Dr. Sebastian Lapuschkin (*né Bach*), *December 16, 1986*

sebastian@lapuschkin.com • +49 (177) 483-2754 • Google Scholar
github.com/sebastian-lapuschkin • linkedin.com/in/sebastian-lapuschkin
Kaiserin-Augusta-Allee 92 • 10589 Berlin • Berlin • Germany

Short Bio

Sebastian Lapuschkin received the 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.

Currently, he is the Head of the Explainable Artificial Intelligence at Fraunhofer Heinrich Hertz Institute (HHI) in Berlin. He 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, or towards the utilization of interpretable feedback for the improvement of machine learning systems and data. Further research interests include efficient machine learning and data analysis, data and algorithm visualization.

Professional Experience

Fraunhofer Heinrich Hertz Institute/HHI — Berlin, Germany

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

Group Leadership and direction of XAI research.

Research: Work towards the next generation of Explainable AI approaches and XAI-based model improvement by, e.g., increasing efficiency (see also) and debugging model training, reasoning and datasets. 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 staff.

Tenured Researcher — *Jan '19 – present*

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: Minor 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 in international events.

Additional supervision by Prof. Dr. Wojciech Samek.

Berlin Institute of Technology/TU Berlin — 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 department for machine learning at TU Berlin.

Teaching: Preparation and lecturing of exercise sessions complementing the lectures "Machine Learning 1" and "Machine Learning 2 – Theory and Application". Visualization and animation of data and learning algorithms discussed in the lecture.
Teaching assistant to Prof. Dr. Klaus-Robert Müller and Prof. Dr. Dr. Franz Király.

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/TU Berlin — 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, 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"
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"
Supervision headed by Prof. Dr. Kristian Hildebrand.

Deutschhaus-Gymnasium/DHG — Würzburg, Germany

Abitur (pre-university secondary education) *1998 – 2007*

Research Projects

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

TraMeExCo (2018-09 – 2021-08, leading role) Transparent Medical Expert Companion

Skills

Technical: Various Software Languages, packages and environments, e.g.,
Python (NumPy, PyTorch, ...), Matlab, Linux, bash, git, subversion, Slurm, Sun Grid Engine, ...

Scientific working and writing: LaTeX, Inkscape, WYSIWYG word processors, ...

Machine Learning: Development, application and evaluation of, e.g.,
SVMs, DNNs, processing pipelines, embeddings, clusterings, ...

Domain Knowledge: text, audio, video, images, time series, biomechanical and biomedical data, ...

Natural languages: German (*native*), English (*full professional proficiency*)

Awards

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 Prize (2016) ICML'16 Workshop on Visualization for Deep Learning

Patents

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)

Talks & Lectures

Talks

excludes internal/confidential events

1. "Zukünftige Trends in der KI und Einsatzmöglichkeiten im Bauwesen" (2022-06-24).
   BIMKIT Jahresveranstaltung 2022, (*keynote*)
2. "Beyond Explaining" (2021-06-03).
   Melanoma Patient Network Europe Meet-up – MPNE meets AI , (*invited talk*)
3. "Beyond Explaining: Explainable AI for Model Improvement" (2021-05-05).
   Sensor and Measurement Science International 2021, (*invited talk*)
4. "Efficient and Effective Neural Network Pruning with Layer-wise Relevance Propagation" (2020-11-12).
   Machine Learning Seminar at Fraunhofer HHI / Technische Universität Berlin
5. "Towards Best Practice in Explaining Neural Network Decisions with LRP" (2020-07-21).
   IEEE World Congress on Computational Intelligence 2020 / IJCNN 2020
6. "XAI for Analyzing and Unlearning Spurious Correlations in ImageNet" (2020-07-18).
   XXAI: Extending Explainable AI Beyond Deep Models and Classifiers, (*ICML 2020 Workshop*)
7. "XAI via LRP and SpRAy" (2020-07-02).
   Ada Day at Ada Lovelace Center / Fraunhofer IIS, (*invited talk*)
8. "Interpretable Machine Learning through Layer-wise Relevance Propagation" (2020-02-18).
   Fraunhofer Symposium Netzwert 2020
9. "Interpretable Machine Learning through Layer-wise Relevance Propagation" (2019-12-12).
   Gesellschaft von Freunden des HHI e.V.
10. "Explainable Artificial Intelligence — Opening the Machine Learning Black Box with Layer-wise Relevance Propagation" (2019-09-26).
    AMA Wissenschaftsrat 2019, (*invited talk*)
11. "Finding Clever Hans" (2019-07-16).
    Universität Bamberg, (*invited talk & press interview*)
12. "AI – Opening the Black Box" (2019-02-25).
    Robert Koch Institut, (*invited talk*)
13. "AI – Opening the Black Box" (2019-02-22).
    Technology Innovation Day – 91 Years HHI
14. "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

Lectures

1. "Explainable AI" (2021-12-13/14/16).
   Universitat de Girona, (*3-day lecture series as part of the Machine Learning Seminar at UdG*)
2. "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*)

3. "Neuronale Netze mit LRP (richtig) erklären" (2020-08).
   KI-Campus | Die Lernplattform für Künstliche Intelligenz
4. "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

Journal Articles

1. 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
2. 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
3. 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
4. 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
5. 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
6. 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
7. 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
8. 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
9. 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
10. 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
11. 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*
12. **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

13. 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*
14. 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*
15. 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*
16. 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
17. **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*
18. **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. 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*
2. 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*
3. 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* 1-7
4. 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*
5. 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*
6. 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*
7. 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*
8. **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
9. 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
10. **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
11. 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*

12. **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
13. 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
14. 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$^{\text{X}}$: 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. Achtibat R, Dreyer M, Eisenbraun I, Bosse S, Wiegand T, Samek W and **Lapuschkin S** (2022).
"From "Where" to "What": Towards Human-Understandable Explanations through Concept Relevance Propagation".
In: *CoRR abs/2206.03208*. *https://github.com/rachtibat/zennit-crp*
2. 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: *CoRR abs/2205.01929*. *Accepted for publication and presentation at CD-MAKE 2022*
3. 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*
4. Weber L, **Lapuschkin S**, Binder A and Samek W (2022).
"Beyond Explaining: Opportunities and Challenges of XAI-Based Model Improvement".
In: *CoRR abs/2203.08008*
5. Hedström A, Weber L, Bareeva D, Motzkus F, Samek W, **Lapuschkin S** and Höhne M-C M (2022).
"Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations".
In: *CoRR abs/2202.06861*. *https://github.com/understandable-machine-intelligence-lab/quantus*
Accepted for publication in JMLR (June 2022).
6. Motzkus F, Weber L and **Lapuschkin S** (2022).
"Measurably Stronger Explanation Reliability via Model Canonization".
In: *CoRR abs/2202.06621*. *Accepted for publication and presentation at ICIP 2022*
7. Pahde F, Weber L, Anders CJ, Samek W and **Lapuschkin S** (2022).
"PatClArC: Using Pattern Concept Activation Vectors for Noise-Robust Model Debugging".
In: *CoRR abs/2202.03482*
8. Hofmann S M, Beyer F, **Lapuschkin S**, Loeffler M, Müller K-R, Villringer A, Samek W and Witte A V (2021).
"Towards the Interpretability of Deep Learning Models for Human Neuroimaging".
In: *bioRxiv 2021.06.25.449906*. *Accepted for publication in NeuroImage (July 2022)*
9. 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*

10. Becker S, Ackermann M, **Lapuschkin S**, Müller K-R and Samek W (2018).
"Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals".
In: *CoRR abs/1807.03418*. *https://github.com/soerenab/AudioMNIST*

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