MERCEL VUBANGSI

+1-641-233-1997 ⋄ Fort Washington, MD
vmercel@outlook.fr ⋄ linkedin.com/in/mercel-vubangsi-68109b23a

Professional Summary

Lead AI Engineer with 10+ years' experience designing production-grade AI and software systems across healthcare, finance, and robotics. Specialized in Deep Learning, Generative AI, Agentic Workflows, Embedded/Edge AI, and AI Ethics. Proficient in GPU optimization and distributed ML systems, with expertise in automating cloud deployments and ensuring system reliability via SLO/SLI and error budget reporting. Strong foundation in Python software engineering (microservices, event-driven systems, APIs, automated testing). Proven success in industrializing ML via MLOps (Kubernetes/Docker, CI/CD, Ray), achieving measurable performance and reliability gains. Committed to innovation, responsible AI practices, and cross-team collaboration in rapidly evolving technological landscapes.

Technical Skills

Programming	Python, C++, Java, Node.js, SQL
ML / AI Frameworks	PyTorch, TensorFlow, Scikit-learn, Reinforcement Learning, Generative AI, Transformers, RAG, Computer Vision, NLP, Agentic AI, crewAI, AutoGen, Prompt Engineering, Ray
Software Engineering	Microservices Architecture, Event-Driven Design, Domain-Driven Design (DDD), REST / gRPC APIs, Automated Testing (PyTest, Unit / E2E), CI / CD, Git, n8n
MLOps & Deployment	Kubernetes, Docker, ML Lifecycle Automation, Model Monitoring, Versioning, Kubeflow, MLflow, GPU Optimization (CUDA, cuDNN), Automated Cloud Deployment
Embedded & Edge AI	TinyML, Model Optimization (Quantization, Pruning, Distillation), Edge Deployment (Raspberry Pi, Arduino)
Data Infrastructure	Apache Spark, Delta Lake, Data Lakes, ETL Pipelines, NoSQL, Data Governance, GDPR Compliance, RabbitMQ, Elasticsearch
Cloud Platforms	AWS (SageMaker, ECS, Lambda, S3), GCP (Vertex AI, BigQuery), Kubernetes, Docker, Terraform, CI / CD
Monitoring & Reliability	Prometheus, Grafana, SLO / SLI, Error Budget Reporting
Ethics & Responsible AI	Fairness, Explainability, Transparency, Accountability, Bias Mitigation
Soft Skills	Critical Thinking, Creativity, Adaptability, Empathy, Curiosity, Lifelong Learning, Cross-Team Collaboration

Experience

Lead Software/AI Engineer 06/2025 – 10/2025
Aliennova, Pittsburgh, PA (Hybrid)

- Architected and deployed a multimodal AI security platform leveraging vision–language models and anomaly detection pipelines for real-time scam and fraud detection, achieving 98% precision in identifying fraudulent behaviors across text, voice, and visual modalities
- Engineered an academic research automation assistant using multi-agent orchestration (Azure AutoGen + LangGraph) that performs end-to-end literature review, data analysis, mathematical computation, and scholarly drafting from a single prompt; reducing research cycle time by ~70%.
- Optimized cross-agent communication, observability, and deployment through containerized microservices (Docker + Kubernetes) and GPU-accelerated pipelines (PyTorch + CUDA), integrating CI/CD validation, API security policies, and centralized analytics pipelines using BigQuery on GCP for telemetry and performance insights, ensuring 99.9% uptime for production workloads.

Lead Machine Learning Engineer 07/2022 – 08/2024
AI and Robotics Institute, Nicosia, Cyprus (Hybrid)

- Architected production-grade agentic AI system (LangGraph + RAG + PyTorch) integrating multi-agent orchestration, external API calls, and secured data pipelines; improving predictive accuracy 25 % in healthcare operations.
- Optimized GPU-bound ML workloads using CUDA and cuDNN, achieving 30% faster training and inference for Ray-based distributed models, improving scalability for real-time healthcare applications.
- Automated deployment and orchestration of ML services on AWS ECS and GCP Vertex AI using Terraform and Kubernetes, streamlining cloud operations and reducing provisioning time by 50%.
- Operated and monitored production environments with Prometheus and Grafana, triaging incidents to maintain 99.9% uptime and implementing SLO/SLI metrics with error budget reporting for customer-facing services.
- Developed Java (Spring Boot) orchestration service enabling cross-language communication between agentic workflows and enterprise APIs; validated interfaces with Python LangGraph agents under CI/CD pipelines.
- Implemented unit and E2E tests using PyTest and Jenkins pipelines for agent communication and service reliability, reducing debug time 25 % and ensuring consistent deployments.

- Collaborated with Engineering, QA, and program management teams to align on system requirements, troubleshoot Ray Serve inference endpoint issues, and deliver customer-facing ML tuning solutions with 98% satisfaction.
- Integrated Elasticsearch for real-time log analysis, enabling proactive issue detection and reducing mean time to resolution by 40%.
- Implemented CI/CD pipelines with SageMaker and Kubeflow, cutting release cycles from two weeks to two days while using MLflow for experiment tracking.
- Championed responsible AI practices, embedding Fairness, Explainability (SHAP), and Transparency into healthcare AI workflows to mitigate bias.

Software Engineer / Machine Learning Engineer 12/2019 – 03/2022
Advanced Analytics Lab, University of Bamenda, Cameroon (Onsite)

- Scaled TensorFlow ensemble models on GCP Vertex AI for 10M+ records/month, sourcing and feature-engineering datasets directly from BigQuery (SQL + scheduled queries), increasing student engagement by 15% and maintaining GDPR compliance..
- Launched a BERT-based healthcare chatbot on AWS Lambda, achieving 98% response accuracy for 50K+ monthly interactions.
- Applied DDD principles and Delta Lake pipelines for modular, scalable data workflows, boosting efficiency by 40%.
- Implemented automated testing with PyTest for E2E and unit test coverage.

Data Scientist (Quantum Signal Processing) 06/2013 – 12/2019
Quantum Systems, Electronics and Signal Processing Lab, University of Dschang, Cameroon (Hybrid)

- Built high-performance **C++** models for quantum signal analysis, improving accuracy by 20% and reducing latency by 30%.
- Developed modular Python services with reusable APIs for quantum data pipelines, instrumenting ingestion to BigQuery for large-scale aggregation and ad-hoc analysis, using test-driven design.
- Designed wireless sensor networks for real-time quantum signal acquisition with microcontrollers + cloud integration, using Arduino for prototyping.
- Prototyped lightweight Scikit-learn models (SVMs, decision trees) on embedded devices for anomaly detection in IoT sensor data.
- Implemented secure communication protocols (MQTT, Zigbee) for distributed sensor networks, ensuring reliability and low latency.

- Leveraged MATLAB and SciPy for signal processing and FFT to prepare quantum data for analysis.

Education

M.S. Computer Science (Online) | Maharishi International University, IA – 2027
M.S. Artificial Intelligence Engineering | Near East University, Cyprus – 2023
Ph.D. Computational Physics | University of Dschang, Cameroon – 2017
B.S. Software Engineering | University of Bamenda, Cameroon – 2020

Projects

- **Distributed ML Inference System** — Built a **Ray Serve**-based inference platform with **Golang** APIs and **RabbitMQ** for event-driven workflows, optimized GPU utilization with **CUDA**, achieving 25% faster inference for 50K+ daily queries.
- **Event-Driven ML Platform** — Designed Python microservices (**FastAPI** + **Kafka**) for distributed ML training/inference; reduced downtime by 40% via automated testing + **CI/CD**. Utilized **n8n** for low-code automation workflows to trigger pipelines.
- **Distributed Deep Learning Training** — Leveraged **Horovod** + **PyTorch** on **AWS EC2**; trained vision models 3× faster, saving 40% in cloud costs using **Terraform** for infrastructure provisioning.
- **Agentic Healthcare Chatbot** — Built a multi-agent system where **LangGraph** orchestrated the primary workflow, while specialized sub-agents were built with **Autogen** and deployed with **RAG** on **GCP**; processed 20K+ daily queries at 97% accuracy.
- **Quantum Edge AI Tools** — Built **C++/Python** hybrid models optimized for edge devices, improving real-time quantum signal analysis by 20%, and leveraging **Qiskit** and **Cirq** for quantum computing simulations.

Achievements

- Presenter, AI & IoT Conference 2023: *"Optimizing Moving target defense for cyber anomaly detection"*.
- Best Paper Award, Intl Conference on Artificial Intelligence in Everything 2022: *"A learning assisted approach to electronic and optical characterization of tin-doped Titanium Oxide"*.
- Published 15+ peer-reviewed articles in Quantum Physics and ML.