Dawood Sarfraz

Artificial Intelligence Engineer

☎ +923061757838 ✉ dawoodsarfraz.cs@gmail.com in Dawood Sarfraz 🌐 Website 📍 Lahore, Pakistan

Education

FAST National University of Computer and Emerging Sciences
Bachelor in Computer Science *Sep 2020 - Aug 2024*

Experience

OrbytLabs **June 2024 – Present**
Machine Learning Engineer *Remote*

- **Developed a vehicle price prediction system**, applying **data validation, feature engineering, and model optimization**. Iteratively improved the model to achieve **98% prediction accuracy**, enabling accurate and reliable price estimations.
- **Built object detection and tracking systems** using **YOLOv10/YOLOv11**, integrating tracking algorithms for real-time visual understanding in **smart security solutions**. **Applied post-training quantization (PTQ)and Quantization-Aware Training (QAT)** to optimize model performance and ensure **efficient, real-time inference on edge devices**.
- Implemented **intrusion detection systems** by combining motion-based region analysis and AI-driven object localization to automatically detect and flag unauthorized entries or restricted zone violations in real time.
- **Implemented OCR pipelines** using **Tesseract** and **EasyOCR** to extract and structure information from **scanned documents and ID cards**, enabling automated data extraction and document digitization.
- **Created and deployed LLM-based chatbots and document understanding systems** using **OpenAI, Claude, and Gemini APIs** with **RAG pipelines** to automate customer support and business operations. **Fine-tuned large language models (LLMs)** for domain-specific use cases, enhancing contextual understanding, accuracy, and efficiency.

Deutics Global **Jan 2025 – July 2025**
Associate Machine Learning Engineer *On-site*

- Developed and optimized a real-time video analytics system by converting RTSP streams to WebRTC for efficient live video processing.
- Implemented advanced tracking algorithms for real-time object tracking across frames, enabling movement monitoring and direction estimation.
- Implemented OCR pipelines using Tesseract, EasyOCR, and custom deep learning models to extract license plate numbers from images and video streams.
- Built and optimized algorithms for wait time estimation, queue detection, speed calculation, and traffic light violation detection to improve traffic efficiency, enforce road regulations, and optimize signal timing.
- Implemented line and zone intrusion detection to monitor restricted areas, enhance security, and enforce access regulations.

FAST NUCES **Sep 2023 – December 2024**
Machine Learning Engineer (Research Assistant) *On-site*

- Conducted research on skin cancer classification using the HAM10000 dataset, focusing on early detection and diagnosis.
- Implemented and evaluated deep neural networks (CNN, NasNet, ShuffleNet) for multi-class skin cancer detection.
- Achieved up to 93% accuracy with NasNet, outperforming CNN (92%) and ShuffleNet (87%).
- Applied data balancing (RandomOverSampler) and model optimization techniques (batch normalization, dropout, Adamax optimizer) to improve robustness and accuracy.
- Demonstrated the potential of advanced neural networks to enhance diagnostic precision and support efficient skin cancer detection solutions.

- Developed an LSTM-based deep learning model to predict future stock prices using historical market data.

- Preprocessed datasets containing features such as opening/closing prices and trading volume, and created training/testing splits for time series forecasting.
- Trained and tuned LSTM models with hyperparameters (hidden layers, neurons, learning rate) while applying regularization and dropout to prevent overfitting.
- Evaluated model performance using MSE, RMSE, and MAE, and visualized predicted vs. actual stock prices to assess accuracy.
- Demonstrated the effectiveness of LSTM networks in capturing long-term dependencies and improving stock price forecasting.

Anonymous Tree — Jun 2023 – Aug 2023

Machine Learning Engineer Intern — *Remote*

- Developed a personalized recommendation system using user interaction data (clicks, purchases, ratings) and item metadata to provide tailored product suggestions.
- Implemented content-based filtering (TF-IDF, Cosine Similarity), collaborative filtering (user/item-based, SVD, KNN), and a hybrid approach to enhance recommendation accuracy.
- Optimized models via hyperparameter tuning (Grid Search, Randomized Search) and evaluated performance using Precision, Recall, F1 Score, and MSE.
- Applied cross-validation and continuous feedback loops to ensure model generalization, prevent overfitting, and improve recommendations over time.
- Improved user engagement and product discovery by delivering relevant recommendations, leading to higher satisfaction and sales.

Projects

Skin Cancer Classification using NasNet and ShuffleNet

- Developed deep learning models (Custom CNN, NasNet, ShuffleNet) for multi-class skin cancer classification using the HAM10000 dataset. Enhanced accuracy with batch normalization, dropout, and Adamax optimizer, achieving the highest accuracy with NasNet and improved efficiency with ShuffleNet.

LlamaAssist Project Link

- LLama Assist is an AI-powered application built with LLama 3.2 3B and LLama 3.2 Vision. It allows users to generate assignments with customizable output length and difficulty, save content into text files, and interact with images via AI-driven chat. Perfect for creating and managing assignments with ease and precision.

RoboText Classifier Project Link

- Built a text classification model using RoBERTa and NLTK. Enhanced performance with dynamic masking, sentence packing, byte-level BPE vocabulary, and larger batch sizes for efficient, accurate classification of diverse text.

Duplicate Questions Pair Project Link

- Built a model to identify and detect duplicate question pairs using Random Forest, XGBoost, and Decision Tree classifiers. Achieved 90% accuracy with XGBoost classifier.

Skills

Programming Languages: Python, C++, JavaScript

Full-Stack Web Technologies: HTML, CSS, Bootstrap, React, FastAPI, Django, Flask, Streamlit, Gradio

Tools & Platforms: Git, Docker, AWS, GCP, Azure

Databases & Vector Stores: MySQL, MongoDB, PostgreSQL, Redis, FAISS, Pinecone, Weaviate

ML/DL Frameworks: PyTorch, TensorFlow, Keras, Hugging Face Transformers, YOLO

ML/DL Libraries: Scikit-learn, NumPy, Pandas, Matplotlib, Seaborn, SciPy, OpenCV, NLTK, spaCy

GenAI Libraries & Tools: LangChain, LangGraph, LlamaIndex, Crew AI

LLM Models: LLaMA, DeepSeek, Mistral, Falcon, Phi, Qwen, Granite, Gemma

Speech & Multimodal Models: Whisper

Techniques: Retrieval-Augmented Generation (RAG), Prompt Engineering, Fine-tuning, LoRA, QLoRA, Quantization, Model Distillation, PEFT