Nitin Sharma

Research Assistant — AI Interpretability
Email: nitinsharma3150@gmail.com
Contact No.: +49-17667635491
Website — LinkedIn — Google Scholar

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

EDUCATION

Year	Degree/Examination	Institution/Board	CGPA/ Percentage
2024	Master of Science	Eberhard Karls University of Tübingen	1.17/4.0
2022	Bachelor of Technology	Indian Institute of Technology, Roorkee	9.57/10
2018	Intermediate (Class XII)	Arcadia Academy (CBSE), Kota	92%

WORK EXPERIENCE

Steering Vectors for Knowledge Access in LLMs | RA, Dr. Wolfers and Dr. Yıldız April 2025 - Present

- Developing activation engineering techniques using tuned lens, causal tracing, and activation patching to localize domain-specific knowledge representations across model layers.
- Analyzing attribute extraction rates and layer-wise knowledge evolution through hook-based interventions to identify targetable directions for systematic model control.

Mechanistic Understanding of Factual Knowledge in LLMs | Master's Thesis, Bethge Lab April 2024 - March 2025

- Developed deterministic pipeline for domain-specific benchmarks from raw corpora; extended work submitted to EACL 2025 (first review round passed, pre-print).
- Conducted large-scale experiments across multiple architectures (1.56M arXiv documents, 8.5B tokens), revealing rapid domain adaptation and layer-wise knowledge representation patterns.

Medical Domain Benchmark Extension | PhD Student Supervision, Mental Health Mapping Lab August 2025 - Present

- Supervising extension of benchmarking framework to medical and mental health domains, focusing on safety-critical evaluation and data contamination effects.

Normative Modeling and GAMLSS Python Package | HiWi/RA, Mental Health Mapping Lab March 2024 - Present

- Developing GAMLSS Python package for neuroimaging with parallel processing and permutation testing.
- Applying toolbox to 25,000-individual lifespan dataset; abstract submitted to OHBM 2025 - paper expected in March.

Nerve Disease Diagnostics using ML | Co-supervisor, Bethge Lab October 2024 - January 2025

- Co-supervising a master's student's lab project focusing on ML applications in nerve disease diagnostics.
- Providing guidance on methodology, implementation, and analysis of ultrasound-based diagnostic tools.

B-cos Learning for rs-fMRI Data Interpretation | HiWi, Mental Health Mapping Lab August 2023 - December 2023

- Reviewed literature on explainable AI methods, focusing on B-cos learning and rs-fMRI analysis.
- Evaluated explainable AI techniques for application to large-scale rs-fMRI datasets in neuroimaging research.

Meta-cognitive Ability in Reversal Tasks | Lab Rotation, Comp. Neuro. Lab November 2023 - February 2024

- Studied decision-making in two-armed bandit tasks with reversal conditions, comparing human and model performance.
- Developed Q-learning and HSMM models to capture nuances of human decision-making and metacognition.

Mechanistic Interpretability of LLMs in Mental Healthcare: A Review | Essay Rotation, Mental Health Mapping Lab September 2023 - November 2023

- Analyzed LLM applications in mental health, exploring their potential for psychotherapy and personalized treatment.
- Focused on mechanistic interpretability to address LLM accountability in privacy, bias, and ethics.

Postoperative Delirium Risk Assessment | HiWi, Mental Health Mapping Lab April 2023 - August 2023

- Developed ML models to predict postoperative delirium in 1,624 elderly patients from five medical centers.
- Applied SHAP values for model interpretation and permutation testing; co-first authored resulting pre-print.

MDD Biomarker Detection | DAAD WISE Scholarship, Friedrich Schiller University June 2021 - August 2021

- Detected MRI-based biomarkers for Major Depressive Disorder using PsyMRI data and connectivity features.
- Applied various ML and DL techniques including ANN, LSTM, and Autoencoder for feature analysis.

RESEARCH PUBLICATIONS AND PRE-PRINTS

- Sharma, N., Wolfers, T., & Yıldız, Ç. (2025). Beyond Benchmarks: A Novel Framework for Domain-Specific LLM Evaluation and Knowledge Mapping. arXiv preprint arXiv:2506.07658.
- Yıldız, Ç., Ravichandran, N. K., Sharma, N., Bethge, M., & Ermis, B. (2024). Investigating continual pretraining in large language models: Insights and implications. arXiv preprint arXiv:2402.17400. (Accepted in TMLR)
- Wu, S. C. J.*, Sharma, N.*, Bauch, A., Yang, H. C., Hect, J. L., Thomas, C., ... & PAWEL Study Group. (2025). Predicting Postoperative Delirium in Older Patients Before Elective Surgery: Multicenter Retrospective Cohort Study. JMIR aging, 8(1), e67958.
- Kim, M., Sharma, N., Leonardsen, E. H., Rutherford, S., Selbæk, G., Persson, K., ... & Moberget, T. (2025). Predicting Mental and Neurological Illnesses Based on Cerebellar Normative Features. Biological Psychiatry: Global Open Science, 5(5).
- Sen, Z. D., Sharma, N., Danyeli, L. V., Colic, L., Opel, N., Chand, T., ... & Li, M. (2024). Ketamine-induced pleasant but not unpleasant dissociation is linked to the functional connectivity profile of the posteromedial cortex
- Li, M., Sharma, N., Danyeli, L., Colic, L., Opel, N., Chand, T., ... & Walter, M. (2023). 56. Ketamine-induced ego dissolution is related to the functional connectivity reconfiguration of the posteromedial cortex. Biological Psychiatry, 93(9), S93.
- Sharma, N., Gaurav, G., & Anand, R. S. (2021, August). Epileptic seizure detection using STFT based peak mean feature and support vector machine. In 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) (pp. 1131-1136). IEEE.

PROJECTS

Understanding the effect of Ketamine on brain | Divyadrishti Lab, IIT Roorkee & Jena University March 2022 - July 2022

- Studied Ketamine's effect on brain connectivity and its potential as a biomarker for Major Depressive Disorder.
- Applied ML for feature refinement and analyzed cognitive questionnaire data; resulted in a pre-print publication.

Deep learning for inter-site heterogeneity in multi-site MRI data | Divyadrishti Lab, IIT Roorkee & Jena University August 2021 - January 2022

- Addressed heterogeneity in multi-site MRI data using fMRI and demographic information from PsyMRI dataset.
- Used unsupervised domain adaptation and XAI to understand heterogeneity sources and improve MDD classification.

Machine learning for Stroke detection | Prof. Sumit Kumar Yadav, IIT Roorkee March 2021 - June 2021

- Conducted statistical analysis and ML-based classification on a Kaggle stroke dataset.
- Improved statistical parameters using imbalance-adjusted ML methods for stroke detection.

GUI for EEG signal processing | Biomedical Instrumentation Lab, IIT Roorkee February 2021 - June 2021

- Developed a Python-based GUI for EEG analysis, catering to both non-programming and programming users.
- Implemented various signal processing and ML algorithms using libraries like MNE, SciPy, and Scikit-learn.

Physiological stress detection | Biomedical Instrumentation Lab, IIT Roorkee December 2019 - July 2020

- Collected and analyzed EEG, ECG, and Pulse oximeter data during stress and relaxation tasks.
- Applied signal processing techniques and feature extraction methods using Python, Matlab, and various libraries.

Epileptic seizure detection using EEG | Biomedical Instrumentation Lab, IIT Roorkee December 2019 - July 2020

- Performed EEG signal analysis to detect seizure onset and classify the EEG epilepsy Bonn dataset.
- Published findings in IEEE conference paper, presented at SPIN 2021 conference in Noida, India.

AWARDS / SCHOLARSHIPS / ACADEMIC ACHIEVEMENTS

- Deutschlandstipendium scholarship (2024): For outstanding academic achievements at University of Tübingen.
- Best Presentation Award (2023): For essay rotation in Neural Information Processing branch, Graduate Training Centre of Neuroscience, Tübingen.

- Department Gold Medal - Physics Department (2023), Indian Institute of Technology Roorkee: Awarded for maintaining the highest academic performance throughout the four-year Bachelor's program.
- Best Bachelor Thesis Award - Physics Department (2023), Indian Institute of Technology Roorkee: Secured the top thesis recognition for an outstanding thesis.
- The DAAD WISE (Working Internships in Science and Engineering) (2021): For summer internship in Germany.
- National Service Scheme, Indian Institute of Technology Roorkee 'Dedicated Member' Award (2019): Recognized for outstanding leadership and active participation in multiple community service initiatives.
- Kishore Vaigyanik Protsahan Yojana (KVPY) fellowship (2018): Prestigious national fellowship for exceptional students in basic sciences, funded by the Department of Science and Technology, India.

SKILLS

Computer languages	Python, C++, MATLAB, Assembly language
Software Packages	PyTorch, TensorFlow/Keras, Transformers, NLTK, spaCy, Pandas, NumPy, SciPy, Scikit-learn, Git, Neuroimaging: SPM12, FSL, MNE-Python, NiLearn
Additional Courses	Deep Learning, NLP, Machine Learning, Feature Selection for ML, Custom Models and Loss Functions in TensorFlow, Principles of fMRI, Fundamental Neuroscience for Neuroimaging
Languages Known	English (Proficient), Hindi (Native)

POSITIONS OF RESPONSIBILITY & EXTRA CURRICULARS

Teaching Assistant | Neuromatch Academy, Deep Learning Course — July 8 - 26, 2024

- Guided international students through complex Deep Learning concepts in an intensive three-week course.
- Facilitated daily tutorials and project work, collaborating with a global team of TAs and instructors.

Teaching Assistant | Academic Reinforcement Program, IIT Roorkee — January 2022 - March 2022

- Assisted freshers with BT-103 (Computer Systems and Programming) coursework.
- Provided programming and theoretical support, occasionally leading summary classes.

Mentor | Student Mentorship Program, IIT Roorkee — December 2021 - May 2022

- Guided first-year students in academic, personal, and professional development.
- Conducted regular meetings to address challenges faced by freshmen.

Executive | National Service Scheme (NSS), IIT Roorkee — July 2018 - June 2020

- Organized various social initiatives including Blood Donation Camps and Ganga Cleanliness Drive.
- Led 'Daan Petika' project to collect and distribute clothes to those in need.

Conference Presenter | SPIN 2021 — August 2021

- Presented paper on "Epileptic seizure detection using STFT based peak mean feature and support vector machine".
- Research based on EEG analysis project completed under Prof. R.S. Anand, IIT Roorkee.

Coordinator | Cognizance, IIT Roorkee — March 2019

- Coordinated Machine Learning and Artificial Learning Workshop at IIT Roorkee's technical fest.
- Managed over 250 students and guests during the event.

REFERENCES

Dr. Çağatay Yıldız
Vernade Lab, Tübingen AI Center
University of Tübingen
cagatay.yildiz@uni-tuebingen.de

Dr. Thomas Wolfers
Department of Psychiatry and Psychotherapy
Universitätsklinikum Tübingen
Thomas.Wolfers@med.uni-tuebingen.de