OJASWA YADAV

New York, NY | +1 (646) 238-7470 | ojaswayadav13@gmail.com | LinkedIn | GitHub | Google Scholar (7 papers, 198 citations)

EDUCATION

Columbia University — M.S., Electrical Engineering (AI & ML Specialization) **New York | Dec 2024**

Vellore Institute of Technology — B.Tech, Electrical Engineering **India | July 2023**

EXPERIENCE

CVS Health **New York, NY**

ML Engineer/ Data Scientist (Contract → Full-time) **Apr 2025 – Present**

- Partnered with **underwriters and business stakeholders across 5+ teams** to replace a manual Excel-and-Slack reporting workflow; gathered requirements, drove agreement on scope and priorities, and shipped the production analytics application that became their primary interface.
- Embedded a Claude Opus assistant that summarizes what changed and why in each reporting cycle, removing the analyst round-trips teams relied on to interpret their numbers.
- Owned the full stack: **Flask/Gunicorn backend, Unity Catalog connection layer, and HTML/CSS/JS** frontend on Databricks Apps, with **OBO token auth** scoping every query and LLM response to the requesting user's data permissions.
- Developed and evaluated Text-to-SQL agentic frameworks (**Google ADK, AgentBricks**) over the profit-and-loss model's output data, benchmarking **query generation accuracy, hallucination rates, and edge-case robustness** against complex production healthcare schemas to establish whether agentic querying could serve production support workflows.
- Cut end-to-end runtime **26%** and cloud compute cost up to **15%** in the Financial Impact Testing forecasting pipeline by mapping inter-module data dependencies and building selective execution logic to recompute only affected modules.
- Sole owner of the **MySQL layer** behind both systems — **schema design, indexing, data-quality checks** — triaging production issues with client teams directly into pipeline and schema fixes.

Palindrome Technologies **Princeton, NJ**

AI Engineer (full-time through Mar 2025; part-time ~20–25 hrs/week thereafter) **Jan 2025 – Jan 2026**

- Owned architecture through deployment of a **multi-agent system** automating authorized web application penetration testing across the OWASP Top 10; code-reviewed weekly and presented monthly to the company for technical critique. Authored the **SBIR white paper** defining its technical approach.
- Built an orchestrator coordinating **10 specialized agents**, each targeting one vulnerability class, calling an LLM with role-specific prompts grounded in a knowledge base of known exploits and endpoint patterns.
- Designed a discovery agent that fingerprints targets before testing — enumerating endpoints, detecting tech stack — so downstream agents generate payloads matched to the actual application.
- Implemented a retry-and-escalation control loop: bounded retries on failure, then escalation via **MCP** to live web search, pulling newly disclosed vulnerabilities into the assessment when known exploits don't land.
- Made the pipeline model-agnostic by decoupling agent logic from the underlying LLM; fine-tuned **Mistral 7B on 50K+ samples** as one backend, improving payload accuracy **15%** over base on a held-out test set.
- Deployed with quantization and structured pruning for **sub-200ms** inference during live scans; built the TypeScript layer connecting **10+ scanning APIs** with request queuing and exponential backoff, cutting analyst manual effort **30%**.

Internet Real Time Lab, Columbia University **New York, NY**

Research Assistant **Aug 2024 – Mar 2025**

- Led research on a cost-aware routing system: formulated **multi-detector classification as an episodic MDP and trained a DQN** to route across five detectors spanning a **50,000× inference-time range — 99.54% accuracy** invoking the LLM detector on only 6% of emails, cutting per-email latency from **3.29s to 208ms.**
- Designed a log-normalized cost reward stable at extreme cost ratios and a **Lagrangian dual mechanism** enforcing a **1.5% FPR** constraint, replacing sliding-window approaches that violate MDP reward stationarity; two-phase curriculum with exploration re-injection at phase transition.
- Built the **Graph-RAG multi-agent detector** being routed to: heterogeneous email graph (5 node types, 7 edge types), dual-mode RAG retrieval, and three parallel role-specialized agents with a verdict-deferring orchestrator. Ablations across three datasets isolate per-agent contribution; isotonic calibration on all detector outputs; honeypot filtering and train/test entity-overlap analysis to quantify label leakage.

SELECTED PUBLICATIONS — 7 peer-reviewed papers, 198 citations

Ramamurthy K., Vaichole T., Kulkarni S., **Yadav O.**, Khan F. "Eff2Net: An efficient channel attention-based convolutional neural network for skin disease classification." *Biomedical Signal Processing and Control*, 73:103406, 2022.

Yadav O., Singh A., Sinha A., et al. "ResNet-50-CNN and LSTM Based Arrhythmia Detection Model Based on ECG Dataset." *Studies in Computational Intelligence*, Springer, 2023.

Yadav O., Singh A., Garg C.V., et al. "Design and Simulation of Solar Power-Based Charging Station Using Irradiance Data Prediction with Deep Learning-Based Model." *Lecture Notes in Electrical Engineering*, Springer, 2024.

Vijayarajan R., **Yadav O.**, Khan F., et al. "VaderLogRest Algorithm: An Ensemble Learning Approach for Sentiment Analysis on Vaccination Tweets." IEEE Conference Proceedings, 2022.

PROJECTS

RankSentry — Silent-failure detection for distributed GPU training **Jun 2026 – Sep 2026**

- NCCL reports only "rank unknown" when a job hangs. RankSentry instruments every rank via a**n LD_PRELOAD C shim, streaming collective timings, heartbeats, and GPU telemetry over gRPC to name the failing rank before the watchdog fires** — with cause, evidence, and recovery path. Validated on **Lambda Labs with injected faults.**

SKILLS

Languages & Data: Python (PyTorch, TensorFlow, NumPy, Pandas, Scikit-learn), SQL, C++, TypeScript/JS, MySQL, MongoDB

ML & LLM: LangChain, LoRA/QLoRA fine-tuning, quantization & pruning, RAG & agentic RAG, multi-agent orchestration, MCP, Google ADK, vector DBs

Infrastructure: Databricks, Spark, Docker, Flask/Gunicorn, Airflow, Git, Azure, GCP