Dr. rer. nat. Sebastian Lapuschkin (*né Bach*)

dob: December 16, 1986 in Würzburg • nationality: german
address: Kaiserin-Augusta-Allee 92, 10589 Berlin, Berlin, Germany
sebastian@lapuschkin.com • +49 (177) 483-2754 • Google Scholar
github.com/sebastian-lapuschkin • linkedin.com/in/sebastian-lapuschkin

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 Berlin Institute of Technology 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 Berlin Institute of Technology, 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 work is focused on pushing the boundaries of XAI, e.g, for achieving human-understandable explanations, and towards the effective and efficient utilization of interpretable feedback for the improvement of machine learning systems and data.

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

Professional Experience

Fraunhofer Heinrich-Hertz-Institute — Berlin, Germany

Head of Explainable Artificial Intelligence — *Jan '21 – today*

Research Group Leadership and direction of XAI research.
(current number of staff: 2 PostDocs, 17 PhD researchers & 15 student research assistants).

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 via Zennit 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 — *Jan '19 – Dec '20*

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 — *Oct '14 – Dec '18*

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.

Berlin Institute of Technology — Berlin, Germany

Research Associate — *Sep '13 – Sep '14*

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 and Prof. Dr. Alexander Binder.

Student Research- & Teaching Assistant *Oct '11 – Aug '13*

Research: Structure and cell type detection in large histopathology images using Bag of Words features and SVM classifiers. Development of XAI for the pipeline.
Research assistant to Prof. Dr. Alexander Binder at the machine learning group at TU Berlin.

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.

Teaching assistant to Prof. Dr. Klaus-Robert Müller, Prof. Dr. Dr. Franz Király, Dr. Irene Dowding (née Winkler) and Dr. Daniel Bartz.

Student Teaching Assistant *Oct '09 – Sep '11*

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

Berlin Institute of Technology — Berlin, Germany

PhD in Machine Learning (with distinction / "summa cum laude") *2013 – 2018*

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 *2010 – 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 *2007 – 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) *1998 – 2007*

Teaching

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

WS 23/24 Machine Learning Seminar.
[**Universitat de Girona.** 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.]

Supervision & Guidance

PostDocs

Dr. R P Klausen (In own group, 2023-). XAI in Physics-Informed Neural Networks

Dr. E Schnoor (In own group, 2023-). XAI in Reinforcement Learning

PhD Students

J Berend (In own group, 2024-*). Latent Spare Restructuring & XAI-based NN Training

E Golimblevskaia (In own group, 2024-*). TBD

B Puri (Technische Universität Berlin, 2024-*, co-supervision). XAI in Natural Language Processing

A Jain (In own group, 2024-*). XAI for Language Transformers

M Weckbecker (In own group, 2024-*). Data-local XAI & Model Correction

G Ü Yolcu (In own group, 2023-*). Data-local & Data-global XAI

O Böhler (In own group, 2023-*). XAI-guided Dark Matter Exploration, Applications of XAI

P Kahardipraja (In own group, 2023-*). XAI in NLP with Transformers

D Bareeva (In own group, 2023-*). Reactive-ClArC for Model Self-Correction

J Naujoks (In own group, 2023-*). Quantum-XAI

P Reinfeld (MPI for CBS, 2023-*, co-supervision). Analysis of Heart-Brain Interactions in ECG and BCI

G Nobis (In own group, 2022-*). XAI in GANs & XAI for Anomaly Detection

R Achtibat (In own group, 2022-*). Methods for Global-Local Concept-based Explanations

M Dreyer (In own group, 2022-*). GLocal XAI for Localization Models, Anomalies and Model Improvement

D Krakowczyk (Universität Potsdam, 2022-*, co-supervision). XAI Quantification (Biomedical Domain)

A Hill (TU Berlin, 2021-*, co-supervision). Open Source Software Frameworks for XAI

A Hedström (Universität Potsdam, 2021-*, co-supervision). Quantification and Evaluation of XAI

F Pahde (In own group, 2021-*). XAI as Part of the AI Life Cycle

J Vielhaben (In own group, 2020-*). Global XAI & Explaining with Concepts

L Weber (In own group, 2020-*). XAI for ML Model Improvement

S M Hofmann (MPI for CBS, 2018-*, co-supervision). XAI for Brain Age Prediction

S Ede (TU Berlin, 2021-2022, co-supervision). XAI in Defense against Catastrophic Forgetting

P Seegerer (TU Berlin, 2017-2021, co-supervision). XAI for Histopathology & Neural Network Pruning

M Hägele (TU Berlin, 2017-2021, co-supervision). XAI in Histopathology

V Srinivasan (Berlin Big Data Center/own group, 2017, co-supervision). Compressed Domain XAI

Master's Students

G Wassiltschuk (In own group, 2024-*). XAI in Physics-Informed Neural Networks

N Hoang (In own group, 2024-*). UI / UX for Effective XAI

P Klinke (In own group, 2024-*). TBD

O Bouanani (In own group, 2024-*). TBD

E Hartig (In own group, 2024-*). Interactive XAI

J Said (In own group, 2024-*). TEMA Project Support + TBD

T Labarta (In own group, 2024-*). XAI Demonstrators & Automation

M A Pham (In own group, 2024-*). Interactive Systems in XAI

E Beceren (In own group, 2024). User Studies in XAI

Y Poupart (In own group, 2023). Explaining Strategic Gameplay in RL

Z Schellin (In own group, 2023-*). Explaining Strategic Gameplay in RL

E Purellku (In own group, 2023-*). Improving Interpretability of Latent Representations

S Gururaj (In own group, 2022-*). XAI for Point Cloud Recognition Models & XAI-guided Input Level Dropoput

S M V Hatefi (In own group, 2022-*). XAI for Adversarial Attack and Outlier Detection

L Hufe (In own group, 2022-*). XAI-based Neural Network Training for Active Learning

J Berend (In own group, 2022-2024). XAI Demonstrators & XAI-based Neural Network Training

A Jain (In own group, 2022-2024). XAI for Language Transformers

L Bürger (In own group, 2023). Relevance Subgraph Disentanglement

N Ayday (In own group, 2022-2023). Kernelized Class Artifact Compensation

B Puri (In own group, 2022-2023). XAI in Natural Language Processing

G Ü Yolcu (In own group, 2022-2023). Neural Network Canonization & Data-Attribution

A Frommholz (In own group, 2020-2021, 2022-2023). XAI in the Audio Domain

E Dolgener-Cantú (In own group, 2020-2023). AI Simulations for Anomaly Detecion in Photonic Wafers

I Eisenbraun (In own group, 2020-2023). Tools for Interactive GLocal XAI

R Achtibat (In own group, 2020-2022). Human-understandable Explanations [...]

M Dreyer (In own group, 2020-2022). Concept-based XAI for Semantic Segmentation [...]

S Baghdadlian (In own group, 2021). XAI in Defense against Catastrophic Forgetting

F W Motzkus (In own group, 2020-2021). XAI Quantification for Measurably Better Explanations

L Weber (In own group, 2019-2020). XAI for Refined Neural Network Training

Bachelor's Students

L Scharff (In own group, 2024-*). XAI Demonstrators and Automation

M Kohlbrenner (In own group, 2017-2020). Composite Rule Sets for Layer-wise Relevance Propagation

Guest Researchers

C I López González (Universidad Carlos III de Madrid, 2023, co-supervision). Factor Analysis of Latent Embeddings

L Tětková (Technical University of Denmark, 2023, co-supervision). Assessment of (Expected) Concept-based Explanations

P M Miguélez (Universidad Carlos III de Madrid, 2022-2023, co-supervision). Evaluating Local XAI against Global Explanations from Polynomial Representations of Neural Network Models

A Wolny (Universität Heidelberg | EMBL, 2021, co-supervision). Understanding Self-Hallucinations in Neural Networks for Cell Segmentation with Tools from XAI

Third-Party Funded Research Projects

TEMA (2022-12 – 2026-11; leading role) — Trusted Extremely Precise Mapping and Prediction for Emergency Management

DAKI-FWS (2021-12 – 2024-11) — Data- and AI-supported Early Warning System

iToBoS (2021-04 – 2025-03; leading role) — Intelligent Total Body Scanner

BerDiBa (2021-01 – 2023-12) — 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; leading role) — Transparent Medical Expert Companion

---

Awards

Stanford Top 2% Scientist Worldwide 2022 (2023)
Among the 2% most impactful researchers of 2022, which is to be taken with a grain of salt. (rank(ns)=136,075)

Best Short Paper Award (2023) — The ACM Symposium for Eye Tracking Research and Applications

Stanford Top 2% Scientist Worldwide 2021 (2022)
Among the 2% most impactful researchers of 2021, which is to be taken with a grain of salt. (rank(ns)=195,784)

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

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

Freunde des HHI (2019) — Förderpreis für exzellente wissenschaftliche Arbeiten am HHI

ERCIM (2019, finalist) — Cor Baayen Young Researcher Award

Best Paper Award (2016) — ICML'16 Workshop on Visualization for Deep Learning

---

Patents

Analyzing an Inference of a Machine Learning Predictor

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

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

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

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)

US 20180018553 A1 "Relevance Score Assignment for Artificial Neural Networks " (published 2018-01-18)

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

Talks & Lectures

Talks

excludes internal/confidential events

1. "Artificial Intelligence We Can Trust – From Explainable to Actionable and Regenerative AI" (2024-02-02).
   MPNE Consensus 2024 Workshop, Berlin, Germany, (*invited talk*)
2. "From Concepts to Prototypes – Towards Minimal Effort Post-Hoc Interpretability" (2024-01-12).
   2nd Machine Teaching for XAI Workshop (*MT4XAI*), *Valencia, Spain,* (*invited talk*)
3. "Explaining AI with Concept Relevance Propagation" (2023-10-06).
   4th Japanese-American-German Frontiers of Science (*JAGFOS*) *Symposium, Dresden, Germany,* (*flash talk & poster, invited*)
4. "Model-Assisted Data Analysis via XAI" (2023-07-05).
   19th Machine Learning in Healthcare Meetup Berlin, Berlin Institute of Health, (*invited talk*)
5. "Accessing the Hidden Space with Explainable Artificial Intelligence" (2023-06-27).
   Informatik-Kolloquium, Universität Bremen, (*invited talk*)
6. "Explainable AI and Beyond with Concept Relevance Propagation" (2023-05-24).
   Data Professional Days / Data4Business Days Köln, (*keynote*)
7. "Beyond Heatmaps – Explaining with Concepts" (2022-10-21).
   BIFOLD Graduate School Welcome Days, (*invited talk*)
8. "Explain to Not Forget: Defending Against Catastrophic Forgetting with XAI " (2022-08-24).
   CD-MAKE 2022, (*paper presentation*)
9. "Zukünftige Trends in der KI und Einsatzmöglichkeiten im Bauwesen" (2022-06-24).
   BIMKIT Jahresveranstaltung 2022, (*keynote*)
10. "Beyond Explaining" (2021-06-03).
    Melanoma Patient Network Europe Meet-up – MPNE meets AI , (*invited talk*)
11. "Beyond Explaining: Explainable AI for Model Improvement" (2021-05-05).
    Sensor and Measurement Science International 2021, (*invited talk*)
12. "Efficient and Effective Neural Network Pruning with Layer-wise Relevance Propagation" (2020-11-12).
    Machine Learning Seminar at Fraunhofer HHI / Technische Universität Berlin
13. "Towards Best Practice in Explaining Neural Network Decisions with LRP" (2020-07-21).
    IEEE World Congress on Computational Intelligence 2020 / IJCNN 2020
14. "XAI for Analyzing and Unlearning Spurious Correlations in ImageNet" (2020-07-18).
    XXAI: Extending Explainable AI Beyond Deep Models and Classifiers, (*ICML 2020 Workshop*)
15. "XAI via LRP and SpRAy" (2020-07-02).
    Ada Day at Ada Lovelace Center / Fraunhofer IIS, (*invited talk*)
16. "Interpretable Machine Learning through Layer-wise Relevance Propagation" (2020-02-18).
    Fraunhofer Symposium Netzwert 2020
17. "Interpretable Machine Learning through Layer-wise Relevance Propagation" (2019-12-12).
    Gesellschaft von Freunden des HHI e.V.
18. "Explainable Artificial Intelligence — Opening the Machine Learning Black Box with Layer-wise Relevance Propagation" (2019-09-26).
    AMA Wissenschaftsrat 2019, (*invited talk*)
19. "Finding Clever Hans" (2019-07-16).
    Universität Bamberg, (*invited talk & press interview*)
20. "AI – Opening the Black Box" (2019-02-25).
    Robert Koch Institut, (*invited talk*)
21. "AI – Opening the Black Box" (2019-02-22).
    Technology Innovation Day – 91 Years HHI
22. "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
23. "Layer-wise Relevance Propagation" (2014-09-10).
    IDA Retreat'14

Invited Lectures

Individual Lectures as Parts of Seminars and Workshops

1. "Explainable AI" (2023-12-14/15/16).
   Universitat de Girona, (*3-day XAI lecture series as part of the Machine Learning Seminar at UdG*)

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. "Explainable AI" (2021-12-13/14/15).
Universitat de Girona, (*3-day XAI lecture series as part of the Machine Learning Seminar at UdG*)
7. "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*)
8. "Neuronale Netze mit LRP (richtig) erklären" (2020-08).
KI-Campus | Die Lernplattform für Künstliche Intelligenz
9. "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

---

Publications

Summary of Scientific Impact

	All	Since 2019
# Publications	68	52
# Citations	14208	13234
h-index	30	30
i10-index	43	41

per Google Scholar, retreived on April 16th, 2024.

Journal Articles

1. 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
2. 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
3. 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*
4. 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
5. 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
6. 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

7. 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
8. 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
9. 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*
10. 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
11. 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
12. 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
13. 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
14. 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*
15. 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
16. 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
17. 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
18. 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
19. **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
20. 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

21. 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*

22. 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*

23. 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

24. **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

25. **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. 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

2. 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*)

3. 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

4. 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)

5. 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

6. 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

7. 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

8. 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

9. 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

10. 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)

11. 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
12. 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)*
13. 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
14. 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
15. 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
16. 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*
17. 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*
18. 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*
19. 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*
20. **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
21. 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
22. **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
23. 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
24. **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
25. 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
26. 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*

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. 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: *CoRR abs/2404.09601*.
Accepted at CVPRW SAIAD 2024
2. Dreyer M, Purelku E, Vielhaben J, Samek W, **Lapuschkin S** (2024).
"PURE: Turning Polysemantic Neurons Into Pure Features by Identifying Relevant Circuits".
In: *CoRR abs/2404.06453*.
https://github.com/maxdreyer/PURE
*Accepted at CVPRW XAI4CV 2024 as a **spotlight paper***
3. Yolcu G Ü, Wiegand T, Samek W, **Lapuschkin S** (2024).
"DualView: Data Attribution from the Dual Perspective".
In: *CoRR abs/2402.12118*.
https://github.com/gumityolcu/DualView
4. 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: *CoRR abs/2402.05602*.
https://github.com/rachtibat/LRP-for-Transformers
5. Bley F, **Lapuschkin S**, Samek W, Montavon G (2024).
"Explaining Predictive Uncertainty by Exposing Second-Order Effects".
In: *CoRR abs/2401.17441*
6. Hedström A, Weber L, **Lapuschkin S**, Höhne M-C M (2024).
"Sanity Checks Revisited: An Exploration to Repair the Model Parameter Randomisation Test".
In: *CoRR abs/2401.06465*
7. Dreyer M, Achtibat R, Samek W and **Lapuschkin S** (2023).
"Understanding the (Extra-)Ordinary: Validating Deep Model Decisions with Prototypical Concept-based Explanations".
In: *CoRR abs/2311.16681*.
https://github.com/maxdreyer/pcx
Accepted at CVPRW SAIAD 2024
8. Nobis G, Aversa M, Springenberg M, Detzel M, Ermon S, Nakajima S, Murray-Smith R, **Lapuschkin S**, Knochenhauer C, Oala L and Samek W (2023).
"Generative Fractional Diffusion Models".
In: *CoRR abs/2310.17638*
9. Weber L, Berend J, Binder A, Wiegand T, Samek W and **Lapuschkin S** (2023).
"Layer-wise Feedback Propagation".
In: *CoRR abs/2308.12053*
10. 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*
11. Pahde F, Dreyer M, Weber L, Weckbercker 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*

12. 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*

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