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 integrating **end-to-end encryption** with **client-side embedding generation**, **cloud-based semantic search**, and **AI-driven fraud detection**.
- Designed, implemented and deployed **Transformer-based models** (E5, MiniLM) for email intent classification, similarity search, and contextual ranking; combined embedding features with tabular metadata in an **XGBoost ensemble** to detect spam, phishing, impersonation, and spear-phishing attacks.
- Implemented **continuous retraining and drift monitoring** (PSI, ROC-AUC tracking), integrated automated evaluation pipelines, and fine-tuned models with **LoRA/PEFT** for domain adaptation.
- Architected a distributed backend in **Go, Python, and TypeScript** with **Kubernetes**, **CouchDB**, and **Pinecone**, supporting real-time synchronization and per-user encrypted data stores.
- Developed the front-end in **React (Nx monorepo)** with secure **WebAuthn/Passkey authentication**, **PouchDB replication**, and privacy-preserving local processing.
- Built event-driven pipelines using **FastAPI**, **Pub/Sub**, and **BigQuery** for logging, model retraining, and behavioral analytics.

Other technologies: *Transformers, LoRA/PEFT, LLM prompting, XGBoost, FastAPI, Kubernetes, Docker, Pinecone, Pandas, scikit-learn, BigQuery, Pub/Sub, WebAuthn/Passkeys, encryption (NaCl, JOSE), CloudRun, CI/CD (GitHub Actions).* **Links:**

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

Principal Engineer — AI & Machine Learning Systems San Francisco
Azumio, CocoonCam, Chrysalis Cloud, HaloVision Jan 2014 – Jul 2025

- Led development of **HaloVision**, an **AI-driven confidential conversation analytics platform** designed to uncover hidden blockers, risk patterns, and sentiment trends within organizations. Architected the system to process voice transcripts using **LLM-powered summarization**, **information extraction**, and **topic modeling** pipelines. Combined **prompt-based reasoning** with **traditional ML techniques** (TF-IDF, clustering, paraphrase similarity scoring, and keyword co-occurrence graphs) to improve factuality, reduce hallucinations, and enable **human-in-the-loop validation**. Integrated **RetellAI**, **Pipecat**, and **AssemblyAI** for real-time transcription and metadata enrichment; deployed orchestration and monitoring through **LangFuse** and custom evaluation scripts for conversational accuracy and drift detection. Delivered the full product lifecycle—from prototype to onboarding the first enterprise clients—in under three months.

  Other Technologies: *Python, FastAPI, Transformers, Pandas, scikit-learn, LangFuse, RetellAI, Pipecat, AssemblyAI, LLM-based conversational AI, information extraction, topic modeling (BertTopic), clustering, paraphrase similarity, hallucination mitigation, human-in-the-loop review, and analytics pipelines for large-scale conversational data.*

  Links:

  - https://halovision.us

- Architected, designed and developed all backend services for **Chrysalis Cloud**, a distributed **video-streaming and ML inference framework** derived from CocoonCam's production infrastructure. Enabled developers and enterprises to stream video securely to the cloud for **real-time AI analysis** using modular, pluggable inference pipelines (object detection, motion tracking, pose estimation, facial recognition, and more). Designed a **containerized edge-to-cloud architecture** supporting on-device preprocessing and adaptive bitrate streaming, enabling both on-edge and cloud-based inference.

  Integrated **Google IoT Core** for comprehensive device management, remote monitoring, and over-the-air software updates. Implemented a centralized control plane that could deploy and upgrade edge devices by remotely updating **Docker images**, orchestrating multiple concurrent video streams per edge device, and ensuring secure communication channels between edge nodes and the cloud. Built real-time telemetry pipelines to collect and visualize system health, performance metrics, and camera diagnostics—enabling proactive maintenance, anomaly detection, and live monitoring of distributed inference workloads.

  Achieved up to **60× cost efficiency** through a **bagging-inspired orchestration algorithm** that dynamically balanced workloads across cloud nodes processing continuous real-time video streams. Deployed across **Kubernetes** clusters with monitoring via **Prometheus**, **Grafana** and custom monitoring UI (similar to NVIDIA Feet Command, but simplified and adopted to Chrysalid Cloud needs).

  Technologies: Go, Python, Docker, Kubernetes, gRPC, Google IoT Core, Google Pub/Sub, Redis, OpenCV, PyTorch, Prometheus, Grafana, real-time video analytics, IaC (infrastructure as Code), remote device management, and a bagging-like orchestration algorithm for managing distributed multi-stream edge inference.

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

- Served as **Lead Backend Engineer** for **CocoonCam**, an **AI-powered baby monitor** platform leveraging computer vision for real-time motion and breathing detection. Architected and scaled the **video ingestion, processing, and ML inference backend**, handling over **15 PB of video data monthly** across tens of thousands of concurrent live camera streams. Implemented the backend services responsible for **camera setup, provisioning, and installation**, enabling seamless onboarding of new devices and automatic configuration of edge connectivity. Integrated the core backend with **etcd**, a distributed and highly reliable key-value store, and built a custom **orchestration layer** on top of it for managing and monitoring over **50K simultaneous live streams on a single cluster**.

  Leveraged a distributed in-memory database (**Redis**) to enable fast, reliable, and efficient video streaming from client devices to the cloud, with the capability to **rewind recent video segments** and ensure seamless continuity and responsiveness of cloud-based ML inference. Extended the pipeline to extract and store **"cute moment" insights**—short, meaningful video clips automatically surfaced for parents from overnight baby activity streams.

  Designed and deployed a **custom monitoring and alerting framework** integrating **Prometheus** and **Grafana** to ensure uptime and performance across global deployments, with multi-region cluster design optimized for **distance-aware latency reduction**—minimizing round-trip time between cameras and cloud inference services.

  Collaborated with ML and mobile teams to deploy **ML inference models** 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 sustaining high user retention.

  Links:

  - CNBC Review
  - YouTube Videos

  Other Technologies: Go, Python, , OpenCV, etcd, Redis, gRPC, Docker, Kubernetes, Prometheus, Grafana, libav (underlying libraries that powers FFmpeg's core functionality), distributed orchestration, and large-scale real-time video analytics.

- Built and deployed **Calorie Mama** at **Azumio**, a leading mobile health and wellness company, developing a deep-learning **food-image recognition and nutrition-analytics platform** featuring over **1M+ food and barcode classes** with rich nutritional metadata.

  Designed, trained, implemented and deployed a large-scale **transformer-based image classification model** on tens of millions of unlabeled food images, operating in conjunction with a **convolutional image classifier and specialized food segmentation model** to achieve **88% top-5 accuracy** across diverse cuisines and lighting conditions.

  Implemented a **transformer-based image classification model** to analyze packaged goods labels—automatically distinguishing front, back, and side panels, and identifying which labels contained barcodes or nutritional facts. Integrated multiple **nutritional data sources**, including **USDA FoodData Central** and **NutritionX**, to enrich product metadata and support accurate nutrient composition analysis. Implemented **barcode scanning within the training pipeline** to pair product images with corresponding nutritional datasets, improving labeling precision and model grounding.

Applied **LLM prompting techniques** to extract and validate additional nutritional information directly from packaging text, improving data completeness and accuracy for unseen products.

Developed a distributed **data ingestion and training pipeline** using **Google DataFlow** and **Cloud Storage**, automating dataset expansion, augmentation, and continual retraining for new foods and regions. Deployed scalable model inference services using **Cloud Run** and **Kubernetes**, with optimized batching and latency-tuned TensorFlow Serving for mobile and API clients.

Helped integrate the technology into **Samsung Bixby (Galaxy S9+)** which as licensed the **Food Recognition API** to enterprise partners.

Links:

- (An Exploratory Approach to Deriving Nutrition Information of Restaurant Food from Crowd-sourced Food Images: Case of Hartford)
- MyDietCam: Development and usability study of a food recognition integrated dietary monitoring smartphone application
- Samsung Adds Food Image Recognition To Bixby Through Calorie Mama API
- Azumio adds DNA-based insights into its Calorie Mama AI app

Other Technologies: Python, Pandas, scikit-learn, TensorFlow, PyTorch, Transformers (Huggingface), Google DataFlow, Cloud Run, Cloud Storage, Kubernetes, REST APIs, Image Classification, Large-Scale Training Pipelines, Data Augmentation, and Model Optimization for Mobile Deployment, LLM prompting

- **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
Specialization: Information Extraction and Envelope Induction

Selected Publications & Projects

- **Mailio:** Privacy-first, end-to-end encrypted email platform integrating **LLM-powered semantic search** and **AI-based fraud detection**. Built distributed backend (Go, Python, TypeScript) with Kubernetes, CouchDB, and Pinecone for scalable, privacy-preserving communication. (Website) (GitHub)
- **Chrysalis Cloud:** Distributed ML video inference platform handling **15 PB+ monthly** with **60× cost efficiency** over AWS Kinesis. (GitHub) (Demo Video)

- **Calorie Mama:** Deep-learning food image recognition and nutrition analytics system (**88% top-5 accuracy**); integrated into Samsung Bixby and Helix DNA personalization platform. (Research Paper)