ABDUL REHMAN

+357-95-949699 | abdulrehmanghani197@gmail.com | Nicosia, Cyprus

LinkedIn | GitHub | Portfolio

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING ENGINEER

AI & ML Engineer with 5+ years of experience specializing in designing and deploying production-grade AI systems. Skilled in Python, PyTorch, and TensorFlow, with expertise in computer vision, NLP, and edge AI optimization. Proven success implementing cloud-based AI pipelines using AWS and NVIDIA DeepStream/TensorRT, improving model performance and deployment efficiency. Experienced in integrating LLMs, API-driven automation, and MLOps workflows to support digital transformation across healthcare, legal, and industrial domains. Eager to apply my expertise in scalable AI architectures to develop and deploy industry-ready AI solutions, contributing to digital transformation initiatives while advancing innovative, real-world AI applications.

KEY SKILLS

- **Languages:** Python, C/C++, JavaScript, SQL, Bash
- **Frameworks & libraries:** PyTorch, TensorFlow, FastAPI, Streamlit, React, OpenCV, GStreamer
- **AI & NVIDIA Tools:** CUDA, DeepStream SDK, Triton Inference Server, TensorRT, Dali, Jetson
- **AI & Computer Vision:** YOLO, DETR, HRNet, FCHarDNet, DeepSORT, Adaface, ONet, OpenVINO, MediaPipe
- **GenAI & NLP:** GPT, Google Gemini, Grok AI, RAG , Prompt Engineering, Text generation , Document Understanding, LLM Integration, LangChain, LlamaIndex, vector databases
- **Model Optimization and compression:** Quantization (FP32 → FP16/INT8), Pruning, Knowledge Distillation, Model Footprint Reduction, Neural Architecture Search (NAS), TinyML Deployment (OpenMV, NCS2, Raspberry Pi)
- **MLOps, Cloud & Deployment:** AWS (Lambda, S3, RDS, Textract, SES, SQS), Azure ML, GCP AI Platform, Apache Airflow, Docker, Triton Server, RESTful API Development
- **Data Engineering & Processing:** Pandas, NumPy, SpaCy, FuzzyWuzzy, PyMuPDF (fitz), PDFKit, Peewee ORM, Boto3, Requests
- **Security & Compliance:** HIPAA-Compliant Data Handling, JWT Authentication, IAM Configuration, Secure File Storage
- **Tools & Systems:** Git, Linux, VS Code, Jupyter, Anaconda, Docker Compose, Performance Profiling & Logging

PROFESSIONAL EXPERIENCE

LvisionAI (Private limited) **Remote**

AI & ML Engineer (Part- time) **Sep 2024 - Continue**

- Building video analytics pipelines for object detection, recognition, and tracking for counting and surveillance applications with real time RTSP streams processing.
- Working on accelerating and bench-marking deep models using Nvidia toolchain, automated deployment via Docker and PyTorch/TensorFlow backends.

AIVStudios **Islamabad, Pakistan**

Computer Vision & ML Engineer **Jul 2024 - Aug 2025**

- Developed a cloud-based AI agent for medical record summarization using Python, react, GPT and Gemini. Implemented secure, serverless workflows via AWS Lambda, S3, RDS, Textract, SES, and SQS, ensuring scalable and HIPAA-compliant data handling. Built RESTful APIs using FastAPI with Peewee ORM and PostgreSQL for backend operations. Enabled seamless automation of medical and legal workflows through Stripe, QuickBooks, SmartAdvocate, and Clio API integrations. (https://acrodocz.com/)
- Developed a cloud-based AI agent for legal assistance using Python, GPT-4, and Streamlit to analyze legal documents, predict potential questions judges may ask, and extract insights from recorded judicial videos. The web service assists lawyers in case preparation by providing automated summaries, predictive analysis, and actionable insights reducing review time by 99% and improving decision-making accuracy.
- Developed a real-time video analytics dashboard for a Saudi-Arabian coffee chain using YOLOv8, NVIDIA DeepStream, Triton Server AND Docker. Optimized GStreamer pipelines and multi-GPU inference to cut latency from 200 ms to 90 ms (≈55% faster), enabling real-time queue tracking and cup counting with high accuracy.

LVisionAI (Private limited) **Islamabad, Pakistan**

AI & ML Engineer **Dec 2022 - Jul 2024**

- Optimized deep learning models for edge deployment on low-power devices (Jetson Nano, Raspberry Pi, Neural Compute Stick) by reducing model footprint and memory usage, achieving up to 90% lower energy consumption compared to CPU/GPU systems for real-time video analytics at remote sites.
- Implemented remote video surveillance using Triton Inference Server and DeepStream, streaming compressed video from low-power edge devices (ESP32, OpenMV) for server-side person and vehicle detection, reducing client-side compute cost by 50% and enabling real-time monitoring in remote locations.
- Applied Neural Architecture Search (NAS) to optimize model design, improving memory efficiency, computational performance, and inference speed for high-performance AI deployments.

DLision (Private limited) **Islamabad, Pakistan**

Machine Learning Engineer **Mar 2021-Nov 2022**

- Developed a multi-camera football analytics platform using DETR for object detection, HRNet for pose estimation, and FCHarDNet for field segmentation delivering real-time player, referee, and ball tracking across 12 synchronized cameras.
- Optimized end-to-end AI workflows with Apache Airflow, NVIDIA DeepStream, and Triton Inference Server on A100 GPUs, achieving high-throughput, low-latency inference and enabling real-time tactical insights for coaches and analysts.

Python Developer **Jan 2020-Feb 2021**

- Optimized deep learning models with OpenVINO & TFLite, increasing inference speed by 40% on resource-constrained platforms.
- Developed an application for car wash monitoring system using YOLO and Python to detect and count incoming and outgoing vehicles, track personnel activity in washing and vacuum zones, and automate occupancy analytics for operational efficiency.

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%.

EDUCATION

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

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

- Advanced studies in AI, machine learning, and computational intelligence with focus on cutting-edge research.

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*

- Learn how to convert existing models to Intermediate representation to run on low cost devices by using Intel OpenVINO toolkit

Intro to Artificial Intelligence — *Udacity, 2018*

- Covered AI fundamentals, including machine learning, probabilistic reasoning, robotics, computer vision, and natural language processing.

Introduction to Computer Vision — *Georgia Tech via Udacity, 2017*

- An introduction to computer vision including fundamentals, methods for application of image processing and machine learning classification.

LANGUAGES

- ENGLISH (CONVERSATIONAL)
- URDU (FLUENT)

VOLUNTEER WORK & OPEN SOURCE CONTRIBUTIONS

- Contributor to Open-Source AI Projects – Developed and optimized AI models for real-world applications on GitHub.
- Mentor for AI & Deep Learning Enthusiasts – Guided aspiring ML engineers in model deployment and NVIDIA SDKs.