Muhammad Muneeb Arshad

Toronto, Ontario | LinkedIn | +1 437 575 6745 | mnbrshd.github.io | mnbrshd@gmail.com | GitHub

Machine Learning Engineer with 6+ years of experience building scalable AI systems, classic ML pipelines, advanced agentic AI workflows, and production-grade document intelligence platforms. Hands-on across the full ML lifecycle — from EDA, feature engineering, and model development to distributed deployment, optimization, and cloud-native orchestration.

Specialized in LLMs, RAG, agent-based systems, NLP for long documents, vector databases, microservices, and event-driven AI architectures.

Skills

ML Engineering: Classic ML, feature engineering, hyperparameter optimization, statistical modeling, XGBoost, SVM, Random Forest

LLMs & Agentic AI: LangChain, LangGraph, OpenAI Agents SDK, Gemini, RAG, memory systems, evaluators, tool-use agents

NLP & Document Intelligence: Text segmentation, summarization, semantic search, entity extraction, contextual reasoning

Vector Databases: ChromaDB, Milvus, Pinecone, Weaviate

Backend / MLOps: FastAPI, Django, Docker, microservices, CI/CD, Kubernetes, Celery, SQLAlchemy

Cloud & Deployment: AWS, GCP, Cloud Run, Lambda, Pub/Sub, Cloud Functions, serverless pipelines

Deep Learning: PyTorch, TensorRT, ONNX, YOLOv8, TransUNet, SAM / MobileSAM

Systems: Event-driven architecture, message queues (RabbitMQ/PubSub), distributed inference, monitoring (Prometheus/Grafana)

Experience

Fullstack AI Engineer (Full Time) **AIXFF Tech Inc,** Toronto, Ontario **Sep 2025 - Present**

• Built a **next-generation multi-collection RAG system** using FastAPI, LangGraph, Gemini, and ChromaDB, improving retrieval precision and throughput for large-scale document reasoning.

• Developed **agentic AI workflows** enabling multi-step reasoning, memory retrieval, evaluators, and tool-use orchestration.

• Deployed the entire platform using **AWS (Bedrock, Lambda, S3) + Terraform**, ensuring scalable, cloud-native, high-availability infrastructure.

• Designed intelligent AI-enabled microservices powering R&D workflows for fragrance formulation.

• Led cross-functional collaboration with engineering & product teams to translate ambiguous requirements into production AI capabilities.

Senior AI Engineer **siParadigm Diagnostic Informatics,** Pakistan **Jan 2022 - Apr 2025**

• Built **document intelligence systems** using spaCy NER, LLM-based extraction, and semantic enrichment pipelines for clinical data.

• Designed **agentic multi-agent architectures** using LangGraph + OpenAI Agents SDK, including evaluator agents, memory routing, fallback logic, and tool-use orchestration.

• Developed a full **RAG pipeline** using embeddings + Milvus/ChromaDB for semantic search, long-document processing, summarization, and contextual reasoning.

• Conducted extensive **EDA and preprocessing** on noisy EHR/PDF-extracted data to ensure consistency in downstream ML/NLP workflows.

• Implemented advanced **chunking strategies for long, mixed-format clinical PDFs**, including multi-column layout handling and table structure extraction.

• Integrated vector databases into production inference stacks to ensure high-performance retrieval.

MLOps Engineer **siParadigm Diagnostic Informatics,** Pakistan **Jan 2022 - Apr 2025**

• Architected **event-driven ML infrastructure** using microservices, RabbitMQ, Pub/Sub, and serverless compute for high-throughput document ingestion and inference.

• Deployed real-time **ASR + TTS pipelines** on AWS Lambda & S3 for clinical workflow automation.

• Delivered ML-driven triage systems improving diagnostic decision accuracy by **30%**.

• Built chromosome workflow automation using **MobileSAM**, reducing review time by **50%**.

• Optimized YOLOv8 and TransUNet inference using **16-bit ONNX + C++**, reducing latency by **40%**.

• Owned ML lifecycle end-to-end: versioning, monitoring, evaluation loops, CI/CD for models, and long-term maintenance.

AI Developer **Jul 2019 - Jan 2022**

- Developed and deployed a **YOLOv5-based WBC detection system** using Python, Django, and Docker.
- Created a PyQt5 clinical decision-support tool with iteration loops informed by physician feedback.
- Built **classic ML pipelines** with feature engineering, hyperparameter tuning (XGBoost, SVM, RF), and comparative model evaluation.
- Collaborated with clinicians and engineering teams to operationalize AI outputs in diagnostic workflows.

Education

- **Robotics and AI (Major)** - Master of Science (2018-2022)
  National University of Sciences and Technology, Islamabad, Pakistan

- **Mechanical Engineering (Major)** - Bachelor of Engineering (2014-2018)
  National University of Sciences and Technology, Islamabad, Pakistan

Publications

Stain Normalization of Hematology Slides using Neural Color Transfer

arxiv.Electrical Engineering and Systems Science.Image and Video Processing · Sep 10, 2024