Nardi Xhepi

Experience

Doctolib Apr 2026 – Present

Machine Learning Engineer (intern, then full-time) *Paris, France*

- **Phone Assistant (production):** Shipped a streaming end-of-utterance detection model on continuous audio, raising recall from **56% to 90%**.
- **Consultation Assistant (clinical safety):** Built evaluation pipelines on a production golden dataset, with calibrated LLM-as-judge scoring and statistical drift monitoring, to measure model behavior and catch regressions in a regulated healthcare setting.
- **Speech stack:** Benchmarked and fine-tuned STT foundation models (NVIDIA NeMo, Parakeet) and evaluated TTS models (Gradium), weighing accuracy, latency and serving cost; deployed in production with Ray Serve.
- **Data quality & reliability:** Ran Kili annotation campaigns with human review, and diagnosed a pipeline bug that was corrupting **12%** of evaluation data, restoring trust in the metrics.

Dassault Systèmes Feb 2025 – Aug 2025

Research Scientist Intern, LLMs & AI Agents (ENOVIA) *Paris, France*

- **Hybrid retrieval:** Built RAG pipelines on Qdrant and knowledge graphs, combining BM25 with neural early/late-interaction retrievers.
- **Agents & extraction:** Developed LLM agents for entity–relation extraction on enterprise data; benchmarked open-weights models (Mistral Small, Llama 3.1, SmolLM 3B) to optimize the accuracy/compute trade-off.

EDF R&D Jul 2024 – Jan 2025

Research Scientist Intern, Time Series & Anomaly Detection (DOAAT) *Paris, France*

- **Benchmark to production:** Designed ensemble detectors for multivariate industrial time series, improving AUC-PR by **15%** over baselines, with up to **€60M** in avoidable annual losses at stake. Published at **ICDE 2026**.
- **Representation learning:** Built feature-based latent representations of time series to drive detector aggregation and transfer from benchmarks to production.

Skills

Production AI	Prototype-to-production delivery, evaluation harnesses, observability, latency/cost optimization, streaming inference, regulated-data environments
LLMs & Agents	RAG (hybrid retrieval), AI agents, fine-tuning, knowledge distillation, pseudo-labeling, open-weights model benchmarking
Evaluation	LLM-as-judge calibration, golden datasets, drift monitoring, human annotation (Kili)
Speech & Time Series	STT/TTS (NeMo, Parakeet, Gradium), multivariate anomaly detection, representation learning
Engineering	Python, PyTorch, Hugging Face Transformers, scikit-learn, Ray Serve, Qdrant, Git, Bash, Linux
Languages	French, English (fluent) · Albanian (native)

Publications

N. Vautier, P. Caron, **N. Xhepi**, F. Bizeul, M. Boumghar, et al. *From Benchmarks to Production: Transferring Time Series Anomaly Detection Methods for Electricity Production Monitoring.* **IEEE ICDE 2026**, Montréal, Canada. [hal-05630277]

Education

École Polytechnique Sep 2024 – 2026

M2 Data Science, highest honors *Palaiseau, France*

Télécom SudParis Sep 2022 – 2026

Diplôme d'ingénieur, Computer Science & Data Science *Évry, France*

Achievements & Interests

- **Bending Spoons First Ascent France** (Nov 2025): one of 20 students selected nationwide for the Milan finals.
- **D4GEN Hackathon, Genopole** (May 2025): 2nd place out of 100+ participants, patient-specific breast cancer treatment solution.
- **Hi!Paris Hackathon** (2024 & 2025): Technical Excellence Award, twice (highest technical score).