Samyak Shah

Senior ML Engineer | Production ML, Modeling, Real-Time Inference, LLM Systems
Dallas, TX • samyakmehulshah@gmail.com • (410) 330-0487 • linkedin.com/in/samyak-shah-68410561

Skills

Modeling: Python, scikit-learn, PyTorch, TensorFlow, XGBoost, random forest, isolation forest, SVM, PCA, clustering, HMMs, MCMC, probabilistic modeling, adaptive deep learning

ML Systems: MLflow, DVC, DAGsHub, Spark, ClickHouse, model evaluation, model monitoring, drift detection, experiment tracking, real-time inference, MLOps

LLM / Applied AI: RAG, prompt orchestration, fine-tuning, LangChain, NeMo Guardrails, context-aware retrieval, citation-grounded generation, Whisper

Backend / Cloud: AWS, AWS CDK, FastAPI, Flask, Spring Boot, PostgreSQL, MongoDB, REST APIs, CI/CD, Bitbucket Actions, Fastlane

Frontend / Product: TypeScript, Next.js, React, Swift, Xcode, Objective-C

Experience

Co-Founder, Principal ML Engineer *Aug 1, 2017 – present*
EpiWatch
Co-founded EpiWatch as a Johns Hopkins spinout during my graduate research and continued development alongside roles at Johns Hopkins and Orba, serving as the primary technical owner for ML, product, and regulatory work.

- Developed a novel real-time seizure detection algorithm using only sensors available on consumer wearables, achieving 10x better performance than the prior state of the art. Algorithm has been validated in a large scale clinical trial and received FDA clearance.
- Designed a dynamic patient-adaptive focal seizure detection approach using probabilistic modeling, latent-variable methods, and adaptive deep learning. Conceived, designed, and wrote the $300,000 NIH STTR grant that funded this program.
- Led the technical work for successful 510(k) application for the EpiWatch Seizure Monitor, including all core engineering documentation (SRS, SDS, Design Controls, Risk/FMEA, SDLC) and the initial QMS/SOPs, completing this in ∼2 months and helping deliver the program 60% faster than planned on under 50% of the budget.
- Designed and prototyped the deployment strategy for a per-patient adaptive focal seizure detection system intended for FDA-regulated use, including a code-promotion architecture in place of direct model promotion to reduce release risk and improve auditability once deployed.
- Built monitoring and incident-response workflows for a frozen, on-device seizure detection algorithm, tracking input-data drift and seizure-to-detection latency via replay tests and catching a hardware-induced latency regression before release, preventing deployment to hundreds of devices.
- Developed a distributed Spark-based training and simulation framework for seizure detection models, parallelizing multiple training / simulation jobs and reducing experiment turnaround time from 1 month to 10 days.
- Established reproducible ML workflows with DVC, MLflow, and DAGsHub, supporting 200+ tracked experiments and reducing time to compare model variants from days to hours.
- Built seizure detection models using classical ML approaches including isolation forest, random forest, XGBoost, SVM, PCA, and clustering, evaluating multiple modeling approaches with ROC/AUC, F1, and leave-one-user-out cross validation to select robust candidates.
- Implemented and optimized seizure detection algorithms on Apple Watch for periodic real-time alerting, keeping CPU and memory usage low enough to maintain over 10 hours of battery life, even on older devices.
- Served as technical lead for a succesful 2-year, 6-hospital multi-center clinical trial validating seizure detection algorithm efficacy, and built a Python/SQL reporting system that maintained trial blinding while automating compliance checks and surfacing data quality issues in near real-time.
- Developed the statistical analysis plan for clinical trial alongside a biostatistician to ensure unbiased evaluation of seizure detection performance, including binomial proportion methods for paired sensitivity analysis.
- Designed the architecture, resource plan, and budget for migrating seizure-sensor data to a ClickHouse time-series database, improving raw physiological data query times by over an order of magnitude in proof-of-concept tests on ∼1 TB of data and informing the in-progress full migration.
- Led full-stack development of the EpiWatch product from the ground up, including iOS and WatchOS frontend (Swift, Objective-C, Xcode), backend services (Java, Spring Boot), and cloud infrastructure (AWS, AWS CDK).

- Developed multi-developer SOPs and CI/CD workflows across the application stack, including Fastlane and Bitbucket Actions, reducing release overhead by 40%.
- MedTech Innovator accelerator participant (4% acceptance rate) — worked with digital health leaders including DexCom.

Co-Founder, Applied ML Engineer *Aug 1, 2022 – June 1, 2024*
Orba

- Built production LLM and agentic workflows for customer-facing automation, including RAG pipelines with citation support and context-aware retrieval over CRM and internal data sources, improving answer relevance by ∼30% and reducing hallucinations by ∼50%.
- Implemented NVIDIA NeMo Guardrails to enforce safety and formatting policies, improving response consistency and reducing obvious failure modes in production LLM workflows.
- Fine-tuned GPT-3.5 and designed prompt-chaining workflows for production use with LangChain-based orchestration, improving task completion or response quality by 15%.
- Engineered a self-hosted real-time AI voice system with sub-200ms latency, integrating Whisper transcription, context-aware retrieval, and response generation for customer support use cases.
- Developed backend and application infrastructure using Python, FastAPI, AWS, PostgreSQL, and MongoDB, including a distributed web-scraping system to process large customer websites.
- Launched as #2 Product of the Day on Product Hunt, accepted into Jason Calacanis' Founder University, and featured in popular industry newsletters.

Senior Software Engineer *June 1, 2019 – June 1, 2024*
Johns Hopkins University

- Implemented Agile methodology for the research group, improving team efficiency by 30% and streamlining project workflows.
- Developed and evaluated real-time brain-signal decoding models for cursor control using HMM and deep learning, implemented in a Kedro-based pipeline
- Applied Bayesian and probabilistic methods, including MCMC, to seizure detection models, using posterior analysis to understand parameter uncertainty and improve model robustness across multiple datasets.
- Led development of an LLM-based brain–computer interface application with long-term memory that enabled a brain-implant user to communicate intent and hold real-time conversations, in collaboration with Johns Hopkins APL and presented at SfN 2023.
- Developed a production-ready modeling and real-time inference pipeline with sub-70 ms end-to-end latency using Python and C++, enabling scalable experimentation across multiple decoding pipelines and integrating with user interfaces.
- Built user-facing React/Next.js applications for neural decoding and assistive communication, including interfaces used by a brain-implant user to test speech and motor decoding models and support ALS patient interaction and control.
- Built a real-time WebRTC-based audio visualization tool for voice synthesis modeling in Next.js and C++, providing low-latency feedback for end users and serving as the primary interface for training a real-time speech synthesis model.
- Published 8 peer-reviewed papers with 256 citations and presented research at Society for Neuroscience 2023.

Education

Johns Hopkins University *Sept 1, 2022 – May 1, 2024*
Coursework Toward MS in Computer Science and Data Science

Johns Hopkins University *Aug 1, 2017 – June 1, 2019*
MS in Biomedical Engineering (Data Science Concentration)

University of Glasgow *Aug 1, 2013 – June 1, 2017*
BS in Biomedical Engineering

Selected Highlights

- Inventor (royalty-bearing), patent published — Methods and systems for physiological detection and alerting.
- Co-inventor, patent pending — Neurally augmented virtual interface.
- Primary technical author and key personnel on a $300,000 NIH STTR grant for adaptive wearable seizure detection, secured in 2 weeks (75% faster than expected).
- Speaker — Society for Neuroscience (SfN) 2023 (audience of 200+).