2401.06122*.
https://github.com/dilyabareeva/grad_slingshot
Accepted for publication at NeurIPS 2025
14. Zverev E, Kortukov E, Panfilov A, Volkova A, Tabesh S, **Lapuschkin S**, Samek W and Lampert C H (2025).
"ASIDE: Architectural Separation of Instructions and Data in Language Models".
In: *CoRR abs/2503.10566*
15. Arras L, Puri B, Kahardipraja P, **Lapuschkin S** and Samek W (2025).
"A Close Look at Decomposition-based XAI-Methods for Transformer Language Models".
In: *CoRR abs/2502.15886*
16. Yolcu G Ü, Weckbecker M, Wiegand T, Samek W and **Lapuschkin S** (2024).
"DualXDA: Towards Sparse, Efficient and Explainable Data Attribution in Large AI Models".
In: *CoRR abs/2402.12118*.
https://github.com/gumityolcu/DualXDA
17. 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*
18. 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*

19. Schwenk G and **Bach S** (2014).
"Detecting Behavioural and Structural Anomalies in Media-Cloud Applications".
In: *CoRR abs/1409.8035*