

Varun Sai Raigir

+1-945-273-0734 | Raritan, NJ | varunraigir@gmail.com | varunsairagir.com
[LinkedIn](#) | [GitHub](#) | [Google Scholar](#)

SUMMARY

AI/ML engineer building systems that reach real users. Currently developing production edge AI algorithms running on health wearable hardware. Previously built and shipped a clinical AI platform used by physicians in active pilot, generating documentation accepted with zero manual edits. Published researcher in multimodal cancer survival analysis. Works across the full stack: edge inference on constrained hardware, agentic AI and RAG systems, evaluation infrastructure, and cloud deployment.

SKILLS

Languages: Python, Java, C++, SQL, TypeScript, JavaScript, R

ML & Deep Learning: PyTorch, TensorFlow, Scikit-learn, CNN, GAT, GRU, LNN, Fine-tuning, Transfer learning

GenAI & Agentic AI: LangChain, LangGraph, LlamaIndex, RAG pipelines, FAISS, Pinecone, OpenAI GPT, Anthropic Claude, Gemini, MCP, Prompt engineering

Edge AI: TFLite Micro, ONNX Runtime, TinyML, Quantization, Pruning, CUDA, MPS, On-device inference

Backend & Frontend: FastAPI, Flask, Django, Spring Boot, React, Next.js, REST APIs, Microservices, PostgreSQL, MySQL, Supabase

Cloud & DevOps: AWS (S3, EC2, Lambda), Azure, GCP (Vertex AI, Gemini), Docker, Kubernetes, Terraform, CI/CD, GitHub Actions, PySpark

WORK EXPERIENCE

AI/ML Engineer | INVIZA Health

Jan 2025 – Present

- Building production edge AI algorithms for smart insole wearables: step count detection, gait analysis, and fall detection using sensor fusion and lightweight ML models compiled with TFLite Micro for real-time on-device inference without cloud dependency.
- Own the full pipeline from raw IMU sensor data collection through model training in PyTorch and TensorFlow, quantization-aware optimization, and embedded deployment on constrained hardware. Benchmark accuracy and latency trade-offs across quantization levels against target performance budgets.

AI/ML Engineer | mednabdh.ai

Mar 2025 – Aug 2026

- Designed and deployed a clinical AI platform that reached active pilot with physicians: real-time speech-to-text transcription, LangGraph-orchestrated multi-step LLM reasoning, RAG retrieval over a Pinecone vector store, structured SOAP note generation, and automated EMR field population via REST API. Physicians accepted output with zero manual edits.
- Built the evaluation infrastructure that made clinical deployment possible: offline golden-set evaluation against clinician-reviewed cases, LLM-as-judge scoring against documentation standards, CI-integrated regression tests gating every pipeline version, and responsible AI guardrails catching hallucinations before output reached physicians.
- Built FastAPI backend, PostgreSQL data layer, React frontend, and MCP servers for standardized tool connectivity. Deployed on Azure and GCP using Vertex AI and Gemini APIs with Docker, CI/CD, structured logging, observability dashboards, role-based access controls, and audit logging for PHI compliance.

ML Research Associate | University of South Florida

Dec 2023 – Dec 2025

- Designed a Graph Attention Network ensemble for precision medicine treatment stratification on multi-omics cancer data, achieving 92% accuracy across patient subgroups by modeling relationships between genomic, transcriptomic, and clinical features as a graph.
- Built automated data extraction and preprocessing pipelines using Python, Selenium, and NoSQL to integrate heterogeneous clinical datasets. Used PySpark for distributed processing across multi-node GPU clusters. Published MADSurv at ACM BCB 2025.

IT Enterprise Services Intern | Regeneron Pharmaceuticals Inc.

May 2024 – Aug 2024

- Automated ECR-to-Nexus container migration using Jenkins pipelines and Python, cutting deployment costs by over 15%. Built Splunk dashboards monitoring AWS asset utilization and compliance across enterprise infrastructure. Used Terraform for infrastructure-as-code provisioning.

Software Developer | Amadeus

Feb 2022 – Aug 2023

- Built modular Java Spring Boot microservices with MySQL and MSSQL backends for enterprise travel clients. Designed scalable REST APIs, diagnosed production issues including memory leaks and connection pool exhaustion using heap and thread dump analysis, and led testing with JUnit and PyTest. Reduced production bugs by 10% and response time by 5%.

Data Scientist Intern | Facedapter

Jan 2021 – Jul 2021

- Developed a CNN-based deep learning system to detect mask-based spoofing and presentation attacks in biometric authentication, increasing accuracy to 89%. Engineered training datasets improving generalization to real-world attack scenarios.

PUBLICATIONS

- E. Zhang, V.S. Raigir et al. "MADSurv: An Uncertainty-Aware Framework for Multimodal Cancer Survival Analysis." ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2025.
- V.S. Raigir. "Integrative Multi-Omics and Clinical Data Analysis for Predicting Recurrence and Survival in Uterine Cancer." M.S. Thesis, University of South Florida, 2025.

PROJECTS

- EdgeShield AI: Behavioral Anomaly Detection for Medical Devices

Jan 2026 – Present

 - Built an ML pipeline monitoring networked medical device telemetry: Isolation Forest on normal-operation baselines plus a TrendAnalyzer using rolling-window regression to catch gradual degradation before threshold breaches. Enforces a hard human-in-the-loop constraint at the API layer, informed by FDA 524B guidance on human oversight in software-assisted clinical systems.
github.com/VarunSai02/EdgeShieldAI
- Toothfairy Dental Insurance Advisor

Jan 2025 – Present

 - Production RAG application for insurance plan recommendation. Document ingestion, FAISS vector index, OpenAI embeddings, retrieval and ranking, and LLM generation with source-grounded recommendations. Python Flask backend, PostgreSQL, React frontend. Live at toothfairydentalinsurance.netlify.app.
- Shri AI: Cancer Risk Prediction Platform

Jan 2025 – Present

 - Deployed MIT's Mirai breast cancer risk model as a production REST API on AWS EC2 with Docker. Ported from CUDA to MPS and CPU backends, resolving hardware-specific numerical inconsistencies through PyTorch kernel-level analysis. Live at shri-ai.org.
- OwnerConnect

Jan 2025 – Present

 - Independently architected and shipped a full-stack real estate agent workflow platform: REST API backend, Supabase PostgreSQL data layer with schema design, authentication, and React frontend. Owned the system from database design through production deployment. Live at ownerconnect.net.
- LongRAG Dual-Perspective Medical QA

Sep 2024 – Dec 2024

 - Fine-tuned LongRAG on PubMedQA using FAISS and LangChain. Built a full agentic pipeline covering query understanding, multi-document retrieval, dual-perspective reasoning over conflicting evidence, and answer generation, improving QA accuracy from 55% to 57% through iterative prompt and retrieval optimization.
- Edge AI Health Monitor

Apr 2025 – Jun 2025

 - Real-time physiological tracking pipeline using Liquid Neural Networks, CNN, and GRU. Compiled to ONNX and optimized with TinyML quantization for deployment on resource-constrained hardware, profiling inference latency and memory footprint across optimization levels.

EDUCATION

M.S., Computer Science | University of South Florida
B.S., Computer Science | GITAM University, India