

Madhav Suri

(214) 208-3176 | Surimadhav06@gmail.com | linkedin.com/in/madhav-suri | github.com/madhav-suri | U.S. Citizen

EDUCATION

The University of Texas at Dallas

Richardson, TX

Bachelor of Science in Computer Science and Mathematics | GPA: 3.8/4.0

Expected Dec. 2027

– **Relevant Coursework:** DSA, OOP, Discrete Mathematics, Database Systems, Software Engineering, Operating Systems

– **Certifications:** AWS Cloud Practitioner, Google Project Management Certificate, CAPM in Progress

Portfolio: madhavsuri.com

SKILLS

Languages: Python, Java, JavaScript, TypeScript, C++, Dart, SQL, Bash, Go, HTML, JSON

Frameworks/Tools: React, Next.js, Node.js, FastAPI, REST APIs, Git, Firebase, Postgres, Docker, Kubernetes, Jenkins, Linux, AWS

EXPERIENCE

Frontend Developer

Jan. 2026 – Present

ACM UTD

Richardson, TX

- Built and shipped the **React/TypeScript** frontend of D.I.Y.A., an **AI-powered academic-support platform** used by 350+ live users across 5+ product workflows, including scheduling, forum Q&A, self-check grading, and professor analytics.
- Integrated **AWS**-backed application workflows across S3, API Gateway, Cognito, **Lambda**, **Aurora PostgreSQL**, to support authenticated routing, notifications, and role-based university access.
- Built **RAG** and intelligence-layer features using **OpenSearch vector retrieval**, S3 course documents, Lambda ingestion/chunking, and **Claude/OpenAI** workflows, achieving 92% query-routing resolution, 95% retrieval relevance, and 45+ reusable UI/API flows.

Software & AI Engineering Intern

Aug. 2025 – Dec. 2025

SoftBank Robotics America

Southlake, TX

- Developed EdgeSight, a **real-time assistive vision system** running fully on a resource-constrained **NVIDIA Jetson Orin Nano** with 8GB unified memory, Linux/JetPack 6.2, and **CUDA 12.6**.
- Integrated **YOLOv8 OIV7 600-class** detection, multi-object tracking, a **multimodal VLM** assistant, and a **FastAPI** dashboard, sustaining 15+ FPS under CPU/GPU memory, power, and thermal limits.
- Reduced on-device VLM query latency by ~8x from 74s to ~7–9s after benchmarking 5 VLMs, including Gemma-4B, Qwen2.5-VL-3B, moondream, and Gemma-4 E4B/E2B, across **Ollama**, **vLLM**, and **llama.cpp**.
- Deployed a 4-bit-quantized Gemma-4 E2B model on **llama.cpp** with **FlashAttention** and q8_0 KV-cache, while accelerating detection through **TensorRT**, FP16 mixed precision, and MAXN_SUPER clock tuning.

Data Engineering Intern

May 2025 – Aug. 2025

Riddell

Irving, TX

- Engineered automated **ETL pipelines** in **Python** using pandas, NumPy, and SQL to process 5,000+ helmet-impact sensor readings per session, reducing manual data-preparation time by 55%.
- Built **scikit-learn**-powered telemetry reconciliation workflows against performance baselines, improving data quality checks and reducing product-iteration reporting time by 82%.

PROJECTS

StreamForge | *Go, Kafka, gRPC, Protobuf, Parquet, S3, Apache Iceberg, Prometheus*

[LINK](#)

- Built a **distributed stream processing engine** in Go (**Kafka** ingest, coordinator-managed worker pool, keyed **gRPC/Protobuf** shuffle) sustaining ~40K events/s at $p99 \leq 25\text{ms}$ across 3 workers.
- Added two-phase aligned **checkpointing** with **heartbeat failure recovery**: reassigns a dead worker's partitions and key-buckets and restores state from the last checkpoint with **zero data loss** in ~2.8s.
- Reached **effectively-once** output via staged two-phase commit with idempotent (key, window) dedup, plus **event-time watermark windowing** and an **Apache Iceberg lakehouse** sink supporting time-travel queries.

FinHack (MLH) 3rd Place, RiskOS AI | *FastAPI, React, SQLAlchemy, Spark, DuckDB, scikit-learn, Pytest, S3*

[LINK](#)

- Architected a **fintech fraud-operations platform** using **FastAPI**, **React**, and SQLAlchemy ORM to process and route a 6.3M-row PaySim-simulated transaction dataset into analyst review queues, placing 3rd among 150+ participants.
- Built an analytics layer using **Parquet** datasets in S3, queried through **DuckDB** and **Spark**, to support experiment tracking and case-review workflows under hackathon time constraints.
- Implemented **explainable risk scoring** with **scikit-learn** Logistic Regression and 7 deterministic validation rules, achieving 0.901 ROC-AUC, 82.6% fraud recall, and 94.3% automated routing coverage.

– Added **RBAC**, manager-only overrides, immutable audit trails, and a 111-test **Pytest** suite in a **GitHub Actions CI/CD** pipeline.

RepoPulse Agent | *Next.js, TypeScript, OpenAI API, GitHub API, Vitest, Docker*

[LINK](#)

- Engineered an **agentic AI codebase-intelligence platform** in **Next.js** and **TypeScript** with a 7-stage workflow, audit trail, and human validation gates.
- Built **semantic code search** using **OpenAI embeddings**, syntax-tree parsing, GitHub API patch testing, **Docker** sandboxes, and 200+ Vitest checks.
- Sustained sub-1.5s latency across benchmarked 5,000 indexed files while cutting token spend by 35%.

LEADERSHIP AND ACTIVITIES

Research Team Lead

Jan. 2026 – Present

The Erik Jonsson School of Engineering and Computer Science at UT Dallas

Richardson, TX

- Led undergraduate research under Prof. Yi Ding on **Trusted Execution Environments and confidential ML inference**, building secure C/C++/Python pipelines with **ARM TrustZone** and **Intel SGX**; presented at the Capital of Texas Undergraduate Research Conference; limited overhead to ~12% across 5 edge-AI workloads with a unified cross-platform codebase portable across SGX and TrustZone via the **Open Enclave SDK**.

Preprint: <https://doi.org/10.5281/zenodo.20840404> — **Code:** <https://github.com/Mxs8513/EnclaveSplit>.