Aravinda Raman Jatavallabha

Raleigh, NC | aravindaraman14@gmail.com | (919) 327-0958 | aravinda-1402.github.io | linkedin.com/in/aravinda-jatavallabha

EDUCATION

Master of Computer Science | North Carolina State University, Raleigh, NC — Aug 2023-May 2025
Course Specialization: Data Science — **GPA: 4.0/4.0**

B. Tech in Information Technology | Manipal Institute of Technology, Manipal, India — Jun 2019-Jul 2023
Course Specialization: Big Data Analytics — **GPA: 8.64/10.0**

TECHNICAL SKILLS

- **Programming & Backend:** Python, SQL, TypeScript, JavaScript; Flask, Angular, REST, Event-Driven Systems
- **Machine Learning:** PyTorch, TensorFlow, Scikit-Learn, Transformers, CNNs, GNNs, Time Series Forecasting
- **LLMs & Retrieval:** OpenAI API, Anthropic Claude, HuggingFace, RAG pipelines, Vector Databases, Model Context Protocol (MCP)
- **Data & MLOps:** Feature Engineering, Model Evaluation, Cross-Validation, Drift Detection, CI/CD, Docker
- **Cloud:** AWS (S3, SageMaker, Lambda, ECS, Textract), Linux, Git

WORK EXPERIENCE

AI Software Engineer | Long Health, Raleigh, NC — Jun 2025-Present

- Architected and deployed **serverless ML pipelines** (AWS Lambda, ECS, S3) powering end-to-end clinical documentation workflows; processed **10K+** healthcare documents weekly and improved throughput by **35%**.
- Designed asynchronous, **event-driven inference workflows** using RabbitMQ for OCR (AWS Textract), **RAG processing**, LLM summarization, and ICD-10 inference; reduced processing latency by **30%** with **99.9%** uptime.
- Engineered **RAG-based impairment scoring** and clinical report generation workflows with traceable guideline grounding; achieved score variance within **5%** of clinician-calculated benchmarks.
- Deployed **OpenAI Whisper** for real-time transcription and automated clinical report generation from physician–patient conversations.
- Designed and implemented **OpenAI- and Anthropic Claude–powered pipelines** for structured extraction, real-time summarization, and context-aware Q&A; reduced documentation and triage time by **50%**.
- Built and maintained multi-user physician platform (Angular, NestJS) with **4-role RBAC**, WebSocket-based real-time updates (Soketi), Stripe-based billing workflows, and standards-based impairment calculation engine with auditable reporting.

Data Scientist/Machine Learning Engineer Co-op | SmartProtect Public Safety Solutions, Wilmington, DE — May 2024-Jun 2025

- Developed and A/B tested time series forecasting models (**ARIMA, Prophet, LSTM**) on **1.2M+** emergency call records to predict demand surges and optimize staffing, improving scheduling accuracy by **20%**.
- Designed **clustering-based optimization algorithms** for dynamic staff allocation based on call volume trends and anomalies, reducing overtime by **18%** and increasing resource utilization by **22%**.
- Productionized **ML pipelines** using Flask APIs, AWS SageMaker, and SQL-driven feature engineering with CI/CD automation, reducing retraining time by **35%** and improving deployment reliability.
- Fine-tuned **Azure OpenAI LLMs** and implemented **RAG** over dispatcher transcripts for real-time anomaly summarization and context-aware Q&A, reducing incident triage time by **35**%.

Machine Learning Engineer Intern | Defence Research and Development Organisation, Bengaluru, India — Jan 2023-Jun 2023

- Engineered **Temporal Graph Neural Network (GNN)** leveraging continuous temporal features to predict evolving user interactions, improving model accuracy by **2%** over baseline benchmarks [Paper].
- Integrated **Incremental BERT** (iBERT) with the Temporal GNN to model semantic drift in real-time text streams, achieving **3.19 perplexity (6% improvement over prior SOTA)** in masked language modeling; results published in Springer ICPR 2024 [Paper].

PROJECTS (SELECTED)

- **TrustBench: Benchmarking Trustworthy Large Language Models** [Code]**:** Evaluated GPT-4.1, GPT-4o, GPT-5, and Claude Sonnet 4.5 across stability, ambiguity handling, repeatability, and cost-effectiveness metrics for deployment readiness.
- **CoveredAI - Health Insurance Assistant** [Code]**:** Built a full-stack LLM + RAG system (React, Flask, LangChain, OpenAI, FAISS) enabling semantic search, document summarization, plan comparison, secure OAuth authentication, and exportable reports.
- **Multimodal Conversation Derailment Detection** [Paper]**:** Developed a hierarchical transformer combining BERT and Faster R-CNN for multimodal Reddit thread modeling; achieved 71% accuracy and 78% AUC (+6% over text-only baselines).
- **Privacy Awareness in Large Language Models: Input Regurgitation and Prompt-Induced Sanitization:** Studied regurgitation risk and prompt-based sanitization strategies for privacy-sensitive LLM deployments under HIPAA/GDPR constraints.
- **Store Demand Forecasting with CNN + BiLSTM** [Code]**:** Hybrid deep learning model for item-level sales prediction; outperformed ARIMA and XGBoost baselines with lowest MSE on multi-store retail dataset.
- **COVID-19 Detection using Chest X-rays** [Code]**:** Built and trained a CNN-based multi-class classifier (256×256 input) on Coronahack dataset to distinguish Normal, Pneumonia, and COVID-19 cases; achieved **89.5% validation accuracy** after learning-rate tuning and deployed via Flask for real-time inference.