Anany Sharma

Jacksonville, Florida | anany.sharma@ufl.edu | Linkedin.com/in/ananyd36/ | GitHub.com/ananyd36 | AnanySharma.com

EDUCATION

University of Florida   *Expected Graduation: May 2026*
Masters of Science in Artificial Intelligence Systems | 3.8 GPA

SKILLS

Languages   Python(5yrs) · Javascript(3yrs) · C++(2yrs) · Dart(1yrs) · Java(1yrs) · SQL(6yrs) · NoSQL(2yrs)
Software   AWS · Azure · Langchain · Docker · OpenAI · TensorFlow · JupyterNotebook · Pytorch · Supabase · Git · Pinecone

EXPERIENCE

Machine Learning Engineering Research Fellow   *March 2024 - Present*
UNIVERSITY OF FLORIDA   *Gainesville, Florida*
- Developed and deployed multiple edge AI models on AHA board utilizing 5+ sensors, achieving 90%+ accuracy
- Optimized edge AI models through Edge Impulse, reducing inference time by 40% while maintaining model accuracy
- Leading development of an edge AI curriculum for 200+ students across K-12 and undergraduate levels

Software Engineer   *Jan 2022 - Aug 2024*
UNITED HEALTH GROUP
- Implemented RAG agents using Azure AI Search and OpenAI, increasing retrieval speed by 40% and adoption by 50%
- Automated claims validation with Azure Document Intelligence and NLP, improving processing accuracy by 30%
- Created real-time Power BI dashboards for KPIs, reducing reporting latency by 40%
- Automated biweekly financial reports for 1.5K+ users, cutting generation time by 20%
- Architected ETL pipelines to process 10M+ monthly claims using IBM WTX on UNIX AIX

Machine Learning Engineering Intern   *July 2021 - Nov 2021*
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 CRF-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*
- Used CrewAI, LangChain to engineer agent systems, with multiple iterations of prompt refining using CoT, ReAct techniques
- Created an end-to-end solution complete with booking automated personalized appointments for customer
- Engineered consulting agents, news agents and search agents for better medical assistance
- Architected logging doses/prescriptions/reports and chatting for insights using Retrieval Augmented Generation

CORAS - Context-Based Intelligent Knowledge Retrieval System | Individual Project *(~40 hours)* - GITHUB   *Sept 2024 - Dec 2024*
- Built a Retrieval-Augmented Generation (RAG) system using React, Pinecone, Flask, OpenAI, integrating multimodal capabilities to handle both text and audio data.
- Leveraged Prometheus monitoring system performance and Grafana for visualizing traffic and system metrics

CAPTIVISION - Image Captioning System | Individual Project *(~15 hours)* - GITHUB   *Mar 2025 - Mar 2025*
- Developed an image captioning system using a Vision Transformer (ViT) encoder and GPT-2 decoder, leveraging Hugging Face's transformers for feature extraction and text generation
- Deployed the model with Streamlit for interactive, real-time image captioning via a web interface

RESILITREE - Tree Fall Prediction/Safety Assistant | Team Project (~10 hours) UF/IBM Hackathon - GITHUB   *Sep 2024 - Oct 2024*
- Developed AI-driven disaster preparedness system utilizing PyTorch for tree fall risk prediction and an IBM Watson-based chatbot for real-time safety guidance
- Implemented a Streamlit-based web app for real-time tree stability prediction and disaster relief, leveraging image processing and contextual AI responses for user engagement

VESSELVIEW - Ship Classification in Satellite Imagery | Individual Project *(~10 hours)* - GITHUB   *Oct 2024 - Oct 2024*
- Developed 3 iterations of a ship detection system using dimensionality reduction (PCA, Isomap) and machine learning models (Logistic Regression, Random Forest) to classify ships in satellite imagery with a 95% accuracy
- Implemented PCA and manifold learning to reduce computational complexity and improve inference time, achieving an optimal F1 score of 0.91 and 0.1126 seconds inference time

PATENT

Crowd Detection and a Method Thereof - IIP(2021) | Published Patent *(~80 hours)* - PATENT   *Jan 2021 - July 2021*
- Developed a YOLOv5 multimodal detection system with high-risk crowd detection using Python, Flask, OpenCV
- Implemented real-time alerts and optimized model performance through transfer learning and mAP evaluation
- Deployed the system with Docker for scalability, achieving an average precision of 60% for PPE, 85% AP for mask, and 80% AP for no mask.