Yasser El Haddar

Tech Lead - Machine Learning Ops Engineer

yasserelhaddar@gmail.com +447765231969 Manchester, UK

[Personal] github.com/Yasserelhaddar | [Mindtrace] https://github.com/YasserElHaddar16

PROFESSIONAL EXPERIENCE

Mindtrace Limited, *Tech Lead - Senior Machine Learning Ops Engineer* 03/2022 – present | Manchester, UK

Contributed to the development and delivery of asset inspection and manufacturing defect detection solutions for prominent companies such as IBM, NV5, ESmart, and SharperShape. Here are some tasks I worked on:

• Implemented and deployed deep learning models using YOLO and Dinov2 for defect detection, achieving 92% accuracy in defect classification.
• Designed an innovative self-learning pipeline, resulting in up to 10% improvement in model performance across multiple iterations.
• Enhanced point cloud segmentation pipeline implementing SOTA models (PointNet2, SPVCNN, Superpoint Transformer), optimizing processing time by 50% and reducing operational costs by 60%.
• Developed and implemented an Apache Airflow workflow to automate Jira updates, streamlining the entire point cloud segmentation process.
• Pioneered 3D reconstruction pipeline using COLMAP and Hierarchical-Localization (HLoc) for creating precise silo point cloud representations from images, achieving sub-centimeter accuracy.
• Deployed and maintained server-less production pipelines in GCP, ensuring high availability and scalability of services.
• Utilized RESTful APIs in micro-services architecture with Docker and FastAPI for efficient system management, achieving high-throughput and low-latency performance.
• Contributed to internal framework development, implementing robust data lake architecture and database solutions using SQLAlchemy and Alembic with GCP SDK and Huggingface dataset.
• Developed CI/CD pipelines, facilitating automatic updates of Docker images and streamlining the deployment process for multiple ML models used in inferencing.
• Implemented a serverless PDAL-based preprocessing module in GCP.
• Contributing to the development of a new framework providing a unified interface for managing data, models, and physical infrastructure, including datalake and model bank management.
• Leading initiatives to integrate cluster orchestration, industrial hardware connectivity, and streamlined annotation pipelines into the framework, simplifying and accelerating ML operations for enterprise clients.
• Contributed to manufacturing defect detection pipelines, adapting vision models and infrastructure to handle high-speed production line imagery with strict latency requirements.
• Architected and implemented comprehensive camera management ecosystem including web-based configurator app, CLI tools, and streaming services with MJPEG support, achieving real-time video processing with configurable quality/FPS control and RESTful API integration.
• Designed and implemented complete sensor data ecosystem with MQTT support, featuring AsyncSensor framework with multi-backend architecture (MQTT, HTTP, Serial), sensor simulators for integration testing, and Model Context Protocol (MCP) integration, achieving 92% test coverage across 170+ tests.
• Added industrial automation support in Mindtrace, building initial PLC communication components with Allen-Bradley PLCs to enable machine-to-software integration.
• Worked on Poseidon, the customer-facing portal for asset inspection applications, implementing modules for data acquisition, hardware management, user management, inference and deployment, and full model lifecycle support.
• Assumed Tech Lead role, directing project strategy and mentoring team members in advanced technical areas, ensuring high-quality deliverables and efficient problem-solving.
• Conducted technical interviews and contributed to team building and talent acquisition, playing a key role in team growth and talent development.

Tech Stack: Python, Bash Script, PyTorch, GCP, GitHub Actions, Docker, PDAL, Pandas, Airflow, MLFlow, Huggingface dataset, RESTful APIs, Hydra, SQLAlchemy, Alembic, COLMAP, HLoc, Reflex, OpenCV, Reflex

Vitesco Technologies, *Data Scientist* 06/2021 – 03/2022 | Regensburg, Germany

- Developed a machine learning-based approach to predict defects in SCR Injector manufacturing, analysing key parameters such as welding voltage, current and feed rate.

- Implemented and compared various machine learning models, including Random Forest, SVM, and XGBoost, to identify the most effective approach for defect prediction.
- Utilized SPOT for balancing the dataset and applied LIME for model explainability.

Tech Stack: *Random Forest, SVM, XGBoost, SPOT, LIME,* scikit-learn, KNIME, Pandas, NumPy, Matplotlib.

Technology Campus of Plattling, *Master Thesis* 01/2021 – 06/2021 | Plattling, Germany

- Experimented with GANs to enhance simulated LiDAR and RADAR point cloud data for realistic environmental representation.
- Transformed low-density LiDAR/RADAR data into higher density formats to improve annotation and segmentation.

Tech Stack: *GANs (Generative Adversarial Networks), Keras, PyTorch, Carla Simulator.*

PROJECTS

ShopSense-AI, *Intelligent Shopping Assistant Platform* 01/2025 – Present

- Architecting end-to-end microservices-based AI shopping platform with three independent services: Knowledge Engine (LLM training/inference), Discovery Engine (product data collection), and Advisory Engine (user-facing recommendations)
- Implementing custom LLM fine-tuning pipeline using QLoRA on Mistral-7B-Instruct with WandB experiment tracking, achieving production-ready shopping consultation capabilities
- Building real-time product discovery system integrating multiple APIs (Appify, Best Buy, RapidAPI) with Qdrant vector database for semantic search, processing 100+ products per query with sub-50ms response times
- Designing production-grade infrastructure with Docker containerization, multi-database architecture (PostgreSQL, Redis, Qdrant), and comprehensive health monitoring across distributed services
- Developing intelligent recommendation engine combining vector similarity search with LLM-generated advice, featuring conversational AI interface with context-aware product suggestions

Tech Stack: Python, FastAPI, Docker, Qdrant, PostgreSQL, Redis, OpenAI API, Hugging Face Transformers, WandB, Mistral-7B, QLoRA, sentence-transformers, async/await, UV

CacheFuse, *LLM Caching Framework* 02/2025 – 08/2025

- Developed production-ready caching framework for Large Language Models, achieving 60-90% API cost reduction and 100x latency improvement (3ms vs 2-5 seconds) for cached queries
- Implemented intelligent cache invalidation with TTL management, tag-based bulk operations, and stampede protection for concurrent requests
- Built multi-backend architecture supporting SQLite (local) and Redis (distributed) with privacy-compliant hash-only mode and deterministic cache key generation
- Designed drop-in decorator system (@llm, @embed) with zero-configuration setup and comprehensive CLI tools for cache management

Tech Stack: Python, Redis, SQLite, OpenAI API, Anthropic API, async/await, pytest, SHA256 hashing

SkewSentry, *ML Feature Validation* 03/2025 – 10/2025

- Created automated feature parity validation system preventing training/serving skew in ML deployments, addressing 70% of production ML failures through pre-deployment detection
- Implemented multi-source adapter pattern supporting Python functions and HTTP APIs with configurable tolerance thresholds for numeric and categorical features
- Built comprehensive validation pipeline with CI/CD integration, generating HTML/JSON reports and providing exit codes for automated deployment gates
- Designed YAML-based configuration system for feature comparison rules with support for time-series windowing, categorical validation, and null policy management

Tech Stack: Python, FastAPI, Pydantic, pandas, PyYAML, pytest, HTML reporting, CI/CD integration

MCP-DS-Toolkit-Server, *AI-Powered Data Science* 05/2025 – 09/2025

- Built standalone Model Context Protocol (MCP) server enabling AI assistants (Claude, Cursor) to perform complete data science workflows through natural language
- Implemented 29 specialized tools covering data loading, preprocessing, model training (14+ algorithms), and experiment tracking with SQLite-based persistence
- Designed local-first architecture ensuring data privacy while providing production-quality code generation and comprehensive performance metrics

- Created intelligent recommendation system for algorithm selection, hyperparameter tuning, and preprocessing strategies with automated experiment lineage tracking

Tech Stack: Python, MCP Protocol, SQLite, scikit-learn, pandas, uvx/uv, pytest, cross-platform deployment

Football Smart Tactician, *AI powered analysis of football footages* 08/2024 – 03/2025

• Developed an end-to-end ML-powered football analysis system using YOLOv11 for multi-class detection (players, ball, referees, goalkeepers) and pose estimation for field keypoints, achieving real-time processing capabilities.
• Implemented advanced computer vision pipeline incorporating homography projection to transform detected entities into a 2D tactical view.
• Engineered comprehensive analytics system generating heat maps, formation analysis, team pressure metrics, and pass networks, providing deep tactical insights.
• Integrated OpenAI LLM for automated tactical analysis, creating a novel system that transforms computer vision analytics into detailed tactical commentary and insights.
• Designed robust MLOps pipeline using Label Studio for annotation workflow, implementing automated model training and deployment through MLFlow, enabling continuous model improvement through human-in-the-loop feedback.
• Built scalable FastAPI backend with modular router architecture handling video processing, model fine-tuning, authentication, and annotation workflows.
• Implemented asynchronous video processing and model training pipeline using Celery for background task management, enabling efficient handling of compute-intensive operations.
• Developed authentication system using OAuth2, ensuring secure access control and user management for the platform.
• Created interactive React dashboard for tactical visualization and analysis, providing intuitive interface for football analytics insights.
• Architected containerized infrastructure using Docker Compose with Nginx reverse proxy, enabling SSL termination, load balancing, and serving static content.

Tech Stack: Python, FastAPI, React, Docker, Nginx, YOLOv11, OpenAI API, Label Studio, MLFlow, Celery, OAuth2, PostgreSQL

Finance Web App 04/2024 – 06/2024

- Development and deployment of a Streamlit web application for visualising stock prices and technical indicators.
- Creation of a real-time webpage for monitoring intraday stock prices, foreign exchange rates, and market news.
- Development of a RAG-enhanced chatbot powered by GPT-3.5 Turbo, designed to address queries related to EDGAR filing data.
- Design and construction of a prediction pipeline for daily stock market forecasts using the Temporal Fusion Transformer

Tech Stack: Python, Streamlit, LangChain, OpenAI API, PyTorch Forecasting, Alpha Vantage API, Pandas, NumPy, scikit-learn

EDUCATION

Technical University of Deggendorf, 03/2020 – 03/2022 | Deggendorf, Germany
MSc in Electrical Engineering and Information Technology

CERTIFICATES

- Coursera - Generative AI with Large Language Models: MYSZH9VYRE4J
- Coursera - IBM AI Engineering Specialization: 5BTJSWY5FFD7
- Coursera - Introduction to Self-Driving Cars: JN8ERNP86BR5