Mohsin Khan

Lahore, Pakistan

📞 +92-314-941-5367   ✉️ mohsinkn007@gmail.com   in linkedin.com/in/mohsin-khan   🌐 mohsinml.vercel.app

Education

National University of Computer and Emerging Sciences   **Aug 2020 – May 2025**
Bachelor of Science in Computer Science

Technical Skills

Languages: Python, C++, SQL, Bash
Frameworks & Libraries: FastAPI, LangChain, TensorFlow, PyTorch, OpenCV, Scikit-learn, XGBoost, Pandas, NumPy, Wandb, FAISS, Vector Databases
Databases: PostgreSQL, Supabase, MongoDB
Tools & Platforms: Git, Docker, Kubernetes, AWS, Hugging Face, Ollama

Experience

AxcelerateAI   **Lahore**
Machine Learning Engineer   **Jul 2025 – Dec 2025**

- Fine-tuned BERT for Spanish real estate ad classification, achieving 95% precision on a diverse dataset. Containerized and deployed the model via FastAPI on AWS EC2 for scalable batch processing.
- Built a mobile-integrated computer vision solution for real-time golf ball detection and tracking in cluttered outdoor environments.
- Trained and optimized a YOLO-based model on a custom dataset with diverse terrains, lighting conditions, and occlusions to support lightweight iOS deployment.
- Developed a computer vision–based restaurant monitoring system to analyze customer activity and service patterns from video streams.
- Designed a YOLO-powered detection and tracking pipeline using ByteTrack and DeepSORT to identify entry events, serving actions, and table interactions.
- Built and Dockerized a document OCR system using PaddleOCR and FastAPI to extract structured information from images.
- Applied image preprocessing (contrast enhancement, denoising) and integrated Supabase (PostgreSQL) for persistent storage.

Projects

Proactive Autoscaling of Cloud Apps Using ML (FYP) | *TensorFlow, Kubernetes, Docker, React*   **Sep 2024**

- Built a forecasting model to predict container workload trends 5 minutes in advance and drive automated scaling, reducing resource waste.
- Developed a custom Kubernetes autoscaler for dynamic replica adjustment and validated scaling policies under varying workloads using Minikube.
- Integrated a FastAPI backend with a React dashboard for real-time visualization of pod metrics and autoscaling performance.

PSX Stock Price Forecasting using Deep Learning | *Python, TensorFlow, yfinance*   **Jun 2025**

- Developed an LSTM-based time-series model to predict next-day PSX stock prices using historical OHLC and volume data to outperform ARIMA
- Built a preprocessing and sequence windowing pipeline to capture temporal dependencies and evaluated model performance on real market trends.

ClinQA: Fine-tuning LLM using QLoRA | *Gradio, Transformers, PEFT, Gemma2*   **Feb 2025**

- Fine-tuned Google's Gemma model with QLoRA on medical QA dataset for clinical question answering.
- Implemented LoRA adapters to reduce memory usage and accelerate training.
- Built a Gradio interface for real-time evaluation of fine-tuned responses.

Vehicle Orientation Classifier | *Python, TensorFlow, scikit-learn*   **Oct 2024**

- Built CNN-based image classifier for vehicle orientation detection; achieved 95% accuracy.
- Implemented hybrid model using VGG16+ SVM, improving precision to 96%.
- Curated real-world dataset from PakWheels for realistic classification performance.

Certificates

Machine Learning Specialization (Coursera)
Deep Learning Specialization (Coursera)
Get Started with Databricks for Machine Learning (Databricks)