

Akhil Reddy Danda

☎ Phone: +1 (224) 999-9944 ✉ Email: akhilreddydanda1@gmail.com 🔗 [Linkedin](#)

❑ EDUCATION

Master's in Computer Science, Texas A&M University, Texas, United States

Dec 2021

Bachelor's in Computer Science, SRM Institute of Technology, India

May 2019

❑ TECHNICAL SKILLS

Programming Languages
Cloud Platforms

Python, Java, C, C++, Go, SQL, JavaScript, Bash, Ruby
Azure: Data Factory, Data Lake, Synapse, ML Studio, Event Hubs, Stream Analytics, Kafka, Load Balancing; **AWS:** Lambda, S3, SQS, SNS, EC2, SageMaker, CloudWatch, EKS, IAM. REST APIs, gRPC, Spring Boot, Microservices Architecture, Event-Driven Design, HTML5, CSS3, JavaScript, React, Angular
PostgreSQL, DynamoDB, MySQL, Azure SQL Database, MongoDB, Redis, Data Lake Storage, Cosmos DB
Docker, Kubernetes, Jenkins, Git, GitHub, Azure Devops
Power BI, DAX, Pandas, NumPy, Azure Stream Analytics
TensorFlow, PyTorch, Scikit-learn, OpenCV, NumPy, Pandas, Hugging Face, Model Deployment, A/B Testing, LangChain, LangSmith, LlamaIndex, Hugging Face, RAG, Vector Embeddings(Pinecone, ChromaDB, FAISS), PEFT, LoRA, QLoRA, Embeddings, Semantic Search, Fine-tuning(SFT, RLHF)

Web Technologies

Databases

CI/CD & DevOps Tools

Data & Analytics

Machine Learning/AI

LLM & GenAI

❑ PROFESSIONAL EXPERIENCE

Senior Software Engineer, HHA Medicine, Corpus Christi, TX

Jan 23 – Present

- Architected large-scale **distributed data infrastructure** on Azure processing **100GB+ daily** healthcare data with sub-second query latency, **99.9% uptime**, and **fault-tolerant design**
- Built scalable **ML pipelines** serving 15+ hospitals using Azure ML and **Azure Event Hubs**, implementing event-driven message processing for real-time healthcare data ingestion
- Built **RAG (Retrieval-Augmented Generation) systems** using Azure Cognitive Services and **LangChain** to analyze clinical data and provide intelligent insights improving response relevance by 40%
- Developed **LLM** powered clinical decisions support system using GPT-4 and custom **embeddings**, serving 50K+ daily clinical decisions.
- Implemented vector database (Pinecone) with horizontal scaling for semantic search across medical records, reducing document retrieval time by 60%
- Implemented **vector database (Pinecone)** for semantic search across medical records, reducing document retrieval time by 60%
- Established **A/B testing framework** for ML model iterations, improving prediction accuracy by 23% over 6 months

Software Development Engineer Amazon, Seattle, WA

Feb 22 – Jan 23

- Architected **event-driven microservices** using AWS Lambda, SQS, SNS, and **Kafka** (like Azure Service Bus/Event Hubs) for real-time **GNSS data** processing, reducing data pipeline latency by 65% and handling 100K+ concurrent requests/second at peak load
- Designed containerized workloads on **Kubernetes/EKS** (transferable to AKS) serving 1000+ fleet vehicles with auto-scaling, load balancing, and high availability, improving deployment frequency by 40% and reducing infrastructure costs by 25%
- Implemented **distributed caching layer** using **Redis** to reduce database load by 50% and achieve **p99 latency** under 50ms for critical API endpoints
- Developed and **deployed CNN** based computer vision models processing **500K+** daily images achieving 95% accuracy (up from 72% baseline) through feature engineering and **hyperparameter tuning**.
- Built ML deployment pipeline using **SageMaker for edge inference**, cutting model inference latency by 40% and enabling real-time object detection at 30 FPS on edge devices
- Implemented automated **model retraining pipeline** processing 100K+ labeled samples weekly, improving model accuracy by 23% over 6 months through continuous learning
- Reduced **camera feed analysis** processing time by 60% by optimizing video streaming architecture and implementing **batch inference strategies**

❑ ACADEMIC PROJECTS

Advanced Driver Assistance System (ADAS) using Computer Vision

Publication: “**Advance Driver Assistant System using Artificial Intelligence**” (IJRASET, Apr 2019), DOI: 10.22214/ijraset.2019.415

- Published research on **real-time driver monitoring** achieving 92% accuracy in drowsiness/distraction detection using **CNN, YOLO v3, OpenCV, TensorFlow** on Raspberry Pi edge devices.
- Implemented **multi-threaded architecture** processing 30 FPS for lane departure warning, collision avoidance, traffic sign recognition

Transformer Based Sentiment Analysis for Financial Markets

- Built **BERT-based sentiment analysis model fine-tuned** on 500K+ financial news articles and SEC filings to predict stock price.
- Implemented attention visualization to identify key phrases driving sentiment predictions, achieving 82% accuracy on market direction forecasting. Developed **custom tokenizer for financial domain vocabulary**, improving model F1-score by 18% over baseline BERT