

VINAY JOGANI

joganivinay@gmail.com | San Francisco, CA | [Linkedin](#) | [Github](#) | [Portfolio](#)

EDUCATION

Northeastern University, Boston, MA

September 2023 - August 2025

Master of Science in Information Systems

Focused Coursework: Advance Data Science & Architecture, Parallel Machine Learning & AI, LLM with Knowledge Graph DB, Natural Language Engineering, AI Generative Modeling

Veermata Jijabai Technological Institute, Mumbai, India

August 2019 - June 2023

Bachelor of Technology in Information Technology

Focused Coursework: Data Structures and Algorithms, Linear Algebra, Discrete Mathematics, Machine Learning

SKILLS

ML & Deep Learning: PyTorch, TensorFlow, Scikit-learn, XGBoost, LightGBM, CNN, Vision Transformers, Distributed Data Parallel, Multi-GPU Training, Transfer Learning, Ensemble Methods, Optuna

MLOps & Model Serving: Docker, Kubernetes, MLflow, CI/CD, GitHub Actions, Model Monitoring, Drift Detection, Prometheus, Grafana, Terraform, FastAPI, REST APIs

LLM & Agentic AI: Claude API (tool calling, MCP), GPT-4o, LangChain, RAG, FAISS, ChromaDB, Hugging Face, Prompt Engineering, LLM-as-a-Judge Evaluation, Agent Orchestration, Evals

Data Engineering: Apache Spark, PySpark, Kafka, Airflow, Dask, Stream Processing, ETL Pipelines

Programming & Cloud: Python, SQL, C++, TypeScript, PostgreSQL, Redis, Snowflake, BigQuery, AWS, GCP

CV, NLP & Explainability: OpenCV, Med-SAM, Image Segmentation, BERT, RoBERTa, LIME, SHAP, Grad-CAM

PROFESSIONAL EXPERIENCE

Machine Learning Engineer | RSK IT International | Remote

September 2025 - Present

- Built low-latency streaming audio pipelines for drive-thru, toll-booth, and parking-lot voice agents across 3 live sites, with 80-120 ms chunking, VAD gating, ASR, and intent routing, holding end-to-end turn times under ~500 ms
- Deployed serving topologies for 3 production customers and 2 POCs across on-prem edge and hybrid edge/cloud, containerizing ASR/NLU microservices and gRPC/WebRTC streaming clients with latency monitoring
- Shipped a RAG-based QA helper over ~75 internal runbooks and a three-agent ticket-routing layer (docs, infra, model), cutting misrouted tickets and cross-team handoffs

AI/ML Research Assistant | Amal Lab, Northeastern University | Boston, MA

June 2024 - August 2025

- Trained a production deep-learning classifier to 97.28% accuracy on skin-cancer detection via comparative analysis of ResNet-34, EfficientNet-B1, and Vision Transformers with ensembles and external validation
- Engineered a Med-SAM segmentation system for MRI, CT, and histopathology data with automated preprocessing, improving lesion boundary detection through fine-tuned transformers
- Scaled infrastructure processing 18TB+ TCGA multi-omics data on Spark and Dask, applying clustering and dimensionality reduction (PCA, t-SNE, UMAP) across 33+ cancer types

Research Data Scientist | Brigham and Women's Hospital | Boston, MA

August 2024 - December 2024

- Developed a clinical-trial analytics pipeline over 10,000+ pharmaceutical reports using BERT and RoBERTa to link media sentiment with regulatory failure rates
- Designed a HIPAA-compliant meta-analysis framework across ClinicalTrials.gov and PubMed using random forests and gradient boosting

TECHNICAL PAPERS

Analysis of Explainable AI Methods on Medical Image Classification (IEEE) | [Link](#)

May 2023

Adversarial Attacks and Defenses for Skin Cancer Classification (IEEE) | [Link](#)

April 2023

PROJECTS

Bank Reconciliation Agent: Production-Grade Agentic Matching System | [Link](#)

May 2026

- Architected a 5-step agentic pipeline pairing rule-based matching with a Claude tool-calling fallback, reaching 90.6% accuracy and 92.5% F1 on a 32-case golden evaluation suite
- Added step-level traces, on-demand replay, and invariant enforcement with Postgres and Prometheus metrics

NBA Injury Prediction System | [Link](#)

October 2025

- Served real-time predictions via FastAPI at sub-100 ms p95 latency, sustaining 1000+ requests per second
- Built MLOps stack with Docker, Kubernetes, and MLflow versioning across XGBoost, LightGBM, and PyTorch models

CodeBuddy: Fine-Tuned CodeT5 Code Explanation with RAG | [Link](#)

August 2025

- Fine-tuned CodeT5 with a FAISS-backed RAG pipeline over 2.1M code-documentation pairs, improving BLEU 24.1% over a GPT-4 baseline with sub-35 ms retrieval across 2M+ code snippets

Skin Cancer Classification using High Parallel Machine Learning | [Link](#)

April 2025

- Trained a 35-class classifier to 91.8% accuracy on 245,000 images using EfficientNet-B3 and DenseNet-121
- Cut training time 3.37x with PyTorch DDP on 4x NVIDIA A100 GPUs and AMP, reaching 84.3% parallel efficiency