Muhammad Sameer

AI/ML Engineer | ML Systems | MSc Data Science
Berlin, Germany (Open to relocation within Germany) • **+49 176 62376462** • sameermubasher99@gmail.com
Portfolio: muhammadsameer.de • **GitHub:** github.com/mirzasameer2000
LinkedIn: linkedin.com/in/mirzasameerbaig99
Xing: xing.com/profile/Muhammad_Sameer033677
Languages: English (C1) · Urdu (Native) · German (A2)

PROFILE

I build AI systems end to end, from data pipelines and model training through to APIs and deployed services. At Della.hu I designed and shipped a production LLM orchestration layer integrating 10+ AI providers, with async task pipelines, rate limiting, and audit logging. I also built a full RAG system over biomedical literature using LlamaIndex and Falcon-7B. My dissertation applies ML and DL to large real-world atmospheric and flight datasets. I am comfortable working across the full stack of an AI system, and I genuinely enjoy translating ambiguous research needs into something that actually runs.

MASTER'S THESIS RESEARCH (ACTIVE — ARDEN UNIVERSITY BERLIN)

Contrail Formation Prediction & Climate-Aware Flight Trajectory Optimisation Using ML/DL *MSc Dissertation* *Arden University Berlin | Supervisor: Dr. Yasser Shokr | Target: Sep 2026*

- Researching whether aircraft will form contrails based on atmospheric conditions — contrails trap heat and may contribute more to global warming than aviation CO_2 itself; built ML/DL models to predict contrail formation and optimise flight paths to avoid contrail-forming zones while minimising fuel cost.
- Collected and merged two large real-world datasets independently: **OpenSky Network** flight data (36 CSV files, 15th of each month 2023–2025, 80,000–138,000 flights/day) and **ERA5 atmospheric reanalysis data** from Copernicus (36 NetCDF files, variables: temperature, humidity, wind components, cloud cover, geopotential at cruise pressure levels 200–300 hPa).
- Performed full data pipeline: raw file ingestion, merging flight records with ERA5 atmospheric readings by timestamp and pressure level, subsampling strategy to handle memory constraints, and feature engineering for contrail-relevant atmospheric conditions.
- **ML phase:** Random Forest and XGBoost for contrail formation classification and comparison.
- **DL phase:** LSTM and Deep Neural Network on the same merged tabular dataset — enabling direct ML vs DL performance comparison on identical data.

KEY ML/DATA SCIENCE PROJECTS

Flight Price Prediction — Data Analysis & ML | Arden University (Machine Learning) | *2024–25*
Python · Pandas · scikit-learn · XGBoost · EDA · Feature Engineering · Statistical Testing

- Collected and cleaned a 300K-row aviation dataset from publicly available online resources; performed full EDA including distribution analysis, correlation heatmaps, and outlier detection using seaborn and matplotlib.
- Applied ANOVA and Kruskal-Wallis hypothesis testing to identify statistically significant features before model training — airline, departure time, number of stops, route.
- Built end-to-end ML pipeline: LabelEncoder feature engineering → Random Forest (R^2 0.985, RMSE ₹2,768) outperforming XGBoost (R^2 0.966) and Linear Regression (R^2 0.904).
- Findings show that with richer flight telemetry or satellite-sourced data, model depth can increase significantly — identified as the core motivation for further aerospace ML research.

Handwritten Digit Recognition — Deep Learning | Arden University (AI & Neural Networks) | *2025*
PyTorch · MLP · Custom Preprocessing · Ablation Study

- Designed 6-layer MLP from scratch in PyTorch with a fully custom preprocessing pipeline: MaxFilter, connected component extraction, centre-of-mass shift, 28×28 canvas normalisation.
- Ran activation ablation study across ReLU, Tanh, Sigmoid — achieved 98.07% MNIST accuracy. Demonstrated the value of preprocessing quality on model performance.

MedAI — RAG Research Chatbot | *2025*

LlamaIndex · Falcon-7B · RAG · HuggingFace · Gradio

- Built full RAG system over CORD-19 biomedical dataset using LlamaIndex VectorStoreIndex, all-MiniLM-L6-v2 embeddings, and Falcon-7B-Instruct (4-bit NF4 quantisation).
- Handled real challenges of working with large unstructured scientific corpora: chunking strategy (150 words), persistent index, memory-buffered chat engine (2048 tokens).

Medical ML — Heart Disease & Diabetes Prediction | *2025*

scikit-learn · Logistic Regression · SVM · StandardScaler · PIMA Diabetes Dataset · Cleveland Heart Disease Dataset

- Built binary classifiers on two medical datasets: heart disease prediction (Logistic Regression, 82% test accuracy, confusion matrix analysis) and diabetes prediction (SVM with linear kernel on PIMA dataset, 77% test accuracy with StandardScaler preprocessing).
- Applied stratified train/test splits, data standardisation, and predictive system deployment for single-instance inference.

WORK EXPERIENCE

Django Developer — AI Systems & Backend | Della.hu (Remote, Budapest) | *Sep 2024 – Present*

Python · Django · LLM APIs · MongoDB · Celery · Redis · Docker · AWS

- Built a production orchestration system in Django integrating 10+ AI providers, with rate limiting, API key rotation, and audit-trail logging.
- Built AI Jogász, a full-stack contract management platform with LLM-driven workflows, REST APIs, Docker deployment, and a React frontend - demonstrating end-to-end ownership from backend to user-facing tools. Managing deployments at Railway.
- Designed Celery/Redis task pipelines for high-concurrency AI inference workloads, showing asynchronous processing and infrastructure thinking.

EDUCATION

MSc Data Science | Arden University Berlin | *Sep 2025 – Oct 2026*

- **Current average:** 74% — 5 Distinctions, 1 Merit (Arden University Level 7 scale)

BSc Computer Science | University of Faisalabad, Pakistan | *2018 – 2022*

- **CGPA: 3.45**

TECHNICAL SKILLS

ML / DL: PyTorch · TensorFlow · scikit-learn · XGBoost · HuggingFace · LangChain · LlamaIndex · RAG · NLP · SVM · StandardScaler · Weights & Biases

Data & EDA: Pandas · NumPy · Matplotlib · Seaborn · Bokeh · Power BI · Tableau · Statistical Testing (ANOVA, Kruskal-Wallis)

Cloud & Infra: AWS (Lambda, S3, Athena) · Docker · GCP · Snowflake · Hadoop · PySpark · Git · CI/CD

Languages: Python · SQL · R · Bash (basic) · C++ · JavaScript · React

CERTIFICATIONS

ML Specialization — Stanford / DeepLearning.AI (Dec 2024)

Deep Learning in Computer Vision — HSE University (May 2021)

IBM Data Science Professional Certificate (Nov 2025)

Django Web Framework — Meta/Coursera

Data Analysis with R Programming (Dec 2024)