

Email: sanjulagihan94@gmail.com

GitHub: <https://github.com/sanjula77>

Phone: 0714858298

GIHAN SANJULA

LinkedIn: <https://www.linkedin.com/in/gihan49/>

AI Engineer / ML Engineer / MLOps Developer

Portfolio: <https://polia-portfolio.vercel.app/>

PROFESSIONAL SUMMARY

Results-oriented Information Technology undergraduate specializing in Machine Learning, GenAI, and MLOps. Experienced in building scalable AI systems, RAG pipelines, and predictive models using Python, TensorFlow, PyTorch, LangChain, and FastAPI. Strong engineering mindset with expertise in ML system design, vector search, CI/CD, and cloud deployment. Passionate about bridging the gap between AI model development and production-grade software engineering.

SKILLS

- GenAI & LLM Engineering:** LangChain, RAG Pipelines, LLM Integration, Google Gemini, Vector Databases (FAISS, Redis), Embeddings, ReAct Agents, Semantic Search
- Machine Learning & Deep Learning:** Python, NumPy, Pandas, Scikit-learn, TensorFlow, PyTorch, OpenCV, Model Training & Optimization, Predictive Modeling, MLflow, MLOps
- Backend & Application Development:** FastAPI, Streamlit, Next.js, Node.js, Express.js, REST API Design, Authentication (JWT)
- Cloud & DevOps:** Docker, GitHub Actions, Jenkins, AWS EC2, Oracle Cloud, CI/CD Pipelines, Containerized ML Services
- Databases:** MongoDB, PostgreSQL, MySQL, Firebase, Supabase, Redis
- Tools & Additional Skills:** Git, GitHub, Postman, VS Code, Clean Architecture, Microservices, API Engineering

PROJECT EXPERIENCE

Predictive Maintenance MLOps Pipeline | Python, TensorFlow, Docker, MLflow, Oracle Cloud. [Project Repository](#)

- Built RUL prediction models on NASA CMAPSS, achieving **RMSE 24.04** and **MAE 16.81** cycles.
- Implemented MLflow tracking with automatic model versioning, metadata logging, and best-model promotion.
- Deployed FastAPI inference service using Docker containers with health checks and automated CI/CD pipelines.
- Delivered reproducible MLOps workflow with model registry, automated testing, and real-time predictions on Oracle Cloud.

AI Customer Support Assistant (RAG + LangChain + FastAPI) | GenAI, FAISS, Redis, Next.js. [Project Repository](#)

- Developed an intelligent customer support system using RAG retrieval, LangChain ReAct agents, and Google Gemini models.
- Implemented FAISS vector search and Redis conversational memory for low-latency semantic retrieval.
- Added multi-tool execution for order tracking, ticket creation, and profile updates.
- Delivered a real-time Next.js interface with citations, streaming responses, and session memory.

Web Behavior Insight Engine | Python, XGBoost, Streamlit, FastAPI. [Project Repository](#)

- Engineered an end-to-end ML platform for user behavior analytics and segmentation.
- Developed clustering + predictive models (RF, XGBoost, LR) achieving **93%+ accuracy** and **0.97 ROC-AUC**.
- Built a hybrid recommender system using semantic embeddings.
- Delivered FastAPI backend + Streamlit dashboard with anomaly detection and LLM-driven insights

Voice-Controlled 3D AI Assistant | React, Three.js, Web Speech API. [Project Repository](#)

- Built a real-time 3D AI assistant with voice interaction, lip-sync animation (morph targets), and emotion-based visual responses.
- Integrated multi-model fallback (DeepSeek, Gemini, Llama, Qwen, Mistral) using a modular plugin architecture.
- Achieved **60fps** 3D rendering using React + Three.js with optimized animation pipeline.

Face Recognition Pipeline (InsightFace + ArcFace) | Python, OpenCV, Scikit-learn. [Project Repository](#)

- Designed a full face-recognition pipeline including preprocessing, embedding extraction, and classifier optimization.
- Integrated InsightFace/ArcFace embeddings with tuned classifiers achieving **95%+ accuracy** in live and batch inference.

EDUCATION

Horizon Campus

BSc (Hons) in Information Technology

CERTIFICATIONS

NVIDIA – Transformer-Based NLP

Kaggle – Deep Learning; Data Cleaning; Data Preprocessing

SLIIT – AI/ML Engineer Program (Stage 1)