SOBHAN FOOLADI MAHANI

AI & DEEP LEARNING ENGINEER | RESEARCH & R&D

Barcelona, Spain | fooladi.sobhan@outlook.com | Linkedin | Github

SUMMARY

AI & Deep Learning Engineer with experience building end-to-end ML systems, deploying multilingual LLM pipelines, and developing efficient deep learning models for real-world applications. Strong background in NLP/LLMs, PyTorch-based model development, and ML pipeline engineering across cloud and edge environments. Currently conducting master thesis research at CTTC on energy and performance profiling of deep learning models across heterogeneous hardware. Skilled in designing scalable APIs, optimizing model performance, and delivering production-ready AI features. Motivated to work in research, ML engineering, and R&D roles combining deep learning, efficiency, and practical deployment.

TECHNICAL SKILLS

- **Programming:** Python, Bash
- **Deep Learning & ML:** PyTorch, Scikit-learn, NumPy, Pandas, Matplotlib, CNNs, LSTMs/GRUs, transfer learning, contrastive learning, self-supervised learning, model evaluation
- **NLP & LLMs:** Hugging Face pipelines, multilingual LLM integration, Llama models, embeddings, text classification, sentiment and topic modeling
- **Energy & Performance Profiling:** CodeCarbon, GPU/CPU/RAM profiling, efficiency analysis across edge and server devices
- **Backend & Infrastructure:** Flask APIs, RESTful services, AWS Lambda/API Gateway/S3, Dockerized endpoints, Linux systems, microservice workflow design
- **Tools:** JetBrains IDEs, VS Code, Colab, Jupyter, Git
- **Core Competencies:** Experimental design, ML pipeline engineering, research collaboration, data analysis, reproducible workflows

PROFESSIONAL EXPERIENCE

Master Thesis Researcher – Sustainable AI (SAI)- Part time **Oct 2025 – Mar 2026 (Present)**
CTTC – Centre Tecnològic de Telecomunicacions de Catalunya, Barcelona
Supervisors: Dr. Roberto Pereira, Dr. Paolo Dini

- Conducting research on energy-efficient deep learning, profiling training/inference costs and CO_2 emissions across heterogeneous hardware (Jetson Nano, laptop GPUs, server nodes).
- Benchmarking CNN, LSTM/GRU, and contrastive learning models for image and time-series tasks under diverse training settings.
- Designing measurement campaigns using CodeCarbon and CarbonTracker to evaluate GPU/CPU/RAM usage and energy scaling behavior.
- Deploying models on edge devices to study runtime constraints, memory limits, and batch-size effects.
- Exploring optimization techniques (contrastive learning, reservoir computing, coreset selection) to reduce model training cost while maintaining accuracy.
- Developed an internal tool at CTTC to automate model training pipelines and track energy use/CO_2 emissions during experiments. GitHub: SustainVision

AI & NLP Engineer - Freelance **Apr 2025 – Present**

Moodest, Techno Campus, Mataro, Barcelona

- Developed multilingual LLM and NLP pipelines (Catalan/Spanish/English) for workplace well-being and communication analytics.
- Hosted Llama 3.2 models on Hugging Face AWS Inference Endpoints and integrated them into Moodest's production NLP pipeline with automated context handling and agent routing.
- Designed agent-based orchestration and context-management modules improving processing accuracy and reasoning reliability.
- Achieved 96% message filtering accuracy and 91% engagement-score prediction using transformer embeddings + XGBoost models.
- Built privacy-compliant preprocessing and production-ready REST APIs powering Moodest's in-house AI services.

AI & Backend Engineer (R&D – MAXSENS Project) - Fulltime **Aug 2023 – Aug 2025**

Imaz Tech, Barcelona

- Developed Flask APIs for real-time EMG and motion data processing, enabling scalable modular AI pipelines.
- Built CNN and LSTM models achieving 94% gesture classification accuracy on high-dimensional physiological signals.
- Improved signal-processing efficiency by 30% through optimized filtering and feature extraction workflows.
- Automated analytics report generation (CSV, PDF, JPG), reducing manual handling by 60%.
- Implemented posture/motion detection features that reduced exercise execution errors by 40%.
- Collaborated with clinical and product teams to align ML outputs with real-world use cases.

Software Engineer / Technical Product Owner **Oct 2021 – Dec 2022**

SaminCoin, Tehran

- Designed and maintained RESTful APIs in Python/Django for fraud detection and transaction classification.
- Created detailed system architecture and UML diagrams, improving cross-team communication and reducing development cycle times by 25%.
- Led product roadmap planning and sprint execution, reducing time-to-market by 40%.
- Worked closely with UI/UX teams to optimize onboarding workflows for early users.

EDUCATION

Master's in Telecommunications – Deep Learning for multimedia processing Specialization **2023 – 2025**

Universitat Politècnica de Catalunya (UPC), Barcelona

Bachelor's in Information Technology Engineering **2015 – 2020**

Azad University, Karaj

CERTIFICATIONS & INTERESTS

Languages: English (Fluent), Spanish (A1)

Interests: Deep learning research, LLM systems, efficient ML, edge AI, signal analysis, chess