

Summary

AI/ML Engineer with 4+ years of industry experience in building and deploying ML pipelines, multimodal models, and agentic AI systems. Pursuing an MS in Computer Science (AI/ML) at the University at Buffalo (4.0 GPA). Currently building edge AI/ML for autonomous flight systems and multi-agent governance architectures. Published researcher in molecular property prediction and deep learning, with enterprise experience shipping generative AI, real-time data pipelines, and cloud-native platforms on AWS at scale.

Education

Master of Science in Computer Science (AI/ML), University at Buffalo - NY Aug 2025 – May 2027
GPA: 4.0/4.0

Experience

Lead Technical Advisor, IAIMS - Remote June 2026 – Present

- Leading the technical team in architecting a **multi-agent AI governance pipeline** (Standards Monitoring, Control Generation, Evidence Collection agents) that converts regulatory prose across **3+ frameworks** (NIST AI RMF, ISO 42001, EU AI Act) into versioned, executable evidence artifacts using **LLMs, LangGraph**, and human-in-the-loop review gates.
- Designing open-source **compliance-as-code** infrastructure with reusable control schemas, **OSCAL-aligned** evidence templates, and deterministic evidence collectors integrated into **CI/CD pipelines**, targeting **80%** reduction in manual compliance effort.

Agentic AI & Edge AI/ML Engineer Intern, Kuiper Lab May 2026 – Present

- Developing and evaluating advanced **multimodal models** optimized for **edge inference**, achieving **~3x latency reduction** while maintaining accuracy under mobile hardware constraints.
- Designing **agentic AI workflows** that translate live camera feeds and user audio/text prompts into actionable flight logic, processing **30+ FPS** video streams on edge devices.
- Building end-to-end **data pipelines** to collect and preprocess in-flight imagery, telemetry, and multimodal inputs from **software-in-the-loop (SITL)** environments for model training.

AI Research Assistant, Under Prof Wen Yao Xu, University at Buffalo - NY Feb 2026 – Present

- Conducting multi-modal research on AI-driven clinical sensing, spanning wound infection detection and intelligent ICU monitoring systems.

Research Assistant, CAVAS Lab, University at Buffalo - NY Jan 2026 – Present

- Developing a multi-client distributed simulation framework for Digital Twin-based autonomous vehicle testing using CARLA and ROS 2, synchronizing vehicle state and sensor streams across networked server instances.
- Implementing containerized CARLA deployments using Docker for multi-node simulation orchestration and building telemetry ingestion pipelines for offline analysis and model training.
- Integrating LiDAR and camera sensor fusion pipelines within the ROS 2 middleware stack, supporting perception modules for connected autonomous vehicle evaluation.

Senior Cloud Engineer (ML), LTIMindtree | Portland General Electric - India Jan 2021 – June 2025

- Designed and deployed **cloud-native solutions**, replacing legacy systems with scalable, **event-driven architectures** on AWS, achieving **30% cost reduction** and **99.9% uptime**.
- Built **100+ APIs** and serverless ETL pipelines, improving outage detection and alerting speed by **~40%** for **500K+ customers**.
- Developed distributed data pipelines on **Databricks (Apache Spark, Delta Lake)**, optimizing query performance by **50%**; implemented **anomaly detection** in energy trading, reducing risks by **25%**.
- Deployed **smart meter-based proactive outage alert system** (AWS, Twilio) with fault-tolerant pipelines handling **~2M daily events**, reducing storm-time calls by **35%**.
- Architected real-time **Socket APIs** and integrated **Amazon Kendra** with **AWS Bedrock (Claude)**; delivered a **Generative AI assistant** for natural language queries.

- Automated **CI/CD pipelines** (Jenkins) across all environments, reducing deployment time by **60%**.

Cloud Intern, Grepthor Software Solutions – Hyderabad, India

March 2019 – Oct 2019

- Implemented and maintained **AWS CloudWatch** monitoring alarms and dashboards tracking **15+ key metrics**, reducing incident response time by **30%** and improving application uptime.
- Engineered **event-driven, serverless architectures** leveraging **AWS Lambda, S3, and SNS**, enabling modular, scalable, and cost-efficient cloud solutions.

Projects

PriceWatch, AI-Powered Amazon Price Intelligence Agent - GitHub

2026

Developed an AI-powered Amazon price monitor using Playwright, GPT-4o-mini, and ChromaDB; deployed a scalable, fault-tolerant system via Docker Compose with n8n orchestration and a real-time webhook dashboard.

MedRoute, Hybrid RAG + KG Clinical Decision Support - GitHub

2026

Built a locally deployed patient portal with a trained MLP router (2,000 labelled queries) that dynamically selects between LLM-only, patient EHR retrieval, and a biomedical knowledge graph per query. Matched 96% accuracy while reducing per-query compute by 42% and achieving 99% recall on drug-interaction questions. Stack: Gemma2-9B, MedCPT, FAISS, PrimeKG (4M relations), MedlinePlus, Streamlit.

RefFlag, AI-Powered Chrome Extension - GitHub

2025

Built browser extension using Chrome Extension API (Manifest V3) with background service workers for DOM manipulation and text extraction. Integrated LLM inference endpoints for real-time privacy policy analysis using prompt engineering and semantic chunking. Implemented asynchronous content script injection with MutationObserver for dynamic page monitoring.

Awards and Publications

MolFM-Lite: Multi-Modal Molecular Property Prediction

2025

Developed a lightweight multi-modal deep learning model integrating 1D, 2D, and 3D molecular representations using cross-attention and FiLM conditioning. - arxiv.org/abs/2602.22405

Shooting Star Award, LTIMindtree

2024

Recipient of the Shooting Star Award for key contributions in the Cloud Pillar at LTIMindtree.

An Approach to Translate Human Gestures into Telugu

2022

Published in *International Journal of Scientific and Technology Research (IJSTR)*. [IJSTR Link](#)

Question Answering System using Deep Learning

2021

Published in *Journal of Xidian University*. [Journal Link](#)

Technical Skills

Languages: Python, Java, JavaScript, TypeScript, SQL, Bash/Shell

AI/ML & Data Science: PyTorch, Hugging Face Transformers, LangChain, RAG, PySpark, Pandas, NumPy, Optuna, OpenClaw

Databases: PostgreSQL, MySQL, Snowflake, DynamoDB, Pinecone, ChromaDB

Frameworks & Tools: FastAPI, Spring, React, REST APIs, Postman, JMeter

Cloud (AWS): EC2, Lambda, API Gateway, S3, RDS, Step Functions, EventBridge, SNS, SQS, Glue, SageMaker, Bedrock, Kendra, OpenSearch, IAM, VPC, CloudWatch, CloudFormation

Data Platforms: Databricks, Delta Lake, Apache Spark, Data Lakes, Matillion

DevOps & Infrastructure: Docker, Kubernetes, Terraform, CI/CD (Jenkins), Git, Playwright, n8n

Certifications

- Oracle Generative AI Professional
- AWS Developer Associate
- Oracle AI Fundamentals Associate
- Databricks AI Agent

- AWS Solutions Architect
- Deep Learning (IIT Madras)
- AWS Databricks Platform Architect