KARAN DESAI

AI/ML Engineer (GenAI-Focus)

☎ 9970843464 @ karansd00@gmail.com Linkedin Porfolio Jaysingpur, Maharashtra

SUMMARY

Enthusiastic AI/ML professional with hands-on experience in computer vision, machine learning, and cloud-based model deployment. Skilled in Python, TensorFlow, and AWS, with growing expertise in Agentic AI and Agentic RAG systems. Passionate about building intelligent, adaptive solutions that enhance model performance and drive organizational innovation.

EXPERIENCE

Zensar Technologies

10/2025 - Present

- Supporting model experimentation and fine-tuning for object detection and classification tasks using Python, TensorFlow, and OpenCV.
- Building and iterating on RAG-based and Agentic AI pipelines using LangChain, LangGraph, and LangSmith for intelligent automation and retrieval workflows.
- Collaborating with the ML team to maintain consistent labeling standards, enhance data quality, and optimize training data for scalable AI systems.

ResoluteAI — Gen-ai Intern

June 2025 – Sep 2025

- Built an Agentic Video QA system for offline and non-audio videos using Gemini API and Hugging Face models.
- Developed document processing pipeline with LayoutLMv3 for structured data extraction.
- Implemented real-time multilingual voice transcription for Indian languages using AI4Bharat models.
- Created ML multi-regression model for predicting color tint adjustments.
- Focused on secure offline/local inference with Hugging Face models..

Education

02/2021 - 10/2024

Bachelor of Technology (B.Tech) – Sharad Institute of Technology College of Engineering

CGPA: 7.6 / 10

01/2019 - 01/2020

Higher Secondary Certificate (HSC) – A.N.N. Junior College

Certifications

AWS Academy – Machine Learning Foundation: Fundamentals of machine learning using AWS cloud.

AWS Academy – Cloud Foundation: Hands-on training in deploying ML models on AWS cloud.

Deep Learning using AI-Jetson Nano: Practical applications of deep learning on **NVIDIA Jetson Nano.**

C3SA (Certified Cyber Security Analyst) – CyberWarFare Labs: Training in cybersecurity and threat analysis.

ACADEMIC PUBLICATIONS:

Decentralized File System Using Blockchain — Published in IRJMETS (e-ISSN: 2582-5208), DOI: 10.56726/IRJMETS45522, Sharad Institute of Technology, India.

SKILLS

- **Programming Languages**: Python, C++,System Design,DSA
- **ML & DL Optimization:** Logistic Regression, SVM, Decision Trees, Gradient Descent, Adam, RMSProp, L1/L2 Regularization, Dropout, Early Stopping, Feature Engineering, Evaluation Metrics (Precision, Recall, F1, ROC, AUC), LoRA, QLoRA, NVFP4, DL Quantization
- **Deep Learning:** TensorFlow, Keras, PyTorch, CNNs, RNNs, ViTs, Attention, Transformers, Transfer Learning
- **Computer Vision**: YOLOv5, OpenCV, Supervision, ResNet, PaddleOCR, face_recognition
- **Natural Language Processing**: Hugging Face Transformers, DistilBERT, Spacy, NLTK
- **Generative & Agentic AI:** LangChain, LangGraph, LangSmith, SentenceTransformers, Whisper, N8N, AWS Agentic Stack
- **Audio Processing**: Speech-to-Text (Whisper),FFmpeg,
- **Web & Deployment**: Flask, Streamlit, FastAPI
- **Cloud & DevOps**: AWS EC2,Lambda,AWS sagemaker
- **Database & Retrieval**: FAISS, Pinecone, ChromaDb,Mysql
- **Version Control** : Git, GitHub
- **MLOps**:Agentic AI ,Image/NLP Pipelines, Feature Stores, Vector Search, Chatbots (RAG + Embeddings), Training–Deployment–Monitoring Lifecycle

PROJECTS

Self-Healing Classification DAG with Fine Tuned Model

Developed a Self-Healing Classification DAG using DistilBERT fine-tuned with LoRA, implementing LangGraph's self-healing workflow with confidence checks and human-in-the-loop fallback, ensuring robust sentiment classification with CLI and logging.

Agentic Video QA (Offline & Silent Videos)

•Developed a hybrid agentic system for answering questions about videos with and without audio. Utilized FFmpeg for audio extraction, Whisper for transcription, and BLIP-2 OPT for generating captions from frames in silent videos. Combined captions and transcriptions into a local LLM for question answering, with Gemini API for enhanced generative responses.

Indic Voice Transcription & Analytics (AI4Bharat ASR)

•Designed a multilingual speech-to-text pipeline using AI4Bharat Indic models for real-time voice transcription in Indian languages. Built an analytics layer to detect conversation tone (e.g., rudeness) and infer payment modes (cash vs. online) from call center interactions. Enabled offline inference for secure enterprise deployments.

Document Intelligence with LayoutLMv3 & Gemini 2.5 Pro

•Created a Document AI solution to extract structured information from invoices, contracts, and forms using LayoutLMv3. Integrated Gemini 2.5 Pro for contextual analysis and deployed in a RAG workflow for enterprise knowledge systems.