

Achref Jaray

Senior AI engineer | Computer Vision Engineer

✉ achref.jaray1997@gmail.com



+216 22076218



Achref Jaray

TECHNICAL SKILLS and INTERESTS

Programming & Tools: Python (NumPy, Pandas, Scikit-learn, TensorFlow, PyTorch), LangChain, Docker, FastAPI, Git, MLflow, Weights & Biases, Vector Databases (FAISS, Chroma), Streamlit, Ollama, MCP.

Data Science & MLOps: R&D, Data Analysis, Generative AI Studio, MLOps, Prompt Engineering, Distributed Training

Artificial Intelligence & Computer Vision: Object Detection, VLM (Vision-Language Models), LLMs, RAG (Retrieval-Augmented Generation), Fine-tuning (QLoRA, LoRA), Model Evaluation & Robustness, Hallucination Reduction

Project Management : Jira

Assisted Coding: VS Code, GitHub Copilot, Qwen Code, Kaggle.

EXPERIENCE

- Perfanalysis Consulting

Jan, 2026

Senior AI engineer

Lac, Tunisia

- Led end-to-end development of a Football AI System for performance analysis using Computer Vision and Deep Learning. Enhanced real-time object detection (D-Fine), multi-object tracking (Norfair, DeepSORT), and action recognition to extract tactical, technical, and physical metrics from match/training videos.
 - ♦ Reduced AI model processing time by 75% (from 10 min to 2.5 min per 30s of video) through GPU-optimized batching and pipeline parallelization.
 - ♦ Achieved 80–95% object tracking accuracy with zero identity swaps using ReID-enhanced tracking modes during high-intensity match sequences.
- Designed & deployed an AI application layer based on MCP (Model Context Protocol) to power an intelligent AI Assistant capable of querying PostgreSQL and Neo4j databases, enabling real-time data enrichment, tactical analysis, and decision support.
 - ♦ Delivered sub-3s response times for complex tactical/historical queries, processing 500+ daily requests with 94% coaching staff satisfaction.
- Engineered an advanced Medical & Performance Document Analysis pipeline using LLMs combined with Agentic RAG and GraphRAG architectures. Built a Neo4j knowledge graph from player medical records and performance logs for semantic search, structured reasoning, and automated reporting.
 - ♦ Resolved 88% of medical staff Q&A queries autonomously, reducing manual report generation time by 70%.
- Implemented continuous model improvement mechanisms using structured user feedback loops. Collected annotated feedback from coaches/medical staff to automatically retrain and fine-tune CV and RAG pipelines, leveraging MLflow for experiment tracking, model versioning, and CI/CD retraining.
 - ♦ Reduced model iteration cycles from 3 weeks to 3 days, boosting action recognition F1-score by 18% and RAG retrieval precision by 22% over 6 months.
- Directed QA, technical reviews, and cross-functional delivery, ensuring algorithmic accuracy, data compliance, and seamless AI/backend integration. Generated detailed test reports and maintained daily progress syncs via Jira.
 - ♦ Maintained <2% false-positive rate in tactical event extraction across 500+ hours of footage.
 - ♦ Achieved 98% on-time sprint delivery and reduced post-deployment defects by 65% through rigorous test reporting and daily Jira stand-ups.

Technologies: Python, PyTorch, Ultralytics, OpenCV, Supervision, LangChain, LlamaIndex, Neo4j, Agentic RAG, GraphRAG, Semantic Search, MLflow, FastAPI, Docker, PostgreSQL, MCP.

- Tunisian Armed Forces

Jun, 2021

AI engineer

Tunis, Tunisia

- Developed specialized AI solutions tailored for defense and operational needs, including a Chrome Extension (Manifest V3) for intelligent document analysis enabling real-time Q&A, summarization, and interactive chat with classified documents.
 - ♦ Reduced document review time by 65% while maintaining 94% contextual accuracy, processing 500+ classified reports/month with zero data exfiltration incidents.
- Built a real-time streaming video analysis system capable of full offline inference for object detection, tracking, and anomaly detection directly on edge devices.
 - ♦ Achieved 30 FPS real-time inference with <45 ms latency on constrained edge hardware by leveraging ONNX Runtime, INT8 quantization, and operator fusion.
 - ♦ Attained 91% anomaly detection precision with zero false negatives during live field drills, operating entirely offline.
- Designed and implemented robust offline-first architectures to ensure reliable performance in environments with limited or no internet connectivity.
 - ♦ Maintained 99.8% system uptime during 30-day zero-connectivity deployments, eliminating cloud dependency and enabling seamless local caching/sync with <2% data drift.

Technologies: Chrome Extension API (Manifest V3), JavaScript, ONNX Runtime, Edge AI optimization, Ollama

· Master's Thesis, polytechnic school of Tunisia

Apr, 2024
Tunis, Tunisie

AI Research Internship

Project: Abnormal Detection in Drone Video Surveillance

- Enhanced YOLO World with DCNv3, Coordinate Attention & AMMF modules + SAHI post-processing for robust anomaly detection in low-visibility drone footage (FLAME dataset).
 - ♦ +23% mAP@0.5 (68.4% → 84.1%) and +31% small-object recall; 57% fewer false positives in fog/night conditions.

Technologies: YOLO World, DCNv3, SAHI

· Final Internship, Military Research Center

Feb 2021

AI Engineer Internship

Tunis, Tunisie

- Improved YOLOv5 for mini-UAV detection using advanced augmentation (Mosaic-9, MixUp) and hyperparameter optimization (Optuna) to enhance aerial surveillance accuracy.
 - ♦ Achieved 90% precision and 87% recall (F1: 88.5%) on custom aerial test set; reduced false alarms by 42% vs. baseline.
 - ♦ Enabled real-time inference at 45 FPS on edge GPUs with <30ms latency, supporting 24/7 autonomous airspace monitoring

EDUCATION

Research master's in (Information System Techniques) , National Engineering School of Tunis ,ENIT 2023-2024

Geomatics Engineer, Borj l'Amri Aviation School, EABA 2018-2021

Preparatory cycle: Mathematics and physics, Borj l'Amri Aviation School,EABA 2016-2018

PROJECT

· Aerosol Optical Depth (AOD) Estimation

Solafune

- Achieved 3rd place in Solafune's AOD Estimation competition, achieving 0.9896 public and 0.9893 private scores. Applied a solution using Vision Transformer (ViT) fine-tuned for regression on Sentinel-2 data, supported by 5-fold cross-validation, dynamic learning rate adjustment with a cosine scheduler, and early stopping. Applied data preprocessing techniques, including image normalization and augmentation for better model generalization. Custom loss functions (L1 and correlation-based AoDLoss) were used for optimal alignment with AOD values. Final results averaged across folds for enhanced accuracy.

Technologies: GeoIA, Image processing, ViT

PUBLICATION

- Title: "An Upgraded-YOLO with Object Augmentation: Mini-UAV Detection Under Low-Visibility Conditions by Improving Deep Neural Networks"
- Journal: Operations Research Forum
- Publication Date: 30 Septembre 2022
- DOI : <https://doi.org/10.1007/s43069-022-00163-7>

LANGUAGE

- **Anglais:** professional
- **Français:** professional