# Mohammad Ali Yektaie

## Software & AI/ML Architect

ma_yektaie@outlook.com
linkedin.com/in/ali-yektaie

Experienced engineer adept at Large-Scale Software and AI/ML platforms. Skilled in cost-effective enterprise-level solutions, team leadership, and distributed infrastructure. Proven track record in navigating complex use cases, delivering innovative solutions, and optimizing processes for efficiency and scalability.

## Career Highlights

- Delivered **$1.5B** in **cost reductions** and **revenue gains** through strategic engineering and optimization initiatives.
- Expert in designing, building, and operating **scalable, mission-critical distributed systems.**
- Proven track record of owning and driving end-to-end **delivery** of **high-impact systems**.
- Strong ability to **anticipate risks**, **model scenarios**, and **quantify trade-offs** to guide cross team/organization decisions.
- Skilled at **navigating complex team dynamics**, aligning cross-functional stakeholders, and **driving consensus**.
- Deep foundation in **mathematics** and **statistics**, applying **quantitative methods** to solve real-world engineering challenges.
- Experienced in **AI/ML** systems, including **feature engineering**, **model training**, **evaluation**, and deployment in both **online** and **offline** environments and across different modalities (text, audio, image, video and tabular data).
- Designed and maintained robust **CI/CD pipelines**, significantly reducing defect rates and **improving production stability**.

## Software Engineering Experiences

**Sony Interactive Entertainment**
Staff Software Engineer – Tech Lead of Online Inference
Jan 2023 – Present

Led the online inference team, partnering with AI/ML and product organizations to operate and scale online use case at PlayStation.

- Reduced E2E inference latency to 30ms (**7x reduction**, 200ms SLA), launching **10x** more experiment and contributing to **$450M** incremental revenue. Service had **35-45K RPS**.
- Led initiative to eliminate long standing **fan-out challenges**, **stabilizing** system latency, improving **user experience**, and cutting operational **cost by 20x**.
- Led the redesign of ML model serving using **Seldon** and **Open Inference Protocol**, improving system **stability** and streamlining the MLE experience.
- Partnered with **fraud detection** teams to scale **tampering** detection and **trophy-cheat** detection online services.
- Collaborated with **trust & safety** teams to scale bullying and inappropriate-communication detection for PlayStation chat.
- Worked across **infrastructure** and **security** teams to deliver **CI integration testing** without exposing the Docker root socket, improving **security posture** and **build reliability**.
- Removed **chaotic** and **time-variant caching behavior**, enabling QA to run load tests that reliably mirrored production.
- Owned and operated the **Model Serving Gateway** (MSG), supporting personalization, fraud detection, and payment optimization use cases across the organization.
- Collaborating with other principal/staff/senior level engineers on **strategic roadmap definition** and mentoring younger engineers to make impactful decisions.

## Selected AI / ML Achievements

### AdTech Metric Normalization & Closed-Loop Optimization

Built ML-driven pipelines to **normalize and resample Apple's impression/share metrics** for **AdTech** bidding. Created a closed-loop system connecting streaming events → ML transformations → bidding decisions impacting environment. Pipeline reduced cost by **$690M** while maintaining the same **KPIs**.

### PlayStation Store Personalization – Advanced Modeling Integration

Designed and productionized scalable mechanisms to incorporate bundle-fatigue features, **discreet RNN events processing**, and **multi-armed bandit** reinforcement learning into PlayStation's personalization stack. Worked closely with modeling teams to ensure functional **correctness**, **performance**, and end-to-end **scalability**.

### Fraud & Abuse Detection at Scale

Drove modeling and infrastructure for **tampering** and **trophy-cheat detection**, from deriving statistical features to designing **distributed data plumbing** and inference workflows. Delivered high-throughput detection systems used across PlayStation's **safety ecosystem**.

### Medical Image Segmentation & Cranial Malformation Detection

**Customized U-Net** architecture and **geometric** post-processing pipeline capable of detecting **11 cranial malformations** with specialist-level accuracy. Delivered **real-time** mobile inference and supported clinical workflows that directly improved pediatric outcomes.

**Uber**
Senior Software / Machine Learning Engineer at AdTech Platform
Jan to Dec 2022

Led optimization and scalability improvements for feature-generation pipelines supporting Uber's **AdTech systems**, especially related to the Apple channel.

- Designed and delivered the integration of **Apple's impression count and share metrics** into Uber's bidding architecture using **Kafka** Streams and **Piper** (Uber's **Airflow** equivalent).
- Built data transformations to estimate trends, **align granularities**, and produce bidder-compatible signals from Apple's **raw metrics**.
- Launched the full end-to-end pipeline to production, enabling the Apple ad channel to achieve **$690M in cost reduction (~27%) while maintaining same KPIs.**
- Improved performance of large-scale **HIVE** jobs—processing **few terabytes** of raw data daily—by restructuring **joins**, eliminating **inner queries**, and simplifying query logic, resulting in a **3x reduction** in compute cost and more **reliable execution**.

**PediaMetrix**
Unofficial Chief Technical Officer (CTO)
2019 – 2022

**Automated a fully manual clinical workflow** by building end-to-end pipelines that processed input images, calculated medical measurements by **image segmentation** complemented by **geometrical algorithms**, produced clinician-ready reports, and notified pediatricians for review and approval.

- Led platform architecture by **authoring core design documents** and defining API specifications for internal and external integration and **led a small team** of 6 engineers to deliver the requirements.
- Drove the design and implementation of two **flagship products**—SoftSpot, for cranial malformation detection, and a system for head-circumference **measurement from mobile imagery**.
- Designed **3D** STEP-file visualization tools to support **dataset creation** and generate high-quality ground-truth **annotations for model training**.
- Delivered **real-time image-segmentation** performance on **mobile devices**, achieving 60 FPS on iPhone and 19–33 FPS on Android, and built **backend pipelines** to process clinical report images at scale.
- **Customized a U-Net**-based segmentation architecture for the company's data and led data science efforts to improve model accuracy while developing a targeted, **cost-efficient strategy** for dataset collection.

# Technical Skills

### Distributed & Scalable Systems

Cloud Providers (AWS mainly), **Kubernetes** & ecosystem, Monitoring with **Grafana**, Kafka & message queuing, Hadoop ecosystem (Spark, Hive & HDFS), containerization, Caching strategy and scaling, especially with **Redis** (community and enterprise versions) & **Aerospike**, Autoscaling with **KEDA**, **Synchronization** & Orchestration, Network Protocols (**GRPC**, REST, TCP, common networking troubleshooting tools), Hypervisors (**Proxmox**, ESXi), …

### Best Practices, Security & Compliance

**Bazel** build system**, CI/CD** with **Gitlab**, **Prow**, **GitHub** Actions, **GitOps** at scale and observability, **Infrastructure as code** (Pulumi mainly), DevOps, **GDPR** data compliance, Production access **hardening**, …

### Theoretical Knowledge

Theory of Computation, Formal Languages, Algorithms & Data Structures, Statistics, Calculus, Signals and Systems, …

### Languages

Java, Python, C#, SQL (different flavors), …

### Miscellaneous

Web Development (NextJS), UI Design, Mobile Development with Native & Cross Platform tools, MVC, Low Level 3D and GPU computing, Deep Learning with PyTorch and Keras, Time series processing, JetBrains IDEs …

# Soft Skills

Team Management | Communication | Creativity | Leading Projects | Debugging | Problem Solving

# Education

Master's in Computer Science
**Georgetown University, Washington DC**

Focus on Machine Learning and its applications in Natural Language Processing (NLP) and Signal Processing

**YekiSoft**
Founder
2015 – 2018

Designed and developed iOS and Android applications for foreign-language learning, leveraging spaced-repetition systems (SRS) to improve retention and engagement.

- Built an SEO-optimized online education platform to support content discovery and user acquisition.
- Successfully delivered and maintained 20+ mobile applications, covering both content and learning-tool use cases.
- Deployed machine learning models for user-behavior prediction and personalized content recommendation, improving learning effectiveness and retention.
- Led marketing efforts that attracted 12,000+ users in Iran with a 7% conversion rate, contributing directly to product growth.

**Turned On Digital**
Senior Software Engineer
2014 – 2015

Led the refactoring and modernization of Sibche, an App Store alternative for the Iranian market, serving 400,000+ active users.

Directed the design and development of Kamanche, a full-featured iTunes replacement tailored to regional constraints and user needs.

Contributed to the product design and technical implementation of Fruitcraft, a role-playing mobile game that became viral and reached 1M+ users across Iran.

**Arad Sarmayeh**
Software Engineer
2012 – 2017

Independently designed and built Arad Trader, a full-featured technical analysis and trading platform for Iran's domestic stock market.

Implemented dual Shamsi/Gregorian calendar support, native-language UI, and integrations with local brokerage APIs to enable real-time trading workflows.

Delivered a product that combined the usability and indicators of MetaTrader, the advanced charting capabilities of MetaStock (e.g., XO and Kagi charts), and the Elliott Wave toolset found in Dynamic Trader—fully tailored to regional market requirements.

**HATEF Banking Co.**
Software & Hardware Engineer
2008 – 2013

Progressed from Junior Developer to Head of Software Division within two years, leading a team of 10 engineers in building software controllers for the company's ATM product line.

- Collaborated closely with the hardware division to diagnose and resolve system-level issues across deployed ATM fleets.
- Designed the architectures for multiple strategic software products, setting technical direction and development standards for the organization.
- Developed an ATM controller supporting major industry protocols, including Aptra NDC, NDC 6.0, and WOSA/XFS.
- Significantly reduced development cost and complexity by interfacing native C++ libraries with C# and Java, enabling cross-language reuse of core low-level components.
- Built comprehensive device simulators to streamline development and enable robust end-to-end testing, improving product reliability and developer productivity.