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 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 a tenured researcher and head of the Explainable AI Group at Fraunhofer Heinrich Hertz Institute (HHI) in Berlin. He is recipient of multiple awards, including the Hugo-Geiger-Prize for outstanding doctoral achievement and the 2020 Pattern Recognition Best Paper Award. His current research is focused on actionable eXplainable AI for the interpretation, holistic analysis and rectification of machine learning system behavior. Further research interests include efficient machine learning and data analysis, data and algorithm visualization.

Sebastian loves automation and woodland hiking (not simultaneously).

Professional Experience

Fraunhofer Heinrich Hertz Institute/HHI — Berlin, Germany

Head of XAI Group — *Jan '21 – present*

Research focus: Developement of holistic, automatable and human-centered XAI approaches for understanding, improving and debugging models, model training and datasets.

Tenured Researcher — *Jan '19 – present*

Research focus: Development of (meta-)analysis methods of machine learning behaviour. Improving machine learning predictors and data sources using interpretability feedback. Supervision of PhD students and student assistants.

Research Associate — *Oct '14 – Dec '18*

Affiliation to the newly founded machine learning group at Fraunhofer HHI with simultaneous continuation of PhD studies at TU Berlin.

Research focus: Applications and refinement of the "Layer-wise Relevance Propagation" (LRP) method, resulting in several highly cited publications and multiple open source software tools and repositories.

Other work: Extensions of the h.265 (HEVC) video codec towards the upcoming h.266 standard.
Conceptualization and setup of a HPC cluster with modern GPU hardware.
Implementation of multiple live demos hands-on showcasing the groups' research nation-wide and internationally.

Additional supervision by Dr. Wojciech Samek.

Berlin Institute of Technology/TU Berlin — Berlin, Germany

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

Research focus: Formalization of the "Layer-wise Relevance Propagation" (LRP) concept for explaining individual and nonlinear decisions of machine learning methods, including Neural Networks and kernelized predictors.

Extension of LRP to one class learning and anomaly detection tasks.

Supervision by Prof. Dr. Klaus-Robert Müller and Prof. Dr. Alexander Binder.

Research/Teaching Assistant — *Oct '11 – Aug '13*

Research assistant to Prof. Dr. Alexander Binder at the department for machine learning at TU Berlin.
Tasks: Structure and cell type detection in large histopathology images using Bag of Words image processing pipelines and SVM classifiers.

Teaching assistant to Prof. Dr. Klaus-Robert Müller and Prof. Dr. Franz Király.
Tasks: 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 *Oct '09 – Sep '11*

Teaching assistant to Prof. Dr. Marc Alexa, Prof. Dr. Odej Kao and Prof. Dr. Oliver Brock.
Tasks: Course instruction for algorithmic and practical foundations of computer science (B.Sc.).

Curriculum: Basic and advanced Java development, software engineering and OOP concepts, algorithms on image and graph data, among others.

Education

Berlin Institute of Technology/TU Berlin — Berlin, Germany

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

Research focus on methods and applications of *eXplainable AI* (*XAI*):
Layer-wise Relevance Propagation, Deep Taylor Decomposition, Spectral Relevance Analysis, et cetera.

20 published and peer-reviewed research papers with over 2000 citations according to Google Scholar.

Thesis: "Opening the machine learning black box with Layer-wise Relevance Propagation"

Supervision headed by Prof. Dr. Klaus-Robert Müller.

Master of Science degree in Computer Science *2010 – 2013*

Heavy focus on machine learning, computer vision and large scale data processing. Development dominantly using C++, C#, Java, Matlab and python.

My thesis introduces a precursor concept to "Layer-wise Relevance Propagation" for interpreting predictions of Bag of Words image processing pipelines with-multiple-kernel SVMs.

Thesis: "On Pixel-wise Predictions from Image-wise Bag of Words Classification" (Grade: 1.0 / A)

Supervision headed by Prof. Dr. Alexander Binder.

Bachelor of Science degree in Computer Science *2007 – 2010*

Focus on algorithms, software development and data analysis using imperative (Java, C, C++) and functional (OPAL) programming languages.

Thesis: "Keyword-Based Image Browsing of Large Image Databases" (Grade: 1.0 / A)

Supervision headed by Prof. Dr. Kristian Hildebrand.

Deutschhaus-Gymnasium/DHG — Würzburg, Germany

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

Skills

Technical: Extensive experience in software development using various languages, tools and environments:
(e.g. python, Linux, bash, git, subversion, Slurm, Sun Grid Engine,
Matlab, HTML, C#, C++, C, Java, lua, SQL, ...)

Proficiency in scientific working and writing (e.g. LaTeX, Inkscape, ...),

the development and application of machine learning methods
(e.g. SVMs, DNNs, preprocessing pipelines, embeddings, clustering, ...),

in various application domains
(text, audio, video, images, time series and biomechanical data, ...).

Natural languages: German (*mother tongue*), 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

Relevance Score Assignment for Artificial Neural Networks

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)

Projects

iToBoS (2021-04 -) Intelligent Total Body Scanner

BerDiBa (2021-01 -) Berlin Digital Rail Operations

TraMeExCo (2018-09 -) Transparent Medical Expert Companion

Publications

Journal Articles

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

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

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

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

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

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

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

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

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*

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

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

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*

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*

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

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*

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

Sun J, **Lapuschkin S**, Samek W, Zhao Y, Cheung N-M and Binder A (2022).
"Explanation-Guided Training for Cross-Domain Few-Shot Classification".
In: *Proceedings of the 25th International Conference on Pattern Recognition*

Goh G S W, **Lapuschkin S**, Weber L, Samek W and Binder A (2022).
"Understanding Integrated Gradients with SmoothTaylor for Deep Neural Network Attribution".
In: *Proceedings of the 25th International Conference on Pattern Recognition*

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

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*

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*

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*

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*

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

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

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

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*

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

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

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

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

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

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

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*

Motzkus F, Weber L and **Lapuschkin S** (2022).
"Measurably Stronger Explanation Reliability via Model Canonization".
In: *CoRR abs/2202.06621*

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*

Becking D, Dreyer M, Samek W, Müller K and **Lapuschkin S** (2021).
"ECQ$^{\text{X}}$: Explainability-Driven Quantization for Low-Bit and Sparse DNNs".
In: *CoRR abs/2109.04236*

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*

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*

Rieckmann A, Dworzynski P, Arras L, **Lapuschkin S**, Samek W, Onyebuchi A A, Rod N H, Ekstrom C T (2020).
"Causes of Outcome Learning: A Causal Inference-inspired Machine Learning Approach to Disentangling Common Combinations of Potential Causes of a Health Outcome".
In: *medRxiv 2020.12.10.20225243*

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*

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