ABDUL REHMAN

+92-341-7528497 | abdulrehmanghani197@gmail.com | Islamabad, Pakistan
LinkedIn | GitHub | Portfolio

COMPUTER VISION AND MACHINE LEARNING ENGINEER

Dedicated Machine Learning Engineer with 6+ years of experience specializing in designing and deploying production-grade AI systems. Expert in computer vision and edge computing, with demonstrated success reducing inference latency by 50% through optimization techniques and NVIDIA toolkit (DeepStream, Triton Server, TensorRT). Proven track record of implementing TinyML solutions that reduce costs by 60% while maintaining performance on resource-constrained hardware. Adept at integrating GenAI and NLP into enterprise applications, having developed health care and legal documents summarization platforms using LLMs, such as ChatGPT, Gemini, or Grok, with 85%+ accuracy. Eager to apply my technical expertise in scalable AI architectures as a AI engineer, where I can contribute to business AI initiative while advancing innovative AI solutions that drive business growth.

KEY COMPETENCIES

- **Languages:** Python, C/C++
- **Frameworks:** PyTorch, TensorFlow, GStreamer
- **AI & NVIDIA Tools:** CUDA, DeepStream, Triton Inference Server, TensorRT, Jetson
- **GenAI & NLP:** Large Language Models (e.g., ChatGPT, Gemini), RAG, Prompt engineering, text generation
- **Model Optimization:** Neural Architecture Search, Quantization, Pruning
- **Deployment and Automation Platforms:** Azure ML, AWS, GCP, Apache Airflow, Apache Storm

PROFESSIONAL EXPERIENCE

AIVStudios **Islamabad, Pakistan**

Computer Vision & ML Engineer **Jul 2024-Present**

- Developed an end-to-end medical record summarization platform using Python, Streamlit, and GPT-4, automating PDF extraction, data normalization, LLM-driven summarization, quality assurance review, and secure cloud storage, improving efficiency and accuracy for healthcare and legal workflows.
- Engineered an AI-powered legal document automation platform using Python, GPT-4, and Streamlit, achieving 99% reduction in legal document review time and 85% accuracy in automated decision-making for court case preparation, including predictive analysis of judicial arguments with 65% accuracy.
- Boosted video analytics speed by 50% using multi-threading and deployed fast, scalable AI services with NVIDIA DeepStream and Triton server, reducing latency from 200 ms to 100 ms using the GStreamer with python and c++.

LVisionAI **Islamabad, Pakistan**

Machine Learning Engineer **Dec 2022-Jul 2024**

- Established TinyML pipelines on low-cost devices (Jetson Nano, Raspberry Pi, OpenMV, etc.) for real-time video analytics, reducing cost and energy consumption by over 60% compared to CPU/GPU usage.
- Applied Neural Architecture Search (NAS) to search best model for target devices to reduce memory usage by 30%, improve compute efficiency by 40% and accelerate inference speed by 2× for top performance on resource-constrained hardware.
- Built pipelines on Azure ML and Google Cloud, for advertising agency, analyze the product trends and improve ad campaign accuracy by 25% and decision-making efficiency by 35%.

DLision **Islamabad, Pakistan**

Software Developer **Mar 2021-Nov 2022**

- Trained and evaluated object detection models using NVIDIA DALI, Deepstream SDK, and CUDA Toolkit, achieving a 30% reduction in training time and improving inference speed by 40%. Managed workflows with Apache Airflow, automating 100% of task scheduling for streamlined deployment.
- Implemented real-time video surveillance using Triton Server and DeepStream, offloading inference to the server and reducing client-side cost by 50%.
- Developed demand forecasting models (ARIMA, LSTM, Prophet) for a retail business using historical sales data, improving inventory planning accuracy by 35%.

DLision **Islamabad, Pakistan**

Trainee Developer **Jan 2020-Feb 2021**

- Optimized deep learning models with OpenVINO & TFLite, increasing inference speed by 40% on resource-constrained platforms.
- Contributed to dataset preparation with labelme, Roboflow, run Pytorch models for image classification and object detection (e.g., YOLO, Detr, Mobile Net), and optimizing deployment models.

Interloop Limited **Faisalabad, Pakistan**

Auditor **Jul 2019-Dec 2019**

- Analyzed data from industrial machines, leading to a 15% increase in operational efficiency.
- Generated a web visitor tracking tool using HTML5, CSS, JavaScript, and PHP, helping boost user engagement by 25%.

PROJECTS

- Developed medical records summarization platform (GPT-4 + Streamlit + FASTAPI + REACT) automating PDF extraction, reducing manual review time by 95%.
- Built sports analytics app with YOLO/DETR for player detection and ground segmentation; deployed retail face tracking & counting with DeepStream + Triton Server.
- Achieved 4× faster inference on Jetson using TensorRT for remote object detection; cut latency 35% via remote inference from ESP32 to Triton Server.
- Executed and optimized TinyML computer vision models on Jetson, Raspberry Pi, and Intel Neural Compute Stick 2, using Triton Server and OpenVINO to deliver high FPS on low-power devices.

EDUCATION

M.S. Artificial Intelligence **Nicosia, Cyprus**

University of Cyprus **Sep 2025 - cont.**

- Rigorous program focused on machine learning, natural language processing, and ethical AI development.

B.S. Computer Engineering **Taxila, Pakistan**

HITEC University **Sep 2014 - Jul 2018**

- **Final Year Project:** Vehicle detection & lane tracking for autonomous driving using CNN (Keras + TensorFlow-GPU)

CERTIFICATIONS

- Intel Edge AI for IoT Developer — *Udacity, 2021*
- Intro to Artificial Intelligence — *Udacity, 2018*
- Introduction to Computer Vision — *Georgia Tech via Udacity, 2017*