Anany Sharma

+1-3523286457 | Gainesville, Florida | anany.sharma.ufl@gmail.com | linkedin.com/in/ananyd36/ | github.com/ananyd36

EDUCATION

University of Florida ***Expected Graduation: May 2026***

Masters of Science in Artificial Intelligence Systems | 3.8 GPA

SRM Institute of Science and Technology ***Graduated: July 2021***

Bachelors of Technology in Computer Science | 3.7 GPA

SKILLS

Languages Python · TypeScript/Javascript · C++ · Dart · Java · SQL · NoSQL

AI/ML PyTorch · TensorFlow · JAX · Hugging Face · scikit-learn · OpenCV · Generative AI · NLP · Computer Vision

MLOps & Data MLflow · Spark · Airflow · Ray · Kafka · PowerBI · Pinecone · ETL

Software & Cloud AWS · Azure · Docker · Kubernetes · REST/gRPC · React.JS · Jenkins · Supabase · Git

EXPERIENCE

Machine Learning Engineering Research Fellow ***March 2024 - Present***

UNIVERSITY OF FLORIDA *Gainesville, Florida*

- Optimizing **SLM** inference on **Jetson Nano**, conducting edge-cloud benchmarking to improve AI efficiency on the edge
- Built and deployed **5+ edge AI** pipelines on ESP32 MCU integrating light, motion, and ToF sensors with **90%+ accuracy**
- Reduced inference time by **50% (3 ms to 1.5ms.)** via quantization **(float32 to int8)** on ESP32 MCU deployment
- Designed hands-on AI curriculum to **100+** high school students, combining ML, sensors, and embedded systems

Software Engineer - AI/ML ***Jan 2022 - Aug 2024***

UNITED HEALTH GROUP

- Engineered a **RAG** microservice using **vector DB and Azure AI Search**, enhancing retrieval speed and efficiency by **40%**
- Architected a claims validation service with **Azure Document Intelligence**, improving claim adjudication rate by **30%**
- Designed low-latency data pipelines for real-time **KPI dashboards**, reducing reporting latency by **40%** using **PowerBI**
- Optimized reporting pipelines (**SSIS/SSRS**) to generate finance reports for **1.5K+** clients, to cut generation time by **20%**
- Architected high-volume ETL pipelines to process **10M+** monthly claims, focusing on data integrity and throughput

Data Engineer ***March 2021 - Dec 2021***

INFOCEPTS

- Optimized **PySpark** ETL pipelines on **Amazon EMR** to process 5TB+ monthly data, reducing batch runtime by **20%**
- Built custom **Lambda** functions with **S3 triggers** to ingest and preprocess data from **20+ client sources** into EMR
- Designed **10+ Airflow DAGs** for incremental data loads, cutting ingestion latency by **40%** and **99% SLA** adherence

Machine Learning Engineering Intern ***July 2020 - Nov 2020***

MFIT TECHNOLOGIES

- Developed an NLP-powered financial data system with **85% field-extraction accuracy** for transaction monitoring
- Engineered document extraction system supporting **10+ formats** using hybrid NER and layout-aware frameworks
- Built dynamic **Conditional Random Field** plus **Spatial** model to parse financial statements across **4 major banks**

PROJECTS

SMIRE AI - Medical Multi AI Agent System | Individual Project *(~60 hours)* - GITHUB ***Feb 2025 - Present***

- Architected a **RAG service** using **Pinecone** to enable conversational access to medical reports and prescriptions
- Engineered an agentic system with **CrewAI** and **LangChain**, leveraging **Chain of Thought (CoT)** and **ReAct** prompts
- Integrated an appointment booking system with **Twilio** and a **calendar API** to automate personalized scheduling
- Deployed the full-stack platform via Docker on Vercel, with **Prometheus** and **Grafana** for system monitoring

EMOGEN - Emotion Aware Multilingual Speech to Image Generation | Individual Project *(~30 hours)* - GITHUB ***Mar 2025 - April 2025***

- Engineered a multi-modal generative AI pipeline to integrate speech, emotion, and vision for **contextual image synthesis**
- Fine-tuned a **multilingual ASR model** (OpenAI Whisper) to optimize latency for real-time speech transcription and translation
- Built a **CLIP-guided Stable Diffusion** model for emotionally and semantically aligned image generation
- Integrated **HumeAI** Emotion API to inject voice-based emotional context into image synthesis

TRADE-MCP - Remote MCP Server for Real time Zerodha Trading | Individual Project *(~20 hours)* - GITHUB ***Oct 2024 - Oct 2024***

- Developed and deployed a custom remote **Model Context Protocol (MCP)** server enabling GitHub Copilot and Cursor
- Integrated MCP server with **Zerodha Kite API** for real-time market data analysis, order placement, and portfolio insights
- Built backend using **Next.js** and **TypeScript**, deployed on Vercel for scalable, low-latency performance

PATENT

Crowd Detection and a Method Thereof - IIP(2021) | Published Patent *(~80 hours)* - PATENT ***Jan 2021 - July 2021***

- Developed and patented a **YOLOv5**-based multimodal crowd detection system using **Python**, **Flask**, and **OpenCV**, designed to identify high-risk crowd behavior and Mask/PPE compliance in real time
- Achieved up to **85% AP** for mask detection and **80%** for no-mask cases, optimized model performance with **transfer learning** and **mAP** evaluation, and deployed the solution with **Docker** for scalability