

Yahia Ghallale

+39 339-849-1743 | ghallale.yahia04@gmail.com | linkedin.com/in/yahiaghallale | github.com/yahiag04

Education

- Politecnico di Milano** 2026 - 2028
MSc in High Performance Computing Engineering Milan, Italy
- **Relevant Coursework:** Software Engineering for HPC, Parallel Computing, Computing Infrastructures, Artificial Neural Networks & Deep Learning, Advanced Computer Architectures, Numerical Linear Algebra.
- Università degli Studi di Brescia** 2023 - 2026
BSc in Computer Science and Engineering Brescia, Italy
- **Relevant Coursework:** Machine Learning, Deep Learning, Statistics, Data Structures and Algorithms, Full-Stack Engineering
- Universidade da Beira Interior** 2025
Erasmus+ Blended Intensive Programme | From Smart Sensors to Industrial Applications Covilha, Portugal
- Completed hands-on labs in sensor networks, Industrial IoT, FPGA/PLC systems, industrial standards, and real-time monitoring with participants from five European partner universities.

Experience

- Prospecto** Jan 2026 – Apr 2026
Computer Vision Engineer Brescia, Italy
- Deployed a real-time computer vision pipeline on an NVIDIA Jetson Nano with ROS 2, Docker, MobileNetV3-Small, and ONNX Runtime, achieving 95.9% macro-F1 at 19.4 ms/frame (approximately 51 FPS) with fully on-device inference.
 - Automated field-video analysis with RF-DETR and BoT-SORT, producing annotated footage and structured CSV tracks while eliminating manual frame-by-frame review.
 - Fine-tuned RF-DETR on a custom hand-labeled agricultural dataset, achieving 0.49 mAP@0.50:0.95 and 0.52 AR@100 as the detector baseline.
- GIVI UniBS Motorsport** Mar 2024 – Mar 2025
Embedded Software Engineer Brescia, Italy
- Developed ESP32 firmware for real-time multi-sensor telemetry over Wi-Fi, integrating MLX90614 infrared and accelerometer data during track tests.
 - Built a Node.js backend for telemetry ingestion, storage, and REST access, enabling structured queries and post-session analysis.

Projects

- GPU Memory Broker & Cross-Process Sharing** | C++20, CUDA, Metal, Unix IPC, CMake
- Built a C++20 GPU memory broker with priority scheduling and model eviction to manage memory contention across 3 backends: CUDA, Metal, and mock.
 - Enabled processes to share a single GPU allocation per model via CUDA IPC, avoiding duplicate device buffers and managing imported handles through RAII.
- Open-Source Contribution – PyTorch** | Python, PyTorch, TorchInductor, PyTest
- Submitted a PyTorch core fix for consistent eager and `torch.compile` behavior in `torch.normal`, adding tensor standard-deviation validation to its decomposition and an Inductor CPU regression test.
- Pneumonia X-Ray Classification & Cross-Dataset Generalization** | Python, PyTorch, scikit-learn, FastAPI, PostgreSQL
- Built a reproducible PyTorch pipeline to benchmark five CNN architectures on 7,000 patient-level chest X-rays across RSNA and two external datasets.
 - Achieved 95.0% balanced accuracy on RSNA, 85.8% on Kermany, and 83.0% on Chittagong; used bootstrap confidence intervals and paired statistical tests to quantify cross-dataset robustness.
 - Integrated the trained classifier into Med-AI, an academic FastAPI/PostgreSQL prototype that prioritizes pneumonia-positive cases for review.

Technical Skills

Languages: Python, C++, C, Java, SQL, JavaScript/TypeScript
AI/ML: PyTorch, scikit-learn, OpenCV, ONNX Runtime, MLflow
Data & Backend: FastAPI, PostgreSQL, Redis, ChromaDB, Node.js
Infrastructure: Docker, AWS, GitHub Actions, Linux, Git, Prometheus, Grafana
Robotics & Edge: ROS 2, NVIDIA Jetson, ESP32, STM32