

Professional Summary

Applied machine Learning Engineer with 3+ years of experience developing and productionizing machine learning, generative AI, and data-intensive systems. Experienced in building knowledge graphs, distributed model-serving platforms, RAG applications, computer vision pipelines, recommendation engines, forecasting systems, and agentic tool-calling integrations for enterprise environments.

Technical Skills

Languages & Frameworks: Python, SQL, C++, Bash, PyTorch, TensorFlow, scikit-learn, XGBoost, LightGBM, FastAPI
GenAI & Model Serving: Transformers, vLLM, RAG, MCP, tool-calling, PEFT, LoRA/QLoRA
MLOps, Data & Cloud: MLflow, Docker, Kubernetes, GitLab CI/CD, Airflow, Spark, Hadoop, ETL, PostgreSQL, AWS, Linux

Professional Experience

Prime Healthcare | Data Scientist, AI/ML Los Angeles, CA | Feb 2024 – Present

- Built and scaled **PrimeSpot AI**, an enterprise AI platform serving **2K+ DAU**, integrating LLMs, retrieval systems, database tools, and internal workflows while supporting continuous feature releases using **Azure DevOps CI/CD**.
- Architected a **Kubernetes-based vLLM inference platform** for foundation and QLoRA-adapted models, optimizing batching and concurrency to sustain **500+ aggregate tokens/sec** under concurrent workloads.
- Built a secure **MCP tool-calling layer** for LLM agents to query production databases; implemented authentication, query guardrails, validation, rate limits, and audit logging across **10+ workflows**, reducing manual data lookup by **70%+**.
- Developed a **contract-analysis RAG pipeline** ingesting documents from **NFS and enterprise databases** into **pgvector**; used **BGE embeddings and reranking** to cut average contract search time from **20 minutes to 8 seconds**.
- Productionized a **multimodal invoice-processing pipeline** using vision-language OCR, structured extraction, schema validation, and confidence-based review routing, reducing manual indexing by **80%+**, replacing an external vendor, saving **\$1M annually**.

CarmaCam | Software Engineer - Machine Learning Los Angeles, CA | Aug 2023 – Feb 2024

- Built real-time dashcam vision pipelines using **YOLOv8** for unsafe-driving detection and **ResNet** for road-sign classification; optimized with **TensorRT** to achieve **sub-100 ms latency, 92% precision**.
- Developed a **FastAPI and WebSocket** alerting service to deliver inference results to emergency-response workflows, reducing response time by **30%**.

USC ITS - Office of the CISO | Data Scientist Los Angeles, CA | Feb 2022 – May 2023

- Redesigned a vendor-risk prediction pipeline using **XGBoost** and feature engineering across **28K vendors**; achieved an **F1 score of 0.91** and reduced false positives by **15%** through threshold optimization and holdout validation.
- Productionized an AWS **batch-inference and retraining pipeline** using **S3, ECR**, and feature-drift monitoring, automatically promoting validated model updates and reducing model-refresh time by **30%**.

VIT, Research Lab | Data Science Research Intern Vellore, India | Nov 2020 – Jul 2021

- Developed a **U-Net-based medical imaging model** for COVID-19 and pneumonia detection from chest X-rays; doubled training throughput through model pruning and batch-processing optimization while maintaining **low FLOPs and SOTA efficiency**.
- Optimized a **MicroNet-M3 segmentation architecture**, achieving a **98.9% Dice score** and **0.992 AUC-ROC** with **2x faster inference** and reduced computational cost compared with baseline models.

Arista Networks | Machine Learning Intern Mumbai, India | May 2019 – Jun 2019

- Developed an RF-fingerprinting indoor-localization system using **k-nearest neighbors and Random Forest regression** on Wi-Fi RSSI and inertial-sensor data, achieving a **2–3 meter median localization error**.
- Built signal calibration pipelines to reduce RSSI noise and multipath effects, improving positioning accuracy by **18%**.

Education

University of Southern California	Los Angeles, CA
Master of Science in Machine Learning and Data Science GPA: 3.81/4.00	Aug 2021 – May 2023
Vellore Institute of Technology	Vellore, India
Bachelor of Technology in Electronics and Communication Engineering GPA: 8.95/10.00	Jul 2017 – Jul 2021

Selected Projects

RelayAI Local Speech-To-Text, ASR, wav2vec, Python, FastAPI, Speech Models, LLMs	GitHub 
American Sign Language Detection PyTorch, NumPy, Matplotlib, Computer Vision	GitHub 
Spotify Song Recommendation Engine Python, sklearn, TensorFlow, Keras, NLP	GitHub 
Trojan Map C++, OpenCV, Data Structures & Algorithms	GitHub 