SUDARSHAN RAJAGOPALAN

Software Engineer

sudarshanr2308@gmail.com +91 8072263579 Chennai, Tamil Nadu Linked In Github

PROFILE

Software Engineer with experience building and deploying backend and AI-powered systems on AWS. **AWS Certified Machine Learning Engineer** with hands-on experience using **Python (Django, FastAPI)** to integrate **LLM-based workflows**, including **RAG** systems, into production-grade microservices. Focused on scalable, secure deployments using containerization and CI/CD pipelines.

SKILLS

Programming Languages — Python, Java, C++, Go, JavaScript, TypeScript, SQL, NoSQL, **AI & Machine Learning** — Retrieval-Augmented Generation (**RAG**), Large Language Models (**LLMs**), **AI Agents**, Vector Databases (Pinecone), Deep Learning, NLP, Prompt Engineering, Model Context Protocol (**MCP**), Generative AI, LangChain, LangGraph, Time Series Forecasting, PyTorch, TensorFlow, **Backend & Cloud** — Django, FastAPI, FastMCP, REST APIs, WebSockets, Microservices, AWS (ECS Fargate, S3, SageMaker, Lambda), Docker, PostgreSQL, **Frontend** — React, HTML/CSS, **Tools & DevOps:** — Git, GitHub Actions, Azure Pipelines, CI/CD, Linux

WORK EXPERIENCE

Associate Engineer (AI Platform) 07/2025 – Present | Remote
Norstella

- **Engineered** an automated interview analysis pipeline using **Gemini** and **Python**, slashing processing time by **95%** (from hours to <3 minutes) for high-value medical KOL transcripts.
- **Built** the core **Django** backend for a RAG-powered chat application, designing a normalized **PostgreSQL** schema to handle complex multi-turn conversation history and compliance audit trails.
- Implemented **real-time streaming** of **LLM responses** using **WebSockets** and **Django Channels**, enabling partial responses and improving perceived latency during interactive analysis.
- Developed a standalone **MCP-based retrieval service** with **Pinecone-backed RAG**, exposing **pluggable tool interfaces** reusable across agents and services without modifying core agent logic.
- **Deployed** containerized microservices to **AWS ECS Fargate** using **Docker**, utilizing CI/CD pipelines to automate testing and ensure secure, scalable production releases.

Generative Al Intern 01/2025 – 07/2025 | Bangalore, IN
Titan Company Limited

- Built an **AI-powered Text-to-SQL service** using **LangChain** and OpenAI **LLMs**, achieving **87% query accuracy** and reducing manual query effort.
- Developed a **computer vision–assisted analysis pipeline** using **vision models** to extract structured insights from competitor product images, identifying **three key feature gaps** that informed product strategy.
- Designed and deployed a **real-time analytics dashboard** using **Streamlit**, enabling live sales trend monitoring and improving analyst productivity by **30%**.

Data Science Intern 08/2023 – 10/2023 | Bangalore, IN
Titan Company Limited

- Developed and evaluated **time-series forecasting models** for watch sales prediction across three sales channels, achieving **97%, 94%, and 89% accuracy(low MAPE)**.
- Processed and cleaned **historical sales data (2011–present)**, removing **98% of inconsistencies** through validation and normalization, resulting in a **20% improvement in forecast accuracy**.
- Implemented and tuned **SARIMAX models** for demand forecasting, reducing computation time by **40%** while decreasing overstocking and stockouts by **20%**.

CERTIFICATIONS

- AWS Certified Machine Learning Engineer - Associate
- Microsoft Certified: Azure Al Fundamentals - Microsoft
- Microsoft Certified: Azure Data Fundamentals - Microsoft
- Fundamentals of Deep Learning - NVIDIA

EDUCATION

B.Tech in Computer Science with Specialization in Artificial Intelligence and Machine Learning 2021 – 2025
Vellore Institute of Technology, Chennai
CGPA - 8.92

PROJECTS

Intelligent Email-Tasks Manager

- **Engineered** an automated task extraction agent using **Google Gemini** and **Gmail API**, successfully parsing and scheduling **80%** of actionable emails into Google Tasks without manual intervention.
- **Built** a resilient background processing pipeline (using Cron/Celery) to poll and process emails every 15 minutes, ensuring near real-time task synchronization and reducing user response delays by **50%.**
- **Implemented** entity extraction and intent classification using NLP to prioritize tasks and auto-generate smart reminders, increasing personal workflow productivity by **~35%.**

Speech Emotion Recognition: Dual Model Emotion Detection system

- **Developed** a **multimodal** classification system fusing **BERT** (for text) and **CNNs** (for audio MFCCs/Spectral features), achieving **93.2% accuracy** on text capability and robust performance on noisy audio samples.
- **Optimized** feature extraction pipelines to capture Spectral Centroid and Zero Crossing Rates, improving emotion detection reliability across 6 categories by **40%** compared to single-modality baselines.

Brain Tumor Detection Assistant (BTDA)

- **Designed** a medical diagnostic model using **CNN-LSTM** architecture for MRI analysis, achieving **92.8% accuracy** in classifying pituitary, glioma, and meningioma tumors.
- **Integrated Google Gemma (LLM)** to auto-generate patient-friendly diagnostic summaries from model inference logs, improving non-technical interpretability by **20%.**
- **Implemented Grad-CAM** visualizations to highlight tumor regions in MRI scans, providing **Explainable AI (XAI)** insights to increase radiologist trust in model predictions.