

Muhammad Allah Rakha

Lead AI | Senior AI/ML Engineer | NVIDIA NGC-SDKs R&D | Generative AI Engineer

Location: Lahore, Pakistan • Email: 4444stark@gmail.com • Phone: +923490175636 • WeChatID: AAAASTARK

LinkedIn: [linkedin.com/in/a-a-a-a-stark-69696617b](https://www.linkedin.com/in/a-a-a-a-stark-69696617b) • GitHub: github.com/aaaastark

ORCID: <https://orcid.org/0009-0008-3067-4715>

PROFESSIONAL SUMMARY

Lead AI Engineer, Senior AI/ML Engineer, and Researcher with 3+ years of experience developing scalable solutions across Computer Vision, Natural Language Processing, Deep Learning, and Generative AI. NVIDIA Startup Member, Oracle-OCI AI certified, and AMS SatMOC (NOAA/CIRA) certified in AI Applications Using Environmental Satellite Remote Sensing Data, with a BSCS from FAST-NUCES University, Pakistan (2023). Published researcher in IEEE JSTAR Transactions, AIP Conference Proceedings, and Taylor & Francis Group. Proven track record of leading cross-functional teams, owning AI roadmaps end-to-end from quarterly planning through production release, mentoring junior and mid-level engineers, and reporting delivery, technical decisions, and risk directly to the CEO and CTO. Experienced in deploying high-performance GenAI systems on GPU clusters and cloud production environments, and in driving measurable business impact through innovative solutions.

DOMAIN FIELD

- **Industry Verticals Served:** Real Estate and Property Technology, Media and Entertainment Localization, Environmental and Satellite Remote Sensing, Industrial Manufacturing and Quality Assurance, Enterprise Document and Tendering Automation, IoT and Connected Security.
Agentic AI and Conversational Systems: Multi-agent orchestration, autonomous agent design, tool use and multi-step reasoning, LLM-powered assistants, AI and voice search systems, natural language to structured query translation. Applied in production for a Dubai real estate marketplace and in autonomous research agents.
- **Generative Media and Synthetic Humans:** Enterprise video and audio dubbing platforms, neural voice cloning, lip-sync technology, 3D human avatars, virtual humans, motion capture, facial animation, photorealistic rendering. Built at scale for multilingual content localization.
- **Geospatial and Remote Sensing Intelligence:** Satellite imagery analysis, hyperspectral image processing including classification, denoising, and pansharpening, location intelligence, geospatial analytics for urban and property applications. Grounded in published research and applied SaaS delivery.
- **Computer Vision:** Object detection, semantic segmentation, instance segmentation, image classification, anomaly detection, industrial quality inspection, and defect detection. Spans peer-reviewed research through deployed production systems.
- **Predictive Analytics and Edge Systems:** Time-series forecasting, anomaly detection, Edge AI deployment, IoT security systems, real-time inference on resource-constrained hardware.

PROFESSIONAL EXPERIENCE

XOSOFT TECHNOLOGIES (PROPQA AI, UAE)

Lead AI and Senior AI & ML Engineer

PropQA (XoSoft client), a Dubai-based AI property portal and real estate marketplace offering AI-powered property search, agentic advisory, and community-level market intelligence across sales, rentals, and off-plan inventory, competing with Bayut and Property Finder.

Lahore, Pakistan - One-Site

(Full Time) March 2026 - Present

- Lead the AI and machine learning function for a Dubai property technology company competing with Bayut, Property Finder, and Zillow class portals, owning the AI roadmap from quarterly planning through production release and reporting delivery status, technical decisions, and risk directly to the CEO and CTO.
- Manage a team of 4 engineers end-to-end: scoping and assigning workstreams, reviewing code and technical designs before release, mentoring junior and mid-level engineers one-to-one, and running technical interviews and headcount planning as the team grows.

- Run sprint planning, stand-ups, grooming, and retrospectives with the Product Manager under Scrum, balancing workload across parallel workstreams and surfacing blockers early to protect committed release dates.
- Partner with product, engineering, data, design, and marketing to turn business requirements into prioritized specifications, and present new features and product proposals to the executive team with business cases covering revenue, advertising, and listing monetization, cost, and expected return.
- Led the design and delivery of an advanced agentic AI assistant built with LangChain, LangGraph, Deep Agents, and Claude language models, serving investor, buyer, and tenant personas' conversations with answers grounded in RTA data, Dubai Land Department records, developer and agency inventory, transaction history, live listings, market and off plan intelligence, and location intelligence, each backed by a traceable source.
- Built an AI-powered natural language and voice search that turns free-form or spoken queries into structured filters across rent and buy intent, community, beds, baths, price, handover year, area size, amenities, residential and commercial, off-plan and ready, and furnishing, plus semantic ranking and similar property recommendations.
- Delivered PropQA Area Insights, automatically generating community and sub-community pages across Downtown Dubai, Dubai Marina, Palm Jumeirah, Business Bay, and Jumeirah Beach Residence (JBR) areas from transaction trends, price and rental yield benchmarks, supply pipeline, and liveability signals, supported by ingestion pipelines that deduplicate, resolve, and validate data from government, developer, and agency sources before it reaches production.
- Architected the production platform on AWS using EC2, PostgreSQL on Amazon RDS, Qdrant, and Redis Stack, optimizing for real-time performance through caching, batching, streaming, and model routing, with drift and performance monitoring, rollback procedures, evaluation standards, and documented runbooks now used as the team's baseline practice.

QORDEN AI LTD

Senior AI&ML Engineer

Dubai International Financial Centre (DIFC), UAE - Remote

(Full Time) August 2024 - January 2026

Qorden AI is an AI company delivering multilingual speech analytics, real-time translation, and video dubbing through its Dubbix, QDUB, and QSAP platforms, helping global enterprises break language barriers across voice and video communication.

- Leading AI/ML architecture and development for Dubbix, an enterprise-grade AI-powered video dubbing and creative studio platform that enables content creators and businesses to localize and generate multilingual video content at scale across 10+ languages.
- Architected and deployed end-to-end AI/ML pipelines for Dubbix's flagship products, including Video Dubbing, Speech-to-Speech translation, Text-to-Speech synthesis, AI Avatars, and Video Generation capabilities.
- Engineered production-ready voice cloning and lip-sync systems using advanced deep learning models, achieving studio-quality output with sub-second latency for real-time dubbing workflows.
- Developed multilingual speech synthesis engines supporting 10+ languages with natural prosody, emotion preservation, and accent adaptation capabilities.
- Optimized inference pipelines using NVIDIA TensorRT and Triton Inference Server, reducing video dubbing processing time by 60% while handling concurrent multi-language requests on distributed GPU clusters.
- Implemented CUDA-optimized kernels and leveraged the NVIDIA NeMo framework for efficient large language model fine-tuning and deployment, enabling scalable voice transformation workflows.
- Designed and deployed cloud-native ML infrastructure on AWS with auto-scaling capabilities, supporting enterprise clients with 99.9% uptime SLA.
- Collaborate with cross-functional teams, including product managers, UX designers, and backend engineers, to deliver AI-powered features aligned with business objectives and user needs.
- Implemented a comprehensive MLOps pipeline with CI/CD automation, model versioning, A/B testing framework, and real-time monitoring using MLflow.
- Mentor junior ML engineers on best practices for model development, NVIDIA AI stack utilization, and production deployment strategies.
- Drive technical decision-making for model selection, architecture design, and infrastructure optimization to ensure optimal performance-cost balance.

DATAFY ASSOCIATES

Generative AI Engineer

Melbourne, Australia - Remote

(Full Time) August 2023 - July 2024

- Developed AI-powered SaaS/PaaS solutions for Sustainable Energy Infrastructure and GeoSpatial Intelligence, leading a team of 2 engineers.
- Built production Web applications using GPT-4, Claude, and custom fine-tuned models, delivering \$500K in cost savings for enterprise clients.
- Accelerated NLP models at Lambda Cloud by 65% through NVIDIA TensorRT optimization and Triton Inference Server deployment across A100 GPUs.
- Implemented MLOps workflows with Docker, Kubernetes, and CI/CD automation, reducing model deployment time from weeks to hours.
- Developed real-time analytics dashboards using Django and FastAPI, improving data visibility by 99% and enabling cross-functional collaboration.
- Applied statistical modeling and predictive analytics to drive 25x revenue growth through data-driven decision making.

TECHNICAL SKILLS

Programming Languages: Python, C++,

NVIDIA Technologies: TensorRT, TensorRT-LLM, NeMo, Riva, RAPIDS, CUDA, DeepStream, Triton Inference Server, DALI, NIM

ML&DL: PyTorch, ONNX, JAX, TensorFlow, Keras, Scikit-Learn, Hugging Face Transformers, LangChain, LangGraph, DeepAgents, OpenAI API, Claude AI API, RAG Systems, Fine-tuning LLMs, Model Optimization

Computer Vision: OpenCV, Open3D, YOLO, ResNet, Vision Transformers, Stable Diffusion GAN, UNet

Natural Language Processing: Transformer Models, BERT, GPT, T5, Sequence-to-Sequence Models, Speech Recognition, Text-to-Speech, Multilingual NLP

MLOps & Infrastructure: Docker, Kubernetes, MLflow, Weights & Biases, CI/CD Pipelines, Model Monitoring

Cloud Platforms: AWS (SageMaker, EC2, S3), Google Cloud Platform (Vertex AI, Compute Engine, Maps), Azure (AI Foundry Factory), Oracle Cloud Infrastructure

Distributed Computing: Ray, Hadoop, Multi-GPU Training, Model Parallelism, Data Parallelism

Data Engineering: PostgreSQL, MongoDB, Snowflake, ETL Pipelines, Databricks

Development Tools: Git, GitHub, FastAPI, Django, REST APIs, Redis Stack

Visualization & Analytics: Matplotlib, Plotly, Tableau, Streamlit, Gradio, Power BI

EDUCATION

UNIVERSITY OF MANAGEMENT AND TECHNOLOGY (UMT)

Master of Artificial Intelligence (MS AI)

Lahore, Pakistan - One-Site

(Weekend) March 2026 - In Progress

- Advanced coursework in deep learning, computer vision, natural language processing, and large language models, pursued alongside full-time engineering practice and applied directly to production work.

FAST-NUCES UNIVERSITY

Bachelor of Science in Computer Science (BSCS)

Islamabad, Pakistan

(Full Time) August 2019 - December 2023

- Completed capstone project on Hyperspectral Satellite Image Denoising using deep neural networks
- Published research papers in IEEE and contributed to a CRC Press book chapter on Cyber Security
- Participated in AI/ML research and development initiatives

RESEARCH EXPERIENCE

CHONGQING UNIVERSITY

Researching in Artificial Intelligence and Hyperspectral Satellite Image

Chongqing, China - Remote

(Per Diem) July 2022 - December 2023

- Specialized in hyperspectral satellite image processing, focusing on classification, denoising, and pan-sharpening using deep learning architectures.
- Published peer-reviewed research in the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

- Designed novel CNN and Transformer-based architectures for high-dimensional image analysis, improving performance by 80%
- Collaborated with international research teams on computer vision and remote sensing projects.

PUBLICATIONS

1. **Ashraf, M., Chen, L., Zhou, X., & Rakha, M. A.** (2024). A joint architecture of mixed-attention transformer and octave module for hyperspectral image denoising. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.
2. **Rakha, M. A., Khan, I. U., Ouaisa, M., Ouaisa, M., & Ayub, M. Y.** (2024). Hybrid model for IoT-enabled intelligent towns using the MQTT-IoT-IDS2020 dataset. In Cyber Security for Next-Generation Computing Technologies (pp. 159-176). CRC Press.
3. **Abbas, Z., Qahar, A., Shahzad, M. Y. A., & Rakha, M. A.** (2024). Security and intelligent management. In Cyber Security for Next-Generation Computing Technologies (p. 16). CRC Press.
4. **Shahzad, A. Q., Rakha, M. A., & Komal, I.** (n.d.). Ad hoc aided cellular networks and LPWAN applications: A review. In Low-Power Wide Area Network for Large Scale Internet of Things (pp. 71-85). CRC Press.
5. **Shahzad, A. Q., Nishat, F., Rakha, M. A., & Ayub, M. Y.** (n.d.). LPWAN technology for demand side management. In Low-Power Wide Area Network for Large Scale Internet of Things (pp. 86-99). CRC Press.

CERTIFICATIONS

- AMS SatMOC (NOAA/CIRA) in AI Applications Using Environmental Satellite Remote Sensing Data
- Oracle Cloud Infrastructure 2023 AI Certified Foundations Associate
- Oracle Cloud OCI Services Certified
- LeNet Enabled Intrusion Detection System for IoT Networks (2023)
- Machine Learning with Python - FreeCodeCamp (2022)

CONFERENCES & PROFESSIONAL DEVELOPMENT

- NVIDIA GTC Conference (2023, 2024, 2025)
- MLOps Conference 2023
- Dell Technologies: Accelerate Transformation with PowerEdge
- Weights & Biases: MLOps Fully Connected Conference 2023

ADDITIONAL INFORMATION

Languages: English (Fluent), Urdu (Native), Dutch (Conversational)