Igor Rendulic

Salt Lake City, Utah
igor@mail.io | https://github.com/igorrendulic | https://www.linkedin.com/in/igorrendulic

Machine Learning Engineer — Software Engineer (Python, Go, TypeScript)

Experienced in designing, training, and deploying **ML and LLM-based models** in production environments. Strong background in backend architecture (Go, Python, FastAPI, gRPC), **MLOps and model serving** (Docker, Kubernetes, CI/CD), and applied AI (NLP, embeddings, fraud detection). Passionate about **privacy-preserving AI, multi-modal systems**, and distributed architectures.

Skills

Languages: Python, Go, TypeScript, JavaScript, SQL, Bash
Machine Learning / AI: PyTorch, scikit-learn, XGBoost, Transformers (HuggingFace), LoRA / PEFT, NumPy, Pandas
MLOps / Data Infrastructure: MLflow, BigQuery, Pinecone, Vertex AI, Dataflow, NoSQL databases
Backend / Frameworks: FastAPI, Flask, gRPC, Node.js, Express, REST
Frontend / UI: React, Next.js, Material-UI, Material Design
Cloud / DevOps: GCP, AWS, DigitalOcean, Docker, Kubernetes, Terraform, GitHub Actions, CI/CD, Prometheus
Security / Cryptography: JOSE (JWS/JWE/JWK), NaCl, DIDComm v2, PKI, Secure Email Systems
Other Tools: Git, Linux, VS Code, Jupyter, Postman, Coding AI assistants (Cursor, Copilot)

Professional Experience

Founder / Principal Engineer Feb 2025 – Present
Mailio (personal project) mail.io

Founded and built **Mailio**, a privacy-first email platform reimagining how people communicate by putting users back in control of their data and interactions. Created to address the **trust, overload, and loss of agency** plaguing modern inboxes, Mailio empowers users to define their own communication rules, maintain functional ownership of their data, and prepare for a future where **AI-powered, screenless devices** handle personal communication. Integrated **end-to-end encryption**, **semantic understanding**, and **AI-driven fraud detection** to restore clarity, privacy, and personalization to everyday email.

Developed an intelligent **email classification and ranking engine** using **Python, Transformers (E5, MiniLM), and XGBoost** to detect phishing, impersonation, and spam. Fused embedding vectors with tabular features—such as domain reputation, message frequency, and sender behavior—to boost detection accuracy by over 10% compared to rule-based baselines.

Implemented automated **retraining and model-drift monitoring pipelines** using **FastAPI, Pandas, and scikit-learn**, incorporating **PSI-based drift detection** and **ROC-AUC tracking** to ensure continual reliability as adversarial tactics evolved.

Fine-tuned domain-specific embedding models with **LoRA/PEFT** to improve detection accuracy of **fraudulent intent within emails**. Integrated **client-side Transformers** (via **transformers.js** from Hugging Face) to perform on-device inference, preserving message privacy through full local encryption.

Architected a fault-tolerant, distributed backend built with **Go** (mail delivery, queuing, event dispatch), **Python** (ML inference), and **TypeScript** (Frontend). Deployed across **Kubernetes**, using **CouchDB + PouchDB** for per-user encrypted document stores and **Pinecone** for scalable semantic vector search (RAG)

Developed a secure front end in **React (Nx monorepo)** featuring **WebAuthn/Passkey authentication**, client-side encryption via **NaCl/JOSE**, and background PouchDB replication for seamless local-first performance.

Built event-driven data pipelines using **Google Pub/Sub** and **BigQuery** for structured telemetry, model retraining triggers, and behavioral analytics. Containerized all services with **Docker** and deployed on **Kubernetes** via automated **GitHub Actions CI/CD**, integrating **Hugging Face–hosted Transformer models** managed through private repositories for versioned deployment and inference. **Impact:** Delivered

a full end-to-end encrypted, AI-augmented email system capable of contextual search and real-time fraud prevention, demonstrating how privacy-preserving intelligence can coexist with convenience. Gained early adopter traction among privacy-conscious users.

Links:

- https://mail.io
- https://github.com/mailio

Principal AI Engineer San Francisco, CA
HaloVision Inc. Jan 2025 – Jul 2025

Led development of **HaloVision**, an **AI-driven confidential conversation analytics platform** built to help organizations surface hidden blockers, communication breakdowns, and sentiment trends across teams. The company needed a way to transform unstructured voice conversations into actionable intelligence while preserving privacy and context fidelity.
Designed and implemented voice analytics pipelines in **Python** using **FastAPI** and **Transformers** to perform **LLM-powered summarization**, **information extraction**, and **topic modeling** from transcripts. Combined **prompt-based reasoning** (OpenAI/Claude APIs) with **traditional ML techniques**—including **TF-IDF**, **BERTTopic**, **clustering**, and **paraphrase similarity scoring**—to improve factual grounding, reduce hallucinations, and enable **human-in-the-loop validation** for enterprise clients.
Integrated **RetellAI**, **Pipecat**, and **AssemblyAI** to handle real-time transcription, speaker diarization, and metadata enrichment. Implemented orchestration and performance monitoring with **LangFuse**, custom evaluation scripts, and **drift detection dashboards** for continuous model QA and conversation accuracy tracking.
Built the backend as containerized services with **Docker** and deployed via **Cloud Run** for scalability and rapid iteration. Integrated **LangFuse** for **LLM prompt management, tracing, and debugging**, enabling real-time visibility into model behavior, prompt performance, and conversational accuracy across large-scale deployments.
Delivered the full product lifecycle—from proof-of-concept to onboarding the first enterprise clients—in under three months.

Key Technologies: *Python, FastAPI, Transformers, LangFuse, RetellAI, Pipecat, AssemblyAI, Pandas, scikit-learn, BERTTopic, TF-IDF, clustering, paraphrase similarity scoring, hallucination mitigation, Docker, Cloud Run, Kubernetes, and human-in-the-loop evaluation systems.*

Link: https://halovision.us

Senior AI/ML Engineer Sept 2021 – Dec 2024
Azumio Inc. Food Lens

Built and deployed **Food Lens AI** at **Azumio**, a leading mobile health and wellness company whose biometric products—such as heart-rate monitoring and fitness-tracking apps—have helped millions of users take control of their health. Food Lens extended this mission by addressing a critical challenge: enabling people to accurately recognize foods from photos and receive instant nutritional feedback to make informed dietary decisions. The goal was to automate nutrition tracking for populations where it matters most—**diabetics managing blood sugar**, individuals seeking **weight loss or balanced nutrition**, and users needing to **avoid allergens or specific ingredients**. This innovation helped transform Azumio's suite of health tools into a more holistic, life-improving ecosystem by connecting biometric insight with intelligent, food-based analytics.

Designed, trained, and deployed a large-scale **transformer-based image classification system** in **Python**, **TensorFlow**, and **PyTorch**, combining a **CNN classifier** and specialized **food-segmentation model** to process tens of millions of food images. Achieved **88% top-5 accuracy** across diverse cuisines and lighting conditions, powering personalized diet tracking and nutrition analytics in Azumio's ecosystem.

Implemented a **transformer-based model** using **Hugging Face Transformers** and **OpenCV** to analyze packaged-goods labels—automatically identifying front, back, and barcode panels and detecting which labels contained nutritional facts. Integrated multiple **nutrition data sources** (**USDA FoodData Central**, **NutritionX**) to enrich product metadata and improve nutrient composition accuracy. Added **barcode scanning logic** directly in the training pipeline to pair product images with nutritional datasets, tightening image-to-label alignment and improving model grounding.

Applied **LLM prompting techniques** with domain-tuned language models to extract and validate additional nutritional details from packaging text, improving coverage and accuracy for unseen or regional products.

Engineered a distributed **data-ingestion and training pipeline** using **Google DataFlow** (GPU-enabled) and **Cloud Storage** to orchestrate the entire data life-cycle. Used **Pandas** and **BigQuery** to clean and de-duplicate unlabeled datasets, spun up transient Google DataFlow workers for embedding generation, and employed custom **image-classification micro-models** for label detection (front/back/barcode/nutrition panels). Automated further dataset curation on **Cloud Run** and applied **OpenCV** for image cleanup and normalization prior to model training.

Deployed scalable inference services on **Cloud Run** and **Kubernetes**, using latency-optimized **TensorFlow Serving** with request batching to support both mobile and enterprise API clients. This architecture enabled real-time inference on consumer devices while maintaining cloud scalability.

Integrated the technology into **Samsung Bixby (Galaxy S9+)** and licensed the **Food Recognition API** to enterprise partners, including **Helix** for DNA-based meal personalization—establishing Azumio as an early leader in AI-driven nutrition tracking.

Links:

- Research Paper: "An Exploratory Approach to Deriving Nutrition Information of Restaurant Food from Crowdsourced Food Images"
- MyDietCam: Development and Usability Study of a Food-Recognition Integrated Dietary Monitoring App
- Samsung Adds Food Image Recognition to Bixby via Calorie Mama API
- Azumio Adds DNA-Based Insights to Calorie Mama AI App

Director of Engineering Aug 2019 – Sept 2021
Chrysalis Cloud Chrysalis Cloud

Architected, designed, and developed all backend services for **Chrysalis Cloud**, a distributed **video-streaming and ML inference framework** built to make large-scale video ingestion and analysis **simple, fast, and up to 60× cheaper** than AWS Kinesis. The company's challenge was to reduce the high cost and operational complexity of ingesting and analyzing thousands of live video streams while maintaining real-time inference performance for developers and enterprise clients.
Built the core **video ingestion service** using **Go**, **gRPC**, and **Redis**, with a secure encrypted channel to deliver high-throughput camera streams to the cloud. Implemented the platform on top of **etcd** (a distributed, reliable key-value store) with a custom **orchestration layer**, using **IaC** techniques (**Terraform** and custom cloud integrations) to manage and optimize cluster resources—enabling up to **50K concurrent live streams** on a single cluster.
Integrated and implemented **Google IoT Core** to handle remote device management, telemetry, and over-the-air updates for all deployed edge devices. Built a centralized control plane capable of deploying

new inference modules by remotely updating **Docker containers**, orchestrating multiple concurrent video streams per edge node, and ensuring encrypted, low-latency communication between the edge and cloud.

This hybrid design allowed both edge and cloud inference, minimizing bandwidth usage while preserving low latency and high reliability.
Implemented real-time telemetry pipelines using **Google Pub/Sub**, **Redis**, and **Prometheus/Grafana** to monitor system health, stream performance, and camera diagnostics.
Developed a lightweight custom backend monitoring interface—similar in concept to **NVIDIA Fleet Command**—that visualized cluster activity and helped operators detect anomalies before service degradation occurred.

Links:

- https://www.crunchbase.com/organization/chrysalis-cloud
- https://github.com/igorrendulic/video-edge-ai-proxy
- https://www.youtube.com/watch?v=bWCgTW2Ar5s

Senior Backend Engineer Aug 2018 – Aug 2019
Wearless Tech Inc. Cocooncam

Served as **Lead Backend Engineer** for **CocoonCam**, an **AI-powered baby monitor** platform that used computer vision for real-time motion and breathing detection. The company's challenge was to process and analyze tens of thousands of concurrent live camera streams in real time while keeping latency low and cost manageable.
Architected and scaled the **video ingestion and ML inference backend** using **Go**, **Python**, and **gRPC**, enabling the system to process over **15 PB of video data monthly**. Built backend services for **camera setup, provisioning, and installation**, allowing seamless device onboarding and automatic configuration of edge connectivity.
Implemented the video ingestion pipeline using **Redis** as a distributed in-memory buffer, ensuring fast, reliable streaming from client devices to the cloud with the ability to **rewind recent video segments** and maintain inference continuity during network fluctuations.
Enhanced user engagement by extending the pipeline to extract and store **"cute moment" insights**—short, meaningful video clips automatically surfaced for parents from overnight baby activity streams, showcasing the emotional value of the product.

Designed and deployed a **custom monitoring and alerting framework** integrating **Prometheus** and **Grafana** to maintain global uptime and performance. Optimized latency through a **multi-region cluster architecture** that minimized round-trip time between cameras and cloud inference nodes using **distance-aware load balancing**.
Collaborated with ML and mobile teams to deploy real-time **ML inference models** using **OpenCV** and **TensorFlow** for non-contact breathing detection directly from live video feeds. The platform's technical and commercial success led to its **acquisition by Alarm.com**, generating millions in recurring revenue and achieving strong customer retention.

Links:

- CNBC Review
- YouTube Videos

Senior Backend Engineer Feb 2014 – Aug 2018
Azumio Inc. Health Wellness AI Platform

At Azumio, I architected and scaled backend systems powering AI-driven health, fitness, and nutrition applications—used by tens of millions of users worldwide. I've designed and maintained high-throughput NoSQL database (Google Datastore) for storing biometric and nutrition data at scale. Integrated these with Google BigQuery for analytics, cohort analysis, and user behavior insights across hundreds of millions of records. Optimized performance and reliability, built caching and batch-processing systems to handle 10K+ requests per second with sub-100 ms latency while ensuring data consistency and uptime across global regions. Ensured data privacy and security compliance; Applied encryption at rest/in transit and anonymization techniques to meet GDPR- and HIPAA-aligned data handling standards.
Links:

- https://pmc.ncbi.nlm.nih.gov/articles/PMC6592896/
- https://www.researchgate.net/figure/Azumio-Instant-Heart-Rate-capture-screen$_{fig}$1$_{3}$05663641
- https://covidresearch.ucsf.edu/projects/analysis-azumio-stepcount-heart-rate-and-deidentified-data-covid-19
- Best practices for analyzing large-scale health data from wearables and smartphone apps

R&D Engineer (formerly Junior Software Engineer) Ljubljana, Slovenia
SRC Infonet (formerly InfoNET; acquired by SRC) 2008 – 2012

- Proposed and prototyped a **proxy-based HL7 document exchange platform** (*medGateway*); the solution was productized and deployed within Slovenia's national e-health ecosystem (integrations with ePrescriptions and CRPP) leading to my transition into a dedicated **R&D** role.

Education

Master's in Computer Science 2005
University of Maribor, Slovenia, Universite d'Avignon, France
Specialization: Information Extraction and Envelope Induction

Other Publications & Projects 2006 – 2014

- **Phaidra -** Advanced automation of academic library system and University repository system
- **Motor Vehicle Registry of Slovenia** Worked with Ministry of the Interior as a developer
- **Pfizer -** mobile application for Pfizer's field sales representatives to support biopharmaceutical product promotion and doctor engagement
- **Stanford researchers find intriguing clues about obesity by counting steps via smartphones** Staford report
- **Countrywide natural experiment links built environment to physical activity** Nature Article
- **Comparing Two Commercially Available Diabetes Apps to Explore Challenges in User Engagement: Randomized Controlled Feasibility Study** Research paper
- **Usability Evaluation of Four Top-Rated Commercially Available Diabetes Apps for Adults With Type 2 Diabetes** Researh paper