

MUHAMMAD HISAN USMAN

hisan.usman20@gmail.com
linkedin.com/in/hisan-usman | github.com/hisan-usman

PROFESSIONAL SUMMARY

AI Engineer with hands-on experience building production-grade AI systems across healthcare and enterprise domains. Proven track record in designing multi-agent architectures, HIPAA-compliant voice agents, computer vision pipelines, and end-to-end ML workflows. Proficient in Python, PyTorch, TensorFlow, LangChain, and cloud-based ML infrastructure.

EDUCATION

BS in Artificial Intelligence — FAST-NUCES, Islamabad

2021 – 2025

- Dean's List of Honours — Fall 2023

PROFESSIONAL EXPERIENCE

AI Engineer | Remotebase / Ember AI

Feb 2026 – Present

Pakistan

- Architecting a multi-agent AI system that transforms client requirements into production-ready implementation specifications
- Building a feature-based workflow engine with intelligent document retrieval (RAG), multi-agent deliberation, and real-time streaming
- Optimized AI performance by 10x through strategic prompt engineering and context management, reducing hallucinations

Junior AI Engineer | CareCloud Inc.

Jun 2025 – Feb 2026

Islamabad, Pakistan

- Led a team of 10+ to build HIPAA-compliant voice agents (StratusAI) automating front-desk operations for healthcare providers
- Engineered a conversational pipeline using VAD, turn-detection, ASR/STT, LLMs, and TTS for real-time intent recognition
- Integrated voice agents with EHRs (CareCloud Health & TalkEHR), enabling providers to view AI-created appointments and tasks
- Developed an end-to-end ML pipeline to parse EDI 837 healthcare claims into 40+ fields, achieving 95%+ prediction accuracy
- Specialized in NLP competencies including generative modeling, speech analytics, multilingual support, and healthcare compliance

IT Intern | CareCloud Inc.

Mar 2025 – May 2025

Islamabad, Pakistan

- Built a voice agent analyzing call transcripts using KPI-based metrics to generate ideal conversations and benchmarks
- Produced ideal conversations to coach employees, driving a 20% boost in team efficiency through targeted performance improvement

Machine Learning Intern | Teradata

Jun 2024 – Aug 2024

Islamabad, Pakistan

- Built large-scale financial analytics deep learning models using Teradata Vantage (teradataml) and scikit-learn, enabling in-database model training for high performance
- Developed full data science pipelines from data preprocessing and feature engineering through model evaluation, ensuring robust and reproducible results
- Optimized handling of massive datasets via Teradata's in-database analytics to deliver scalable and efficient model development

AI Intern | Systems Limited

Jul 2023 – Sep 2023

Islamabad, Pakistan

- Automated the data-science pipeline, eliminating the need for ML engineers to write boilerplate ML code
- Built a UI with interactive graphs and evaluation features enabling users to train and save models at the click of a few buttons
- Boosted engineering efficiency: reduced time spent on routine tasks and freed the team to focus on feature crossing and model performance tuning

ML Intern | Rapidev

Jun 2023 – Jul 2023

Islamabad, Pakistan

- Collaborated with ML Engineers to deploy object detection models for a major Saudi project, enabling automated detection of vehicles using drone and camera imagery
- Led annotation efforts across varied angles and lighting conditions, helping improve model accuracy by 53%
- Utilized computer vision techniques to distinguish multiple vehicle types, boosting operational capability and reducing manual monitoring

SELECTED PROJECTS

CCTV Surveillance System (FYP): AI-powered surveillance using Vision Transformer (ViT) for violent activity detection with adversarial training, custom YOLOv8 for real-time object detection, CLIP for scene understanding, and LangChain for automated report generation. Optimized live-stream performance via multi-threaded processing.

Drug Fact-Checking Multi-Agent System: Built a multi-agent system using LLaMa 3.1, PubMedBERT, and ChromaDB to verify drug-related queries against actual medical literature before responding. Designed to eliminate LLM hallucinations in healthcare contexts with a FastAPI backend.

PK Constitution Knowledge Graph: Built a queryable knowledge graph over Pakistan's Constitution using Qwen, Ollama, and KùzuDB, enabling structured natural language querying over legal documents with graph-based retrieval.

Multi-Agent RAG Travel Planner: Multi-agent travel planning system using RAG and vector embeddings to generate cost-optimized vacation itineraries with real-time data retrieval.

AI Agent for University Portals: Automated university portal workflows for instructors using OCR, NLP, and LLM-driven processing for results generation, email drafting, and portal updates.

Deep Learning for Custom Emoji Generation: Live system using GANs to translate users' facial expressions into real-time emojis via video capture and expression recognition.

Sign Language Detection & Speech Generation: Real-time sign-language-to-speech system using CNN for gesture recognition, GPT for sentence construction, and HuggingFace TTS for audio output.

Human Action Recognition with Vision Transformer: Fine-tuned ViT for action recognition on the HMDB dataset, achieving 95% accuracy across 51 action categories.

Diet Planner with AI & LLMs: K-means clustering pipeline to segment patients by health indicators, with HuggingFace LLM generating personalized diet plans.

Stock Price Prediction with LSTM: LSTM-based system for real-time Google stock prediction, trained on historical data and deployed via live API.

Weather Forecasting with MLOps: End-to-end MLOps pipeline using DVC, Apache Airflow, and MLFlow for automated, scalable weather forecasting.

TECHNICAL SKILLS

Languages & Frameworks: Python, SQL, HTML/CSS, React, PyTorch, TensorFlow, scikit-learn, LangChain, LangGraph, Hugging Face Transformers, OpenCV, FastAPI

AI/ML Specializations: Agentic AI, Multi-Agent Systems, RAG, Knowledge Graphs, Generative AI, LLMs, Computer Vision, NLP, Deep Learning, MLOps, Data Science

Tools & Platforms: Git, Apache Airflow, DVC, MLFlow, n8n, Model Context Protocol (MCP), ChromaDB, KùzuDB, NumPy, Pandas, Matplotlib

CERTIFICATIONS

- Building AI Agents with RAG & LangChain — IBM (edX)
- Advanced AI: Transformers for Computer Vision — LinkedIn Learning
- Computer Vision for Data Scientists — LinkedIn Learning
- Machine Learning — Stanford Online (Coursera)