Mustafa Çavuşoğlu

Machine Learning Engineer · MLOps

cavusoglu.dev | github.com/mustafacavusoglu | linkedin.com/in/mustafacavusoglu12

ABOUT

Machine learning engineer with expertise in MLOps, big data pipelines and computer vision. Experienced in deploying and scaling ML models on cloud-native platforms and building end-to-end AI solutions with modern LLM APIs and frameworks.

WORK EXPERIENCE

MLOps Engineer · Intertech

Nov 2025 – Present

- Manage and scale ML model serving infrastructure on **OpenShift** using **KServe**, **Triton** and **vLLM** for high-performance LLM inference.
- Design and maintain CI/CD pipelines with **Azure Pipelines** and **ArgoCD** using containerized deployments with **Docker**.
- Operate and optimize MLOps workflows on OpenShift, ensuring reliable model lifecycle management from training to production.

Data Scientist · Teknosa

Sep 2023 – Oct 2025

- Built classification and recommendation models for campaign targeting on large-scale e-commerce data using **XGBoost** and **Dask** for distributed processing.
- Deployed end-to-end AutoML pipelines on on-prem servers.
- Implemented a **RAG** system with **ChatGPT API** and **LangChain** on complex databases.
- Optimized an ETL pipeline by migrating from Pandas to **Polars**, cutting runtime from **75 min to 15 min**.
- Built a production-ready multi-agent chatbot with **FastAPI**, **Redis**, **PostgreSQL** and **OpenAI API** on self-managed Ubuntu servers.
- Developed reports and dashboards using **SQL**, **Python** and **Streamlit**.

AI R&D Engineer · MLPCARE

Jul 2022 – Sep 2023

- Trained custom deep learning models (**CNN**, **YOLO**) on large-scale medical imaging data for segmentation and classification.
- Developed hybrid models combining deep learning with classical ML techniques.
- Ran parallel training on **HPC clusters** with **SLURM** on multi-channel MRI/DMR data (14–15 channels).
- Built RAG applications with **ChatGPT API**, **LangChain** and **Streamlit** for complex SQL query generation on medical databases.

TECHNICAL SKILLS

- **MLOps & Deployment:** Kubernetes, OpenShift AI, KServe, vLLM, Triton, Docker, Azure Pipelines, ArgoCD, FastAPI, MLflow
- **ML & DL:** PyTorch, TensorFlow, XGBoost, YOLO, Dask, Polars, Pandas
- **Data & Databases:** SQL, PostgreSQL, Redis
- **Tools:** Git, DVC, Streamlit, LangChain, OpenAI API, SLURM
- **Languages:** Python, Rust, SQL · English (professional)

PROJECTS

- **Axon** — Triton-compatible ML model server shipped as a single binary: ONNX Runtime on CPU and NVIDIA GPU (CUDA/TensorRT), KServe v2 HTTP + gRPC, model ensembles, Rhai scripting. *Rust, ONNX Runtime, TensorRT, gRPC*
- **Citewise** — Agentic RAG template that cites and verifies its sources, with a ChatGPT-style UI. *LangGraph, Qdrant, LiteLLM, mem0, FastAPI, msgspec*

PUBLICATIONS

- **2023** — *8th International Conference on Computer Science and Engineering (UBMK)*, pp. 578–583, IEEE. Link
- **2023** — *Breast lesion detection from DCE-MRI using YOLOv7*, AIP Conference Proceedings, Vol. 3030, No. 1, AIP Publishing. Link

EDUCATION

BSc in Geomatic Engineering · Yıldız Technical University *Sep 2017 – Jun 2022* Relevant courses: Image processing, photogrammetry, surveying, satellites