Ali Raza

+92 300 7651218 mirzaalirazafsd@gmail.com ali-raza-0a1282281 alyrraza alyrraza.vercel.app

Summary

ML Engineer with 2 years of experience building and deploying machine learning, deep learning and LLM-based systems, from data preparation to production. Technical lead on a live AI SaaS for US clients. Evaluates every model on held-out data (precision, recall, AUC, mAP) before shipping.

Technical Skills

ML/DL: PyTorch, TensorFlow, Keras, Scikit-learn, HuggingFace, XGBoost, LightGBM, Transfer Learning, Fine-tuning, LoRA, QLoRA, PEFT, Random Forest, feature engineering, Matplotlib, Seaborn
Computer Vision: YOLOv8, U-Net, TransUNet, ViT, SAM, ResNet, DenseNet, BioViL-T, OpenCV, GradCAM, CBAM
LLMs & NLP: LangChain, LangGraph, LlamaIndex, RAG, Generative AI, FAISS, Pinecone, BERT, Embeddings, Prompt Engineering, OpenAI API, Gemini API
MLOps: MLflow, Docker, Kubernetes, Kubeflow, FastAPI, REST APIs, Prometheus, Grafana, GitHub Actions, CI/CD, AWS EC2, ONNX Runtime, TensorRT
Programming: Python, C++, JavaScript, Pandas, NumPy, SQL, PostgreSQL, Git

Experience

ML Engineer — HusQuay Solutions 2025 – ongoing
Remote, US-based Client

- Trained computer vision and classification models for a production AI SaaS platform automating property tax depreciation studies for US clients.
- Integrated open-source detection/segmentation models with third-party **VLM/LLM APIs** into a unified inference pipeline, replacing planned GPU retraining with an API-driven approach.
- Owned ML pipeline architecture end-to-end, from raw image input to report-ready output, shipping features directly into a live production system.

Data Analyst Intern | TAKMIL 2024
Louisville, Kentucky, USA · Remote

- Cleaned and analyzed large-scale datasets via Python (Pandas, NumPy, SQL); built **Power BI dashboards**; proposed optimizations improving workflow efficiency by **30%**.

Research & Key Projects

RadGuard: Multimodal ML System for Medical Imaging — Final Year Thesis 2025–2026
BioViL-T · CXR-BERT · PyTorch · MLflow · Docker · Kubernetes · Kubeflow · FastAPI · AWS EC2

- Detected 4 error types across 14 conditions in AI-generated chest X-ray reports using 74K+ MIMIC-CXR samples, reaching **0.66 validation F1**.
- Deployed a production MLOps pipeline (**Docker, Kubernetes, Kubeflow, MLflow, FastAPI, AWS EC2**) with automated retraining and FAISS-powered RAG querying.

AI Cost Segregation Platform — Production SaaS 2025–Present
YOLOE · SAM2 · Gemini VLM · FastAPI · Firebase · Cloudinary · ReportLab

- Technical lead on a full-stack AI SaaS turning property photos into IRS-defensible tax reports; redesigned a failing classifier into a **3-stage YOLOE + SAM2 + Gemini VLM pipeline** across 312 categories.
- Cut inference cost by moving GPU rental to **CPU-only VPS + hosted VLM inference**; owned the MLOps roadmap (Docker, GitHub Actions CI/CD, MLflow, HITL loop).

PIT-Scorecard: Point-in-Time Credit Risk Scoring System 2026
WOE Scorecard · LightGBM · FastAPI · DuckDB · Docker · Hugging Face Spaces

- Quantified a **0.30 AUC leakage gap** on 2.26M Lending Club loans via a point-in-time feature contract and out-of-time validation split.
- Built a WOE scorecard (**AUC 0.6922**) against a LightGBM challenger, holding a consistent **0.004–0.012 AUC edge** across 5 folds; deployed via FastAPI with drift monitoring on Docker and Hugging Face Spaces.

InferBench: ML Inference Optimization & Serving Platform 2026
ViT-Base · ONNX Runtime · TensorRT (FP16) · FastAPI · Redis · PostgreSQL · Prometheus · Grafana

- Optimized ViT-Base inference with ONNX Runtime and TensorRT (FP16), cutting latency **2.9×–5.6×** with 100% prediction consistency preserved.
- Built a serving platform with custom dynamic batching, Redis caching (**2377× faster** on repeats), and Prometheus/Grafana monitoring, shipped via Docker CI/CD.

Education

BSc Computer Science | FAST-NUCES, Faisalabad 2022 – 2026
Co-Head Hackathon ACM Daira 2025 | TAKMIL Data Science Internship 2024

Publications

Khan MM, **Raza A**, et al. "Generalizable Detection of CIEDs Using a Hybrid Transformer and CNN Framework" — **ACR 2026 [Accepted]**

Ali Raza. "Learn2Breath: Deformable Lung CT Registration" — **Learn2Reg / MICCAI 2026 [Ranked 3rd, Invited Oral Presentation, Strasbourg]**