Aravind Gudikandula

aravind.g1942@gmail.com | +1-669-499-8490 | Linkedin | Github | San Jose, CA

Education

- **San Jose State University** — San Jose, CA
  Master of Science, Software Engineering — *Dec 2025*
- **Malla Reddy College of Engineering and Technology** — Hyderabad, India
  Bachelor of Technology, Information Technology — *Jun 2023*

Skills Summary

- **Languages & Frameworks**: Python, Java, C++, C, Ruby, JavaScript, TypeScript, React, Angular, Node.js, Express, Flask, Django, Rails, HTML, CSS, SQL
- **AI, GenAI & Vision**: PyTorch, TensorFlow, Gemini API, YOLOv8, Stable Diffusion, RAG, LangChain, OpenCV, LLMs, Scikit-learn
- **Cloud, Data & DevOps**: AWS, GCP (Cloud Run/Build), Azure, Docker, K8s, Kafka, Flink, MongoDB, PostgreSQL, Redis, Oracle, Linux, CI/CD, REST APIs, Selenium, Git
- **Certifications**: AWS Certified Cloud Practitioner, Google IT Automation with Python
- **Soft Skills & Tools**: Agile/SDLC, JIRA, Confluence, Product Management, Stakeholder Communication, Problem Solving

Experience

- **Hitachi Vantara** — Remote
  Associate Software Engineer — *Jan 2023 - Dec 2023*
  - Engineered scalable microservices using Python and Java, integrating with AWS cloud infrastructure to deliver enterprise solutions that improved system latency by 35% and fine tuned overall application performance.
  - Developed RESTful APIs and backend services using Flask and Spring Boot, supporting team efforts to process 1M+ requests daily while maintaining 99.9% uptime through efficient database queries and caching strategies.
  - Collaborated in Agile teams to design and deliver full-stack features across web applications, conducting code reviews and writing comprehensive unit tests that increased code coverage from 65% to 90%.
  - Implemented automated testing frameworks using JUnit and Pytest, reducing bug detection time by 40% and accelerating the CI/CD pipeline for faster deployment cycles.
- **F13 Technologies** — Remote
  AWS Intern — *Feb 2022 - Nov 2022*
  - Designed cloud infrastructure solutions using AWS services including EC2, S3, Lambda, and RDS, supporting scalable application deployment for 5+ client projects with 99.8% availability.
  - Automated deployment workflows using CloudFormation templates and CI/CD pipelines, decreasing deployment time by 30% and minimizing production incidents through infrastructure-as-code practices.
  - Monitored cloud resource utilization using CloudWatch and Cost Explorer analytics, identifying cost-saving opportunities that delivered 18% savings in monthly AWS expenditure.
  - Migrated legacy applications to cloud-native architectures on AWS, improving system scalability and decreasing infrastructure maintenance overhead by 25% for client operations.

Projects

- **Building Computer Vision Framework for Smart Vehicles** — Jan 2025 - Dec 2025

  Skills: Python, PyTorch, YOLOv8, Stable Diffusion, CLIP, FAISS, CUDA
  - Architected production-ready computer vision framework integrating Detection, Search, and Generative Insertion modules, exceeding targets with PSNR of 32.44 dB (29% above) and SSIM of 0.9226.
  - Streamlined ML pipeline implementing batch processing and FP16 mixed precision, achieving 12-24x speedup that accelerated processing from 6 hours to 15 minutes with 22 FPS at 95% accuracy.
- **Resolve AI – AI-Powered Visual Repair Assistant** — Dec 2025

  Skills: React 19, TypeScript, Express.js, Gemini 2.5 Flash, Supabase, Google Cloud Run, Docker
  - Built full-stack AI diagnostic application using React 19 and Express.js, integrating Gemini 2.5 Flash for multimodal analysis with 90%+ reliability and real-time WebRTC media capture.
  - Deployed containerized application to Google Cloud Run with multi-stage Docker builds, integrated Supabase PostgreSQL for persistent scan history and enabled auto-scaling capabilities.
- **Medical Domain RAG System** — May 2024

  Skills: BioMistral, Qdrant, LangChain, FastAPI, Llama.cpp, PubMedBERT
  - Established open-source RAG system leveraging BioMistral (7B) and PubMedBERT, achieving 85% relevance accuracy on healthcare queries with local LLM deployment for data privacy.
  - Enhanced embedding pipeline using Qdrant vector database and LangChain with FastAPI integration, maintaining sub-second response times while decreasing overhead by 40%.