GUILLAUME RAKOTONJANAHARY TSANTANIAINA

AI/ML Engineer | Software Engineer | Cloud Architect GCP & AWS | Tech Lead

atr.guillaume@gmail.com | +261 38 66 261 00 | Antananarivo, Madagascar | linkedin.com/in/tsanta1146 | wikolabs.com

PROFESSIONAL SUMMARY

AI/ML Engineer and Software Engineer with 5+ years of experience specializing in Cloud Engineering (GCP & AWS), production-grade AI systems, and full-stack software development. Deep expertise across the full ML stack: Speech AI (ASR fine-tuning, Neural MT, TTS voice cloning), NLP/LLM architectures (RAG, multi-agent orchestration, model fine-tuning), Computer Vision (YOLO, CLIP, SAM), scalable data engineering, and cloud-native deployment (Kubernetes, CI/CD, MLOps). Delivered complex AI systems for international clients in healthcare, fintech, and e-commerce. Tech Lead of Wikolabs, an AI and automation studio. Master in Big Data & AI, ESTIA France (17.2/20). Trilingual: English (C1), French (native), Malagasy.

PROFESSIONAL EXPERIENCE

AI/ML Engineer, Speech AI and Cloud Architecture

April 2026 – Present

Maison du Numerique (MGVaovao) | Antananarivo, Madagascar

Architected and deployed an end-to-end real-time Speech-to-Speech translation system converting audio in 6 languages (FR, EN, DE, ES, IT, PT) into spoken Malagasy dialect output. 4-stage cascade pipeline on Cloud Run GPU (NVIDIA L4 24 GB) with Vertex AI MLOps orchestration.

Pipeline Architecture

- **VAD: Silero VAD v5** (MIT) — CPU-only, 32ms chunks, vad_filter=True for parallel ASR preprocessing; latency ~1ms per chunk, size ~2 MB, no VRAM required
- **ASR: Whisper large-v3-turbo INT8** (Apache 2.0) — 99 languages, quantized via CTranslate2 INT8; VRAM ~1.7 GB, 2x speed vs FP16; fine-tunable via HuggingFace Trainer (not fine-tuned at this stage)
- **Neural MT: NLLB-200-distilled-600M INT8 fine-tuned** (CC-BY-NC) — LoRA rank=16 fine-tuning for FR/EN/DE/ES/IT/PT to plt_Latn (Malagasy Officiel); chrF++ improved +2 to +8 pts over baseline 42-52; VRAM ~0.7 GB
- **TTS: MMS-TTS-MLG VITS fine-tuned** (CC-BY-NC, Meta) — 6 dialect checkpoints (Officiel, Merina, Betsileo, Betsimisaraka, Sakalava, Antandroy); fine-tuned via ylacombe/finetune-hf-vits with 80-150 samples per dialect in 1-2h on L4 Spot; VRAM ~0.5 GB per checkpoint
- Total VRAM footprint: 6.2 GB on L4 24 GB, leaving 17.8 GB margin enabling 3-5 concurrent WebRTC sessions per instance

GCP Cloud Infrastructure and MLOps

- **Cloud Run GPU (NVIDIA L4)**: GA June 2025; scale-to-zero at $0.0001867/sec (~$33/month at 50h); deployed with gcloud --gpu 1 --gpu-type nvidia-l4 --no-gpu-zonal-redundancy
- **WebRTC real-time transport**: Cloud Run CPU for signaling (ICE/STUN/TURN via Cloudflare); RTP Opus 32ms chunk streaming; DataChannel for dialect config and result delivery
- **Vertex AI Pipelines**: Kubeflow DAG (data_validation, nllb_finetune, tts_finetune, evaluate, register, deploy) triggered by Pub/Sub on new GCS dataset upload; $0.03/run
- **Vertex AI Custom Training** on L4 Spot (~$0.28/hr): NLLB LoRA fine-tuning at $1.68-3.48/run; TTS fine-tuning at $0.84/dialect; GCS checkpoint every 30 min

- **Vertex AI Model Registry**: versioned checkpoints with chrF++ and UTMOS metadata; candidate to staging to production lifecycle; 1-click rollback on regression
- **Vertex AI Monitoring**: weekly drift evaluation on 200 fixed reference phrases; chrF++ drop >2 pts triggers NLLB re-FT; UTMOS drop >0.3 triggers TTS re-FT; ~$3/month
- **GCS structured buckets**: datasets/raw/, datasets/processed/, models/ (nllb_lora_v*/ ~200 MB, mms_tts_{dialect}_v*/ ~145 MB), datasets/reference/; $0.020/GB/month
- **CI/CD**: Cloud Build (120 min/day free) and Artifact Registry; blue/green deployment to Cloud Run GPU; 4 GitHub repos (inference-server, ml-training, platform-web, data-pipeline)
- **Budget optimized**: ~$42.46/month total (Cloud Run GPU $33.60, Vertex AI Training $2.24, Monitoring $3.00, Storage $1.00, CI/CD $0.50)

Technologies: *Whisper large-v3-turbo INT8 (fine-tunable), NLLB-200-distilled-600M (LoRA fine-tuned), MMS-TTS-MLG VITS (fine-tuned), Silero VAD v5, CTranslate2, PEFT/LoRA, aiortc, FastAPI, Cloud Run GPU L4, Vertex AI Pipelines, Vertex AI Custom Training, Vertex AI Model Registry, Vertex AI Monitoring, Pub/Sub, GCS, Cloud Build, Cloud Scheduler, Cloud Monitoring, Kaggle T4, Python*

Generative AI Engineer

January 2026 – April 2026

Vohitra MG (Exponent) | Antananarivo, Madagascar

- **AI Email Classification and Routing**: VLM-based multi-format attachment analysis (PDF, images, tables) with Microsoft Graph API; SharePoint document extraction; Calendar-based workload distribution; monitoring dashboard tracking routing decisions and classification accuracy
- **Multi-Agent NL Interface for Facility Management**: LangGraph multi-agent system converting natural language to NoSQL queries; 6-layer RAG context enhancement pipeline with persistent conversational memory; automatic chart/table generation; RAGAS-based quality evaluation
- **RAG Knowledge Assistant for Regulatory Compliance**: Production RAG chatbot over 100+ documents (HTML/PDF/DOCX); hybrid retrieval combining vector similarity, BM25, Reciprocal Rank Fusion, and LLM reranking; daily document sync with hash-based change detection; parallel PDF extraction using Docling for high-fidelity document parsing

Technologies: *Microsoft Graph API, Qwen3 LLM (VLM), LangGraph, LangChain, Docling, FastAPI, React TypeScript, PostgreSQL, Elasticsearch, RAGAS, Docker, Jenkins CI/CD*

Fullstack AI Engineer and Tech Lead

February 2025 – October 2025

FinAlchemy Ltd | Remote, United Kingdom

- AI document processing engine using Vertex AI Gemini Flash for multi-format OCR from pension documents; achieved >95% first-pass accuracy on provider document extraction
- Provider communication automation integrating OpenAI Whisper (STT) and TTS for AI-driven phone call handling, IVR navigation, and automated email sequencing with SLA tracking
- FCA COBS 9/19, PROD, and Consumer Duty compliance framework with real-time breach alerts, immutable audit trails, and automated suitability report generation
- Full-stack platform on FastAPI and React.js deployed on Google Cloud Run with MongoDB Atlas; GDPR-compliant infrastructure with GitHub Actions CI/CD

Technologies: *FastAPI, React.js, Google Vertex AI (Gemini Flash), OpenAI Whisper/TTS, MongoDB Atlas, Cloud Run, Docker, GitHub Actions*

Tech Lead and AI Full-Stack Engineer

April 2025 – Present

Mediwyz.com | Remote, Mauritius / East Africa

- Architected full-stack SaaS digital health platform (Next.js 15, NestJS, Prisma, PostgreSQL) connecting patients with 17 provider types (doctors, nurses, dentists, physiotherapists, labs, pharmacies, etc.) across Mauritius and East Africa
- **Provider and Service Discovery**: Multi-criteria search for healthcare providers with specialty, location, availability, and rating filters; each provider and admin can create services and configurable consultation workflows containing multiple status steps and automated notifications
- **Appointment and Booking System**: Provider time slot availability management with calendar scheduling; patient booking flow with confirmation, reminders, rescheduling, and cancellation; conflict detection and real-time slot locking
- **Video Consultation and Messaging**: Real-time peer-to-peer video calls via WebRTC with Socket.IO signaling and presence management; in-app messaging chat and call features using WebSocket for bidirectional real-time communication
- **AI Provider Onboarding with Document OCR**: VLM-based OCR pipeline for automated extraction and verification of required registration documents (medical licenses, national IDs, professional credentials); validation of required fields and document authenticity checks
- **RAG Health Assistant**: LLaMA-based RAG assistant enabling patients to query health history, lifestyle logs, and receive contextualised wellness recommendations
- **Health Shop and Inventory Management**: Multi-provider health product marketplace with inventory tracking, stock alerts, and category-aware filtering; any provider type can list products and services for sale
- **Laboratory Testing Services**: Lab test ordering, result delivery, and integration with patient health records; lab providers manage test catalogs and result workflows
- **Payment Integration**: MCB Juice mobile money payment gateway integration for consultation fees, product purchases, and service payments
- **Configurable Workflow Engine**: Strategy and registry design patterns powering 33+ consultation workflow types with ~310 status steps, automated notifications, and role-based state transitions
- Shipped 750+ tests and 40+ API routes; owned SEO strategy, OG image design, and VPS deployment; mentored junior engineers on architecture and code quality

Technologies: *Next.js 15, NestJS, TypeScript, Prisma, PostgreSQL, WebRTC, Socket.IO, RAG (LLaMA), VLM (OCR), MCB Juice API, Docker, GitHub Actions, Google Compute Engine, Cloud SQL*

Tech Lead

December 2025 – Present

Wikolabs | Antananarivo, Madagascar

- **Autonomous AI Sales Agent**: Multi-agent pipeline (Google ADK, Vertex AI, LangGraph) automating full B2B sales cycle from cold lead sourcing to deal closing; RLHF-based outreach optimization; real-time bidirectional CRM sync and sales performance dashboard with conversion analytics by agent, stage, and channel
- **Local Service Provider Search by Natural Language or Image Upload**: Multimodal search and recommendation platform for local service providers; users describe a need in natural language or upload a photo of the issue; system returns ranked geographically relevant providers; benchmarked Gemini 2 Multimodal Embedding against CLIP on Vertex AI and selected Gemini 2 as production backbone; BigQuery Vector Search with geolocation-aware clustering and provider metadata ranking; integrated Mobile Money payment gateway (Mvola, Orange Money) for direct service booking and provider payment from search results
- **Intelligent E-commerce Product Catalogue Search**: Image upload or natural language query returning ranked catalogue results with category-aware filtering, stock availability, and commercial metadata; same multimodal embedding architecture as provider search; integrated Mobile Money payments (Mvola, Orange Money) for seamless in-app purchase from search results

Technologies: *Google ADK, Vertex AI, Gemini 2 Multimodal Embedding, CLIP (benchmarking), BigQuery Vector Search, LangGraph, FastAPI, React TypeScript, Next.js, Python, RLHF, Mvola API, Orange Money API*

Data Scientist (Permanent Contract)

August 2025 – January 2026

eTech CDI, TASKFORCE AI Program | Antananarivo, Madagascar

- **Lead Generation Multi-Agent System**: Google ADK and LLM orchestration with multi-source intelligence gathering from social media and company data; pattern-based investment signal analysis
- **Biometric KPI Data Warehouse**: Migrated OLTP PostgreSQL to BigQuery star schema (fact and dimension tables, materialized views); Looker Studio dashboards; BigQuery ML ARIMA+ forecasting for kit usage prediction
- **No-Code E-commerce Product Recommendation App**: NLP and image recognition on GCP; RAG-based chatbot with configurable modes (image, text, combined); Google Sheets catalog integration; under 3s response time
- **SQL Generation Model Benchmarking**: Evaluated SQLCoder-7B-2, GPT-3.5/4, and Claude on architecture, hallucination mitigation, BigQuery syntax, and Spider benchmark; produced multi-cloud deployment guide
- **AWS vs GCP Innovation Benchmarking**: SageMaker vs Vertex AI, Bedrock vs Model Garden, Glue vs Dataflow; TCO analysis and multi-cloud migration guide
- **Real-Time Biometric Enrollment KPI Platform**: PostgreSQL replicated via Airbyte CDC to BigQuery; dbt SQL transformations; Looker Studio real-time dashboards for mobile biometric kit performance monitoring
- **BI Agent Chatbot (Text-to-SQL)**: LangGraph and RAG-powered text-to-SQL solution using Python, FastAPI, and React TypeScript; retrieves SQL patterns, DDL, and metadata for multi-query execution over BigQuery
- **Technical Support RAG Agent**: LangGraph and RAG-powered system for kit usage guidance, troubleshooting, and technical support using biometric process documentation for context-aware assistance

Technologies: *Google ADK, Vertex AI, BigQuery, BigQuery ML (ARIMA+), dbt, Airbyte CDC, Looker Studio, LangGraph, Gemini 2.5 Flash/Pro, SQLCoder-7B-2, FastAPI, React TypeScript, AWS SageMaker, Bedrock, Glue, Terraform, Docker*

Data Scientist Junior (Permanent Contract)

January 2025 – August 2025

eTech CDI, TASKFORCE AI Program | Antananarivo, Madagascar

- **Intelligent Product Catalogue System**: Image-based product search using Gemini 2.5 Pro and Vision API for automatic product description generation; auto-indexing and catalog management
- **Distributed Big Data Clustering**: Apache Spark and Google Dataproc pipeline for distributed clustering of 105,000+ e-commerce product images; automatic categorization with Spark ML and image augmentation
- **Multilingual Airport Conversational AI Assistant**: Deployed on Google Kubernetes Engine (GKE) with WhatsApp Business API integration; NLP for flight status, navigation, and services; BigQuery for query analysis; real-time translation and multilingual NLP
- **Generalized Chatbot Architecture Framework**: Reusable LangGraph design patterns with UML methodology; modular architecture with GoF design patterns, API access, external service integration, and RAG knowledge base
- **Computer Vision for Retail Inventory Counting**: YOLO V8 fine-tuned for object detection with instance segmentation and TensorRT optimization; multi-class detection REST API; real-time inference pipeline; presented at AI Event 2024
- **AI Product Search by Natural Language**: FastAPI and React TypeScript; BigQuery Vector Search RAG; generative AI for intuitive search and product recommendations
- **Project Specification Leadership**: Comprehensive project backlogs with Use Cases, User Stories, Business Rules, Development Effort Estimation, Cloud Cost Estimation, and UML diagrams (Deployment, Activity, Sequence, Use Case, Class, Component)

Technologies: *Python, Apache Spark, Spark ML, Google Dataproc, YOLO V8 (fine-tuned), TensorRT, OpenCV, LangGraph, Gemini 2.5 Pro, Vision API, FastAPI, React TypeScript, BigQuery Vector Search, GKE, WhatsApp Business API, Draw.io, UML*

Data Scientist Junior (Consultant)

December 2023 – December 2024

ETech Consultant | Antananarivo, Madagascar

- **Hybrid RAG Architecture (Vector + Knowledge Graphs)**: Hybrid RAG system combining Neo4j knowledge graph, Pinecone vector search, and Elasticsearch; document and image extraction from cloud infrastructure; served as Master's Thesis in AI
- **Massive Vectorization ETL Pipeline (105,000 images)**: Google Dataflow and Apache Beam distributed ETL pipeline for vectorizing 105,000 images into BigQuery; K-means clustering; vector similarity search across images and text descriptions
- **Automated BI Assistant**: LangGraph with memory persistence, Vertex AI and BigQuery integration, SQLCoder fine-tuning on BigQuery dialect, automatic visualization dashboard
- **CV Recommendation Chatbot**: NER-based skills extraction with spaCy, NLP semantic matching for candidate-job pairing; Python FastAPI and React.js; advanced recruiter interface with ranking
- **Computer Vision for Retail Inventory Counting**: YOLO V8 with instance segmentation and TensorRT optimization; multi-class detection REST API; real-time inference pipeline; presented at AI Event 2024
- **AI Email Classification Agent**: Microsoft Graph webhooks, SharePoint Excel integration, LLM-based routing and triage for a Swiss industrial client
- **AI Product Search by Natural Language**: Automated assistant for e-commerce product search (flowers platform, water services); BigQuery Vector Search RAG and generative AI; Python FastAPI and React TypeScript
- **Malagasy ASR/TTS Architectural Analysis**: Architectural study of ASR and TTS systems for Malagasy as a low-resource language using Chirp (Vertex AI Studio); established the transformer adaptation methodology later applied in full fine-tuning at Maison du Numerique

Technologies: *LangGraph, CrewAI, Neo4j, Pinecone, Elasticsearch, Apache Beam, Google Dataflow, BigQuery, YOLO V8, TensorRT, OpenCV, Microsoft Graph API, Chirp (Vertex AI), FastAPI, React TypeScript, spaCy, Python, MLOps*

Data Engineer and Software Engineer, Health Data Systems

January 2024 – August 2024

JSI Research and Training Institute (USAID Consultant, CHISU Project) | Madagascar, Remote

- Designed comprehensive technical architecture and specifications for new data collection modules supporting malaria bulletin generation with dashboard solutions, enhancing availability and interoperability of health information systems between the Ministry of Health and Program Office Malaria information systems (Microsoft Access and DHIS2)
- Solution used React.js, Chart.js, and LLM models for natural language to SQL query conversion with DHIS2 API management (CRUD operations) and external custom data source integration; collaborated with multidisciplinary team including Data Scientists, Malaria Advisors, Program Officers, ICT Health Informaticians, and Health Informatics Advisors
- Architected and deployed DHIS2 instances in development and pre-production environments; developed custom forms and created data elements and KPIs including WASH indicators, PSERAN 2022-2026 indicators, and PARN/HORS PARN indicators collected at UPNNC level
- Implemented using Java DHIS2 framework, Tomcat server, PostgreSQL database, and Linux server configuration

Technologies: *DHIS2, Java, Tomcat, PostgreSQL, React.js, Chart.js, LLM (NL to SQL), Microsoft Access, Linux, DHIS2 API*

Data and Software Engineer, PMI Measure Malaria Program

March 2023 – September 2023

JSI Research and Training Institute (USAID, Self-employed) | Madagascar

- **COVID-19 Data Anomaly Detection**: DHIS2 platform module using React.js and Node.js automating detection of duplicates, missing values, outliers, and non-applicable entries in COVID-19 health data; maintains dataset consistency for accurate public health assessments
- **DHIS2 Data Validation and Quality Control**: Design and development of quality control modules for DHIS2 data validation and parameter optimization
- **Vaccine and Healthcare Facility Geolocation Application**: Vaccine stock management and healthcare facility navigation tool for the USAID PMM Program; vaccine inventory visualization with search by type and date; geolocation features displaying healthcare and vaccination centers on interactive maps with turn-by-turn directions
- **Digital Health Document Library (Ministry of Health)**: Comprehensive digital repository for the Ministry of Health focusing on digitizing and organizing health documents, research papers, and informational resources; improved accessibility, categorization, and retrieval systems integrated with existing Health Information Systems for evidence-based decision-making
- **Mobile Application Support**: Technical support for expanding scorecard and dashboard mobile application tools across selected regions; contributed to vaccination site visualization mobile app development using React Native
- **Gmail Mailing List Management App**: Individual and group email management application using Gmail API and Google Cloud Service Account; built with Electron.js, React.js, and SQLite
- **Training and Technical Support**: Specialized trainer for health sector civil servants in KoBoToolbox (form design, structuring, visualization); pedagogical facilitator for QGIS cartography training sessions across multiple Malagasy regions

Technologies: *DHIS2, React.js, React Native, Node.js, Gmail API, Google Cloud Service Account, Electron.js, SQLite, PostgreSQL, KoBoToolbox, QGIS, Linux*

Data and Fullstack Software Engineer (Internship)

June 2022 – March 2023

JSI Research and Training Institute (USAID, PMI Measure Malaria) | Madagascar

- Provided technical assistance to project teams for activity implementation at peripheral levels according to project planning schedules
- Supported management and maintenance of COVID-19 vaccination data systems; facilitated user and stakeholder engagement workshops at central and peripheral levels
- Contributed to supervision activities and data regularization processes to ensure data integrity and compliance; participated in data entry processes and activity report capitalization
- Developed personal technical projects and conducted feasibility analysis for geolocation applications; undertook self-directed learning in AI and advanced data science alongside beginning the Master 2 program

Technologies: *DHIS2, React.js, Node.js, PostgreSQL, KoBoToolbox, QGIS*

TECHNICAL SKILLS

Cloud GCP	Vertex AI (Pipelines, Custom Training, Model Registry, Monitoring, Gemini, Model Garden), Cloud Run GPU (NVIDIA L4), BigQuery, BigQuery ML, Dataflow, Dataproc, GKE, Cloud Composer, Pub/Sub, GCS, Cloud Build, Cloud Monitoring, Cloud Scheduler, Artifact Registry
Cloud AWS	SageMaker, Bedrock, Glue, S3, RDS, Redshift, Lambda, ECS, MWAA, CloudFormation (IaC), Terraform
Speech AI	ASR: Whisper large-v3-turbo INT8 (fine-tunable), Whisper Small, Chirp (Vertex AI), Silero VAD v5 | NMT: NLLB-200-distilled-600M (LoRA fine-tuned, plt_Latn Malagasy) | TTS: MMS-TTS-MLG VITS (fine-tuned, 6 dialects), CosyVoice2-0.5B (SFT, zero-shot voice cloning) | Quantization: CTranslate2 INT8 | Metrics: WER, chrF++, UTMOS, BLEU, SacreBLEU
Computer Vision	Object Detection and Segmentation: YOLO V8 (fine-tuned), Segment Anything Model (SAM) | Image-Text Alignment: CLIP | Self-Supervised: MAE, I-JEPA | Backbone: ResNet, EfficientNet, ViT | Optimization: TensorRT | Libraries: OpenCV, TensorFlow, PyTorch | Tasks: inventory counting, OCR, geolocation mapping
Vision Language	Gemini 2 Multimodal Embedding (Vertex AI), GPT-4 Vision, Qwen3-VL, LLaVA | Tasks: document OCR, medical license verification, product recognition, multimodal search | Benchmarked CLIP vs Gemini 2 for production embedding selection
NLP / LLM	RAG (Vector, Graph, Hybrid, Agentic, Multimodal) | Multi-Agent: LangGraph, CrewAI, Google ADK, A2A, ACP, MCP | Frameworks: LangChain, LlamaIndex, Langflow, n8n, Dialogflow | Fine-Tuning: HuggingFace, PEFT, LoRA | Evaluation: RAGAS, Langfuse | Models: Claude API, Gemini, GPT, LLaMA, Qwen3, SQLCoder-7B-2 | Prompt Engineering, RLHF
ML / DL	Supervised: XGBoost, LightGBM, Random Forest, SVM, MLP | Clustering: K-Means, DBSCAN, Agglomerative | Time Series: ARIMA+, Prophet, LSTM | Deep Learning: Transformers, GNN, RNN, CNN | Reinforcement: Q-Learning, PPO, RLHF | AutoML: BigQuery ML, Google AutoML, SageMaker
Data Eng.	ETL/ELT: Apache Spark, Apache Beam, Google Dataflow, Dataproc, AWS Glue | Orchestration: Airflow, MWAA, Cloud Composer | CDC: Airbyte | Transformation: dbt | Vector Stores: BigQuery Vector Search, Pinecone, Elasticsearch, Neo4j (knowledge graphs)
Databases	PostgreSQL, MongoDB, SQLite, BigQuery, Amazon Redshift, DynamoDB, Amazon Neptune, Elasticsearch, Pinecone, Neo4j

Full-Stack	Backend: Python (FastAPI), Node.js, NestJS | Frontend: React TypeScript, React Native, Next.js 15 | APIs: REST, SOAP, GraphQL | Architecture: Microservices, Event-Driven, Pub-Sub, Serverless, GoF Design Patterns
Automation & CRM	Workflow Automation: Make (Integromat), Zapier, n8n | CRM: Zoho CRM, HubSpot, GoHighLevel | Productivity: Notion, Airtable | Payment Gateways: MCB Juice, Mvola API, Orange Money API
Health Data	DHIS2 (Java framework, Tomcat, API, custom forms, KPIs), KoBoToolbox, QGIS, Microsoft Access, COVID-19 data quality pipelines (Isolation Forest anomaly detection)
Automation & CRM	Workflow Automation: Make (Integromat), Zapier, n8n | CRM: Zoho CRM, HubSpot, GoHighLevel | No-Code / Low-Code: Notion, Airtable | Marketing Automation: GoHighLevel pipelines, email sequences, lead nurturing | Integration: REST API connectors, webhook orchestration, bidirectional CRM sync
MLOps/DevOps	Docker, Kubernetes, AWS ECS | CI/CD: GitHub Actions, GitLab CI/CD, Jenkins, Cloud Build, AWS CodePipeline, ArgoCD | IaC: Terraform, CloudFormation | Monitoring: Vertex AI Monitoring, Cloud Monitoring | Document parsing: Docling
Languages	Python, TypeScript/JavaScript, SQL, Java (DHIS2), Bash, C++
Management	Agile/Scrum, PRD and Specification writing, User Stories, Use Cases, Business Rules, Effort Estimation, Jira, Asana, Canvas, Gantt, UML, Draw.io, PlantUML

EDUCATION

Master of Science in Big Data and Artificial Intelligence

IT University & ESTIA France | Grade: 17.2/20 *Dec 2022 – Dec 2024*

Specializations: Big Data, AI, Machine Learning, Full-Stack Development

Bachelor's Degree in Computer Science, Software Engineering

IT University | Antananarivo, Madagascar

Dec 2019 – Dec 2022

Specializations: Full-Stack Engineering, Software Engineering, Web Development, Databases, AI

Scientific Baccalaureate, Series C

October 2019

Sainte Famille Mahamasina | Antananarivo, Madagascar

CERTIFICATIONS AND ACHIEVEMENTS

- **Advanced English C1**, EF-SET Certified
- **Python 3**, CodingGame Certified
- **English for Business and Entrepreneurship**, HP Life Certification
- **IndabaX Madagascar 2023, 3rd Place**: Built a text classification NLP pipeline for pathology prediction from clinical symptoms across 10 classes; achieved 94% accuracy using XGBoost, CatBoost, TF-IDF, Word2Vec, and spaCy preprocessing
- **DataTour 2025 (Pan-African Data Science Competition)**:
  - Content recommendation model for social media video content (~122M user interactions): content-based and collaborative filtering; diagnosed and resolved severe temporal data leakage via causal feature engineering; cosine similarity from embeddings
  - Default risk classification model trained on real bank lending data: credit scoring using historical borrower behavior; feature engineering on loan repayment patterns; evaluated with XGBoost and LightGBM

LANGUAGES

English: C1 Advanced (EF-SET) **French**: Native / Fluent **Malagasy**: Native