

Sreeram Saravana Prasad

sreeram.ss2001@gmail.com | linkedin.com/in/sreeram-s-5454961aa | 6027512565 | github.com/sreeram2001 | sreerams.online

EDUCATION

Master of Science in Computer Science, Arizona State University | GPA : 4.0/4.0

Aug 2025 - Aug 2027

B.Tech in Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham

Jul 2018 - Jul 2022

TECHNICAL SKILLS

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

AI/ML & LLMS: AWS Bedrock, AgentCore, Vector DB, RAG, Multi-Agent Systems, MCP, LangChain, LangGraph, TensorFlow, OpenCV, Mask R-CNN, Scikit-learn, PyTorch, Gemini, Tool-calling

Development: Next.js, React.js, Node.js, Express, Spring Boot, Flask, FastAPI, HTML, CSS, REST APIs

Cloud & DevOps: AWS (EC2, Lambda, API Gateway, S3, SQS), MongoDB, DynamoDB, Jenkins, Docker, Git, CI/CD

WORK EXPERIENCE

AI Full-Stack Developer, ASU AI Cloud Innovation Center (Powered by AWS)

Jan 2026 - Present

- Lead AI/ML based public sector projects on AWS through the AI Development Life Cycle, designing cloud-native serverless architectures (AWS Lambda, API Gateway, S3, DynamoDB) and integrating generative AI capabilities using Bedrock, AgentCore, MCP, and RAG pipelines with Titan embeddings.
- Architected an AI presentation coaching tool on a serverless stack (Cognito, Amplify, DynamoDB, S3), serving 1,000+ students while reducing infrastructure costs by 35%. Leverage Bedrock LLMs for context-aware question generation and post-session analytics, achieving 35% higher accuracy and 40% faster retrieval.
- Developed a real estate coaching tool with multi-persona support, bidirectional voice Q&A AI agents using Bedrock AgentCore, Strands Agents, multi-model orchestration, real-time transcription, edge-based gaze detection, multi-agent orchestration, tool-calling, and MCP integrations for domain-specific knowledge retrieval.

Software Engineer, General Electric (GE Healthcare)

Aug 2022 - Jul 2025

- Developed and integrated MRI post-scan imaging algorithms (ADC, ASL, DWI Propeller) in Java and C++ for clinical pipelines. Optimized ADC computation via multi-threaded pixel-wise diffusion coefficient mapping, improving throughput by 30% and MR Touch liver stiffness accuracy by 25%. Implemented medical image grouping and inline processing.
- Engineered Window Manager and Desktop Management microservices using Java, Spring Boot, RabbitMQ (async message queuing), Eureka (service discovery), and RESTful APIs. Diagnosed and resolved race conditions, memory leaks, and thread deadlocks causing application freezes and blank screen rendering. Reduced response latency by 40% and eliminated 35% of recurring UI freeze incidents through event-driven architecture optimizations.
- Refactored legacy Java and C++ algorithm codebases, resolving 9,000+ static analysis issues, vulnerabilities, code smells. Improved runtime performance by 30% and increased test coverage to 85% using JUnit, Mockito, Cucumber, and Optimus.
- Architected CI/CD pipelines using Jenkins, GitLab automating build, test, analysis phases across microservices. Implemented quality gates, version control workflows, reducing release defects by 50% and deployment rollbacks by 30%.

Software Engineer Intern, General Electric

Feb 2022 - Jul 2022

- Enhanced X-ray Intelligence workflow by engineering automatic camera alignment using Intel RealSense, Detectron2 for anatomical keypoint detection, and Mask R-CNN for instance segmentation (Python, TensorFlow, OpenCV), achieving 90%+ segmentation accuracy and reducing manual calibration time by 60%. Built a FastAPI REST API for real-time ML inference serving segmentation masks and bounding box coordinates with sub-200ms response time.

PROJECTS

SONIC - Voice AI Productivity Agent

LIVE

Built a real-time voice AI agent on AWS Bedrock (Nova Sonic), Node.js, with 12 tool-calling integrations (Zoom, Google Meet, Slack, Calendar, YouTube). Implemented an Agentic RAG pipeline with web grounding, wake-word and browser-native mic capture.

SparkyFi - Financial Wellness App

LIVE

LLM-powered financial AI agent built with Next.js, Gemini API, and Plaid integration with 5+ API integrations. Implemented RAG pipeline over 5,800+ records using Pinecone vector DB for personalized financial advice, insurance recommendations. Features PDF vision analysis, interactive voice agent with ElevenLabs, and real-time anomaly detection with Supabase for auth, data persistence.

Learnify Genie - Text-to-Comic AI Tool for Interactive Learning

GITHUB

Created an educational AI agent using Python, Flask, AWS Nova, Claude LLMs that converts text into comic strips, flashcards. Engineered multi-model pipeline for summarization, script and image generation, deployed on Elastic Beanstalk with Next.js.

Elastic Face Recognition Application on AWS

Built a multi-tier face recognition app on AWS (EC2, SQS, S3, DynamoDB) with Flask, dynamic EC2 scaling based on SQS queue depth, and real-time recognition using FaceNet. Achieved sub-second latency processing 100 concurrent requests with 100% accuracy.