Rajashik Datta

+91 62941 32431 rajashikdatta215@gmail.com Google Scholar LinkedIn Github Portfolio

Research Interests

Trustworthy ML (xAI + robustness), Computer Vision (Hyperspectral/Multimodal), Human-Computer Interaction

Education

Institute of Engineering & Management, Kolkata, India **August 2022 – July 2026**

B.Tech in Computer Science & Engineering (Artificial Intelligence) *CGPA: 9.19 / 10*

Ranked 6th among the top 10% of class by CGPA in Year 3 (AY 2024–25).

Awarded the **Chancellor's Award for Exemplary Research Contribution 2026**, the sole recipient in the department.

Experience

University of Nebraska-Lincoln, USA **June 2025 – Present**

Research Intern *Remote (USA)*

Supervisor: *Dr. Sruti Das Choudhury* (*Offer Letter*)

- Spearheaded an explainable AI + data-storytelling clustering pipeline across precision agriculture and pediatric healthcare—grouping 22 Indian crop types using 7 agro-climatic/soil features and segmenting a 500-record hospital cohort—showing that z-score rescaling + removing binary gender prevents charge-dominated clusters and surfaces clinically meaningful cohorts (LOS up to 29 days; charges up to 34,644) for decision support [P3].
- Co-authored *ReproPheno/ReproPhenoNet*, presenting a large-scale UNL benchmark for reproductive-stage plant phenotyping (∼6 TB) comprising multiview, temporal visible-light, fluorescence, and hyperspectral image sequences of flowers and fruits. Also contributed to a YOLO-based ReproPhenoNet pipeline for flower/fruit detection and temporal phenotype quantification, attaining combined mAP@0.5=0.849 and F1=0.838 [P6].
- Co-developed *PlantPhenoLM*, a retrieval-augmented LLM reasoning framework for phenotype-to-genotype inference that integrates classifier probabilities, nearest-neighbor phenotype evidence, entropy-driven ambiguity diagnostics, and selective prediction to generate auditable genotype reports; achieved top-5 recovery of 0.952 and increased retrieval-augmented top-1 accuracy to 0.429 in 5-fold held-out evaluation [P7].
- Developed a temporal-embedding visual analytics system for 42 plants from 9 genotypes over 25 days, engineering multi-scale phenotype descriptors (growth rates/accelerations, fourier spectra, wavelet energies, distributional stats) and achieving genotype-aligned DTW clustering (ARI 0.30; NMI 0.62) with cross-validated early-prediction curves and SHAP/LIME-linked causal graphs to explain when/why genotypes diverge [P14].
- Implemented an interactive hyperspectral analysis tool, HyperProbe for calibrated datacubes spanning 517-1700 nm (B=243 bands), enabling rapid pixel/ROI annotation, band-difference + Otsu segmentation (IoU/F1 evaluation), and full-scene classification via 3 model families (MLP/logistic regression/random forest) with built-in ablations that log clicks/ROIs under fixed 5-min label budgets to quantify accuracy-per-effort [P15].
- Featured in the university's news story for research contributions: snr.unl.edu (August, 2025)

University of Calcutta **January 2025 – Present**

Research Scholar *Kolkata, India*

Supervisors: *Dr. Arup Kumar Chattopadhyay*, *Prof. Amit Kumar Das*, *Prof. Amlan Chakrabarti*

- Engineered FHFAM (FH-FAM), a fuzzy-hypergraph feature selection algorithm, achieving the best mean accuracy (81.43%) and best mean feature reduction (89.28%) across 15 agriculture/remote-sensing datasets (5/15 wins) with 11.08s average runtime and statistically significant accuracy gains over key baselines (Wilcoxon $p < 0.05$) [P1].
- Proposed SIFHFAM, a stage-wise intuitionistic-fuzzy hypergraph selector with a monotone submodular coverage objective and greedy (1−1/e) guarantee, delivering the top average accuracy (≈84%) while pruning ≈99% features (typically retaining $< 2\%$) across 14 high-dimensional benchmarks in ∼0.1s/run under 10× repeated 75/25 train-test splits [P13].

Generative AI Centre of Excellence, IEM **November 2024 – Present**

Student Research Lead; Senior Research Advisor at GenAI CoE *Kolkata, India*

Established and led GenAI CoE's research sub-committee, end-to-end research execution and operations—recruited and onboarded members via interviews, mentored and staffed project teams, coordinated 10+ journal groups, maintained the CoE website, and launched *ReelBook* (Pearson collaboration) and Medium publishing to scale institute-wide research output and AI upskilling at IEM.

IEM Research Foundation **August 2024 – March 2025**

Project Intern at bair.ai (Certificate) *Kolkata, India*

Built *MemeMetric*, an end-to-end cluster-based cryptocurrency forecasting system by architecting the full data/ML pipeline with automated reporting, and integrated real-time Twitter/Telegram/Reddit sentiment signals via NLP to improve robustness and reduce forecast error/volatility.

Innovation & Entrepreneurship Development Cell (CSE), IEM **March 2024 – August 2024**

Research Assistant (Certificate) *Kolkata, India*

Co-authored an IEM-HEALS 2024 accepted study analyzing Jul 2019–Dec 2022 price dynamics of 20 pharma stocks using multivariate regression, volatility modeling, and event-study methods, and engineered *TraderBot*, a Flask+MongoDB real-time trading simulator wired to Yahoo Finance for live strategy backtesting and portfolio experiments [P12].

National University of Singapore (NUS), Singapore **July 2023 (1 week)**

Study Abroad Program (Certificate) *Singapore*

Studied fundamentals of "Artificial Intelligence, Internet of Things, Machine Learning & Data Analytics", lectured by *Dr. Peter Leong*, *Dr. Eric Cambria*, *Dr. Matthew Chua*, *Dr. Yiliang Zhao*, *Dr. Gábor Benedek*, *Dr. Tan Kian Hua*, *Yong Heng Michael Tan*, *Marton Szel*, *Gillian Cheng*.

Selected Independent Research

- Co-created *MiQ-MCP*, a robust uncertainty estimation framework for high-frequency financial forecasting that integrates Fourier-augmented quantile regression with Mondrian conformal calibration across intraday time segments, enabling reliable prediction intervals for minute-level NIFTY-50 forecasts despite pronounced time-of-day volatility variations [P2].

- Co-developed *PolyJudge-Uncertain*, a multilingual 5,120-example benchmark in English, Hindi, Hinglish, and Bengali for analyzing hedging behavior in LLM-as-a-judge systems; observed that pointwise hedging penalties mostly vanish after template correction, while pairwise evaluators consistently favor more assertive responses when semantics are equivalent [P4].

- Explored garden-path sentence recovery in causal and masked language models using 100 English garden-path/control sentence pairs covering NP/Z, NP/S, and MV/RR structures, demonstrating that decoder-only architectures show stronger online syntactic reanalysis signals, whereas masked encoders remain rather stable due to bidirectional contextual information [P5].

- Introduced *GreedySAT Revision*, a neuro-symbolic approach modeling multi-turn LLM conversations as belief revision, employing an external SAT/SMT-inspired consistency verifier to preserve globally coherent belief states and prevent contradictory dialogue outcomes under adversarial stress conditions [P8].

Publications

Published/Accepted

Journals

P1. **Rajashik Datta**, Sanjan Baitalik, Sruti Das Choudhury, Arup Kumar Chattopadhyay, Amit Kumar Das, "Fuzzy Hypergraph Feature Association Map for High-Dimensional Feature Selection in Agriculture and Remote Sensing", International Journal of Fuzzy Systems, 2026.

P2. Sanjan Baitalik, **Rajashik Datta**, Darothi Sarkar, Ayan Chaudhuri, "MiQ-MCP: Valid and Conditionally Robust Uncertainty Quantification for High-Frequency Financial Time Series via Mondrian Conformalized Quantile Regression", Computational Economics, 2025.

P3. Sruti Das Choudhury, **Rajashik Datta**, Sanjan Baitalik, "Enhancing interpretability through clustering, explainable AI, and narrative visualization: applications in precision agriculture and healthcare patient segmentation", Information, 2025.

Conferences

P4. **Rajashik Datta**, Sanjan Baitalik, "*Confidence as a Tie-Breaker: Reassessing Multilingual Hedging Bias in LLM-as-a-Judge Evaluation*", The Association for Computational Linguistics (ACL Student Research Workshop), 2026.

P5. Sanjan Baitalik, **Rajashik Datta**, "*Garden Path Recovery in Causal and Masked Language Models*", The Association for Computational Linguistics (ACL Student Research Workshop), 2026.

P6. Sanjan Baitalik, **Rajashik Datta**, Utsho Banerjee, Rajarshi Karmakar, Vincent Stoerger, Himadri Nath Saha, Sruti Das Choudhury, "*ReproPheno and ReproPhenoNet: A Large-Scale Multimodal Benchmark Dataset and Deep Learning Framework for Reproductive-Stage Plant Phenotyping*", The Association for the Advancement of Artificial Intelligence (AAAI Workshop on AgriAI), 2026.

P7. **Rajashik Datta**, Sanjan Baitalik, Amit Kumar Das, Sruti Das Choudhury, *"PlantPhenoLM: Phenotype-Genotype Mapping Inference with Multi-Turn LLM Reasoning and Selective Prediction"*, The Association for the Advancement of Artificial Intelligence (AAAI Bridge on Logic & AI), 2026.

P8. Sanjan Baitalik, **Rajashik Datta**, Amit Kumar Das, Sruti Das Choudhury, *"Conversation as Belief Revision: GreedySAT Revision for Global Logical Consistency in Multi-Turn LLM Dialogues"*, The Association for the Advancement of Artificial Intelligence (AAAI Bridge on Logic & AI), 2026.

P9. Sanket Ghosh, Sanjan Baitalik, **Rajashik Datta**, Romit Mukherjee, Darothi Sarkar, Ayan Chaudhuri, *"Explanation-First Agentic Forecaster for Stock Market"*, International Conference on Electronics, Materials Engineering and Nano-Technology (IEMENTech 2026), 2026.

P10. Sanjan Baitalik, **Rajashik Datta**, Sanket Ghosh, Darothi Sarkar, Ayan Chaudhuri, *"Machine Learning-Driven Insights For Stock Market Analysis And Trading"*, International Conference on Interdisciplinary Research in Technology and Management (IRTM 2024).

P11. Sanket Ghosh, Sanjan Baitalik, **Rajashik Datta**, Darothi Sarkar, *"The Pandemic Shock: An Analysis of Impacts and Responses of Indian Stock Market"*, International Conference on Interdisciplinary Research in Technology and Management (IRTM 2024).

P12. **Rajashik Datta**, Sanjan Baitalik, Sanket Ghosh, Saugata Ghosh, Swarnendu Ghosh, "Is Indian Financial Market Ready for Pandemics?", International Conference on Advancing Science and Technologies in Health Science (IEM-HEALS 2024) *Book of Abstracts*.

Submitted

P13. Sanjan Baitaik, **Rajashik Datta**, Arup Kumar Chattopadhyay, Amit Kumar Das, Amlan Chakraborty, "From Graphs to Hypergraphs: Submodular Coverage-Based Feature Selection on Intuitionistic Fuzzy Hypergraphs (SIFHFAM)", Pattern Recognition, 2026.

Manuscripts in Preparation

P14. **Rajashik Datta**, Sanjan Baitalik, Sruti Das Choudhury, Amit Kumar Das, "Visual Analytics of Plant Phenotype-Genotype Dynamics via Temporal Embeddings". Intended for submission to IEEE Transactions on Visualization and Computer Graphics, June 2026.

P15. Sanjan Baitalik, **Rajashik Datta**, Sruti Das Choudhury, "HyperProbe Insight: An Interactive Tool for Exploration of Hyperspectral Image Sequences". Intended for submission to IEEE Transactions on Visualization and Computer Graphics, July 2026.

Skills & Activities

Programming: Python, C, C++, Java, MATLAB, LaTeX
ML/AI: PyTorch, TensorFlow, Scikit-learn, Transformers
XAI: SHAP, LIME
Data: Pandas, NumPy, SciPy
Databases: MySQL, PostgreSQL, MongoDB
Cloud: Google Cloud (Cloud Run/Compute), AWS (S3/EC2)
Visualization: Matplotlib, Seaborn, Plotly, Tableau
Tools: TensorBoard, MATLAB App Designer

Activities: GenSpark 1.0 Ideathon (Organizer; coordinated 50+ teams; shortlisted funded ideas), Jun–Aug 2025; IEM-ICDC 2025 (Conference volunteer; coordination & support), Apr 2025; Department of CSE, IEM (Assisted NBA accreditation documentation), Mar 2024

Projects

Quantization-Aware Momentum | *GitHub*

1-bit Momentum optimizer with error feedback; matches full-precision Momentum SGD. Logistic regression (n=4000, d=2000, 5000 steps): train loss 2.809×10^{-3} vs signSGD 3.7614×10^{-2} (13.39× higher); remains 6–13× better across weight decay sweeps.

Online Learning (VS-AdaGrad) | *GitHub*

Online learning for non-stationary time series via drift-aware, volatility-scaled AdaGrad; on piecewise AR(5) with 5 regimes (T≈5000, 10 seeds), improves regret proxy over AdaGrad by 18.4% (small drift) and 19.8% (medium), and beats tuned OGD by 23.7–63.8% across drift regimes.