

SUNDUS AIJAZ

Data Scientist & ML Engineer · MS Data Science, FU Berlin

Berlin, Germany · sundusaijaz2011@gmail.com · +49 17637665217 · [LinkedIn](#) · [GitHub](#) · [Portfolio](#)

PROFESSIONAL SUMMARY

Data Scientist and ML Engineer finishing MS Data Science at FU Berlin, with 3+ years of professional experience building end-to-end data pipelines, ML models, and cloud-based analytics solutions. Skilled in Python, PySpark, Kafka, AWS, and Azure, with hands-on experience delivering data-intensive systems from initial design through CI/CD deployment. Comfortable translating complex business problems into data science solutions and communicating results clearly to technical and non-technical stakeholders.

WORK EXPERIENCE

Working Student, ML

Dec 2024 – Present

Technische Universität Berlin · Berlin, Germany

- Architected SpkEnsembleCalculator in PyTorch, an end-to-end ML inference tool exposing uncertainty-quantified energies, forces, and stress tensors as structured outputs, enabling active-learning loops for atomistic simulation research.
- Building a generative deep learning model for novel molecular structure synthesis, taking the feature from prototype to active training component, covering the full ML lifecycle from data preparation through evaluation.
- Maintaining and contributing to the open-source SchNetPack repository as core contributor, enforcing code quality, documentation, and reproducibility across releases.
- Unified 8 molecular datasets (QM9, MD17, MD22, rMD17, MPtrj, ISO17, QM7x, Materials Project) into a single reproducible data pipeline, eliminating per-dataset integration work and enabling one-line dataset swapping.
- Shipped pytest suites across three modules (ensemble, data, atomistic-offset), enabling fully automated CI with zero manual test runs per release cycle.

Software Engineer

Oct 2021 – Oct 2023

Lyftrondata · US, Remote

- Reduced data processing time by 30% by engineering Airflow + Kafka ETL pipelines for batch and real-time data integration, unblocking near-real-time analytics across downstream consumer teams.
- Built Python API connectors for heterogeneous data sources, loading structured data into Snowflake and BigQuery using SQL and PySpark, extending the cloud data product catalog for client-facing analytics.
- Translated raw pipeline output into stakeholder-facing Tableau dashboards, communicating data findings directly to business decision-makers across client engagements.
- Shipped data integration features across the full delivery cycle, from requirements gathering and architecture through implementation and deployment in an agile team.
- Trained and mentored interns and junior developers, bridging technical depth with clear, accessible communication.

AI Engineer

Jul 2020 – May 2021

Xehen · Karachi, Pakistan

- Automated data collection across 50+ e-commerce websites by building Python scrapers with Scrapy, Selenium, and BeautifulSoup, storing structured output in PostgreSQL for downstream ML consumption.
- Developed a real-time attendance system using Hugging Face facial recognition and ensemble ML models, delivering a data-intensive solution that eliminated manual check-in processes for a school network.
- Trained a 96-class dog breed classifier with YOLOv3 and landmark-based features, achieving 97% accuracy and demonstrating production-ready model development along with backend on Flask.

PROJECTS

Restaurant Footfall Forecasting

ARIMA · XGBoost · Streamlit

Built a customer-visit forecasting system comparing ARIMA and XGBoost across multiple client scenarios, achieving a ~15% MAE improvement; deployed as a live Streamlit dashboard for continuous business monitoring.

Sales Forecasting System

ARIMA · LSTM · AWS QuickSight

Designed a time-series forecasting pipeline using ARIMA and LSTM to predict sales trends, with results surfaced through AWS QuickSight for stakeholder consumption.

YouTube Transcript RAG

LangChain · FAISS · OpenAI Embeddings · Mistral-7B · HuggingFace

Built a production-style RAG pipeline: ingested YouTube transcripts, chunked with RecursiveCharacterTextSplitter, embedded with OpenAI text-embedding-3-small, indexed in FAISS, and routed queries to Mistral-7B for grounded generation.

Face Detection and Tracking · MTCNN · OCSORT · PyTorch

Built a face detection and multi-object tracking pipeline using MTCNN for frame-level detection and OCSORT for temporal association across video streams.

TECHNICAL SKILLS

- Languages:** Python, SQL
- ML & DL:** Scikit-learn, PyTorch, TensorFlow, ARIMA, LSTM, diffusion models, transfer learning, equivariant GNNs
- Data Engineering:** Apache Spark, PySpark, Kafka, Airflow.
- Backend:** Flask, Django
- Cloud & DevOps:** AWS, Azure, Docker, Git, CI/CD, pytest, MLflow
- Databases:** Snowflake, BigQuery, PostgreSQL, MongoDB
- Data Analytics:** Pandas, NumPy, SciPy
- Visualisation & Dashboards:** Tableau, Streamlit, Matplotlib, Seaborn
- LLMs & AI:** RAG, OpenAI API, LangChain, Hugging Face, prompt engineering

EDUCATION

MS in Data Science Oct 2023 – Expected December 2027

Freie Universität Berlin · Berlin, Germany

Coursework: Machine Learning, Statistics, Data Management, Data Visualisation, LLMs, Semantic Technologies.

MS Thesis, BIFOLD (TU Berlin): *Transfer Learning for Bridging Machine Learning Force Fields and Diffusion Models*

Multi-head model trained jointly on equilibrium-structure and non-equilibrium molecular datasets; evaluated against MLFF and GPFF baselines on QM7x (PyTorch, SchNet/PaiNN).

BS in Software Engineering Jan 2017 – Mar 2021

Sir Syed University of Engineering and Technology · Karachi, Pakistan

Graduated 2nd in cohort, top grade band (1.1, German scale).

LANGUAGES

English (C1, working language) · German (A2, actively improving) · Urdu (native)