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 (Data Science Track) | North Carolina State University, Raleigh, NC Aug 2023-May 2025
Courses - Data Science, Natural Language Processing, Neural Networks, Database Management Systems **CGPA: 4.0/4.0**

B. Tech in Information Technology | Manipal Institute of Technology, Manipal, India Jun 2019-Jul 2023
Minor: Big Data Analytics; Courses - Data Mining, Machine Learning, Pattern Recognition, Algorithms **CGPA: 8.64/10.0**

TECHNICAL SKILLS

- **Programming Languages & Frameworks:** Python, SQL, TypeScript, JavaScript, Spring Boot, Angular, Flask, REST APIs
- **Tools & Platforms:** Docker, Git, Linux, Power BI, Azure OpenAI, AWS (S3, SageMaker, Lambda, ECS, Amplify, Textract)
- **Libraries:** Pandas, NumPy, Matplotlib, Scikit-learn, TensorFlow, Keras, PyTorch, Transformers, HuggingFace, ChromaDB, SciPy, PyG
- **Machine Learning:** Time Series Analysis, Classification, Regression, Convolutional Neural Networks (CNN), Natural Language Processing (NLP), Graph Neural Networks (GNN), Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), Prompt Engineering
- **Training & Certifications:** Deep Learning (deeplearning.ai), Machine Learning (Stanford Online), AI Summer School

WORK EXPERIENCE

AI Software Engineer | Long Health, San Jose, CA Jun 2025-Current

- Designed and deployed **serverless workflows with AWS Lambda, ECS, and S3**, processing **10K+ healthcare documents weekly**, improving throughput by **35%** across patient summary and retrieval pipelines.
- Built and maintained **full-stack physician-facing applications** using **Angular (front-end)** and **NestJS (back-end)**, cutting UI load times by **40%** and backend latency by **25%**.
- Implemented **RabbitMQ-based asynchronous pipelines** for OCR, RAG processing, LLM summarization, and ICD-10 inference, incorporating **ChromaDB** vector storage for fast semantic retrieval, ensuring **99.9% uptime** and reducing processing delays by **30%**.
- Integrated **OpenAI LLMs** into healthcare workflows, powering **real-time** patient **document summarization, medical Q&A, structured data extraction,** and an **intelligent physician chatbot**, reducing documentation burden and triage time by **50%**.
- Partnered with clinical and compliance teams to implement HIPAA and PHI/PII-safe AI solutions, achieving **100% audit compliance** while accelerating adoption across physician groups.

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 identify demand surges and optimize shift planning.
- Deployed **RESTful APIs** for scheduling, improving scheduling accuracy by 20% and cutting dispatcher wait time by 14%.
- Productionized machine learning pipelines using **Flask APIs**, **AWS SageMaker**, and **SQL-driven feature extraction**, implementing **CI/CD** automation to reduce retraining time by 35% and enhance deployment reliability.
- Fine-tuned **Azure OpenAI LLMs** and **integrated RAG on dispatcher transcripts** to enable real-time anomaly summarization and context-aware Q&A, reducing incident triage time by 35% and improving operational awareness.
- Built a **full-stack internal dashboard** using **Spring Boot** and **Angular** to display forecasts, trigger LLM-based alerts, and track scheduling KPIs across 3 regional call centers with adoption by 6+ operational teams.
- 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%.

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

- Engineered a **Temporal Graph Neural Network (GNN)**, leveraging continuous temporal data and node features to predict future user interactions on online platforms, increasing model accuracy by 2% over current benchmarks [Paper].
- Developed and integrated **Incremental BERT (iBERT)** with Temporal GNN to capture semantic drift and enhance real-time semantic understanding of evolving text data, reducing data processing time by 40%.
- Achieved 3.19 perplexity (6% better than SOTA) in masked language modeling, published in **Springer ICPR 2024** [Paper].

PROJECTS & PUBLICATIONS

- **CoveredAI - Health Insurance Analysis App** [Code]: Built a **full-stack** AI-powered app using **React**, **Flask (RESTful APIs)**, **LangChain**, and **OpenAI GPT** to analyze, summarize, and compare health insurance documents. Integrated **RAG** (semantic search + chunking via FAISS) for natural language Q&A and plan comparisons. Enabled PDF/DOCX uploads, secure **Google OAuth**, and exportable reports.
- **Multimodal Conversation Derailment Detection** [Paper]: Built a hierarchical **transformer** combining **BERT, Faster R-CNN, and GRU** for multimodal Reddit thread modeling, integrating text and visual cues. Achieved 71% accuracy and 78% AUC, outperforming text-only baselines by 6% in conversational derailment detection.
- **Legal Query AI Assistant** [Code]: Built an AI assistant using LLMs (**OpenAI GPT/LLaMA**) and **RAG** to deliver accurate legal query responses. Combined vector-based retrieval with semantic understanding and deployed a lightweight **Flask** interface for real-time contextual Q&A.