Rahul Ray

rahulraypm2002@gmail.com | +31 6 20023962 | Den Haag, Netherlands | Portfolio | Rahulrayy | rahul-ray-b67741222

MSc Data Science student at Leiden with a habit of building things to answer questions I couldn't find clean answers to. Interested in what actually happens when you push a method past its assumptions.

Education

MSc Computer Science – Data Science *Leiden University, Netherlands* Sep 2025 – Aug 2027 (expected)

BTech Computer Science *SRM Institute of Science and Technology, Chennai* 2021 – 2025

Work Experience

Machine Learning Intern *Mar – Aug 2025 · Fox Solutions Pvt. Ltd., India*

- Developed a ViT-based computer vision pipeline end-to-end — from model training through to edge deployment with ONNX, optimising for inference latency and memory footprint on constrained hardware.

Projects

Can LLMs Make Tabular ML Better? writeup GitHub

Python · XGBoost · LLaMA 3.1 · sentence-transformers · Optuna · SHAP

- Tested LLM-generated semantic features as input to XGBoost via nested CV (3×5-fold, leak-free PCA, Optuna); the hypothesis failed on both datasets — documented honestly with Wilcoxon tests, Cohen's d, and SHAP analysis.

AutoInsight – End-to-End AutoML Web App writeup GitHub

Python · Streamlit · XGBoost · LightGBM · CatBoost · SHAP · Optuna · Groq API

- Trains 7 models in parallel with stratified CV, SHAP explainability, and Bayesian tuning; privacy-preserving LLM integration (aggregated metrics only, never raw data); live on Hugging Face Spaces.

Quake – Learned Heuristic A* on BSP Geometry writeup GitHub

Python · PyTorch · XGBoost · NetworkX · BSP v29 parsing

- Parsed Quake 1 BSP maps from scratch, generated 625k Dijkstra ground-truth pairs, and trained MLP/XGBoost heuristics cutting A* node expansions by 42% with paths within 2% of optimal.

Epoch-wise Double Descent Under Label Noise – ResNet-18 / CIFAR-10 writeup GitHub

Python · PyTorch · ResNet-18 · CIFAR-10

- Reproduced Nakkiran et al. (2019) under η up to 30%, isolating the corrupted-accuracy inflection as the mechanistic peak marker; label smoothing hurt, Mixup partially helped — both reported honestly.

Newsphere Live – Semantic 3D News Map writeup GitHub

Python · sentence-transformers · UMAP · HDBSCAN · Three.js · GitHub Actions

- 11 RSS feeds fetched every 6 hours, embedded via MiniLM, reduced with UMAP, clustered with HDBSCAN, rendered in Three.js — fully serverless, zero backend, zero cost, via GitHub Actions.

JEPA-IR – Benchmarking I-JEPA for Image Retrieval writeup GitHub

Python · PyTorch · I-JEPA ViT-H/14 · CLIP · FAISS

- Benchmarked I-JEPA vs CLIP vs ResNet-50 as frozen extractors on Oxford5k/Paris6k; I-JEPA trailed by ∼30 mAP points and near-zero query expansion gain revealed a fundamental embedding space issue, not a tuning problem.

Text-JEPA – Porting the JEPA Objective from Vision to Language writeup GitHub

Python · PyTorch · DistilBERT · WikiText-103 · EMA · SST-2 · STS-B · SVD geometry analysis

- Implemented a text adaptation of I-JEPA (Assran et al. 2023) on DistilBERT: instead of predicting masked tokens, a predictor is trained to match EMA target encoder representations at masked spans — training for 5000 steps on WikiText-103 with contiguous span masking and a cosine loss in representation space.
- The JEPA objective substantially reduced embedding anisotropy (avg pairwise cosine sim 0.93 → 0.60) but degraded downstream task performance (SST-2 81% → 75%, STS-B Pearson 0.48 → 0.16), revealing a clean tradeoff between geometric isotropy and semantic task utility.

Technical Skills

Core	Python, PyTorch, pandas, NumPy, scikit-learn, experiment design, statistical testing, nested CV
Tabular ML	XGBoost, LightGBM, CatBoost, SHAP, Optuna, feature engineering & selection, Wilcoxon, Cohen's d
Deep Learning	Vision Transformers, ResNet/CNNs, self-supervised learning, Mixup, label smoothing, generalisation theory
RL	Q-Learning, DQN, Policy Gradients, PPO, Actor-Critic, Gymnasium
NLP	sentence-transformers, TF-IDF, BM25, c-TF-IDF, topic modelling, NER, UMAP, HDBSCAN
IR	FAISS, dense retrieval, CLIP, multimodal retrieval, mAP, NDCG, Recall@k, query expansion, re-ranking
MLOps	Automated ML pipelines, edge deployment, model compression, heuristic & graph search, Hugging Face, Streamlit

Languages

English – C1 Professional Proficiency *(IELTS Certified)* | **Hindi** – Native