Maruri Sai Rama Linga Reddy

+91 7893865644 | sairam.maruri@gmail.com | saiii.in | LinkedIn | GitHub

Education

Vellore Institute of Technology Andhra Pradesh, IN

Bachelor of Technology in Computer Science and Engineering | CGPA: 8.31/10.0 *Sep 2022 – Aug 2026*

Relevant Coursework: Deep Learning, Reinforcement Learning, Data Structures & Algorithms, Neural Networks, Operating Systems, Software Engineering (Design Patterns, System Design), Computer Networks, Machine Learning, Natural Language Processing

Technical Skills

Languages: Python, Java, C++, GO, SQL

ML/AI Frameworks: PyTorch, TensorFlow, Keras, scikit-learn, ONNX, Hugging Face Transformers

LLMs & GenAI: GPT-4, Claude, BERT, T5, QWEN, LangChain, RAG, Prompt Engineering, Fine-tuning, RLHF

ML Infrastructure: FAISS, Vector Databases (Pinecone, Chroma), Neural Network Optimization, Model Deployment

Cloud & DevOps: AWS (EC2, Lambda, RDS, S3), Docker, Kubernetes, Terraform, CI/CD, Git

Research Tools: Jupyter, Colab, Weights & Biases, TensorBoard, Matplotlib, Seaborn

Research & Engineering Projects

End-to-End AI Research Platform with LLM Integration *May 2025 – Present*

- Architected and deployed scalable research platform (orravyn.cloud) on AWS supporting 100+ concurrent users with multi-role authentication, reducing AI pipeline latency by 30% through optimization of LLM inference
- Engineered Retrieval-Augmented Generation (RAG) system using FAISS vector embeddings and custom ranking algorithms, indexing 100+ research papers and achieving 45% faster semantic search with 4-5 relevant paper retrieval per query
- Implemented multi-document summarization using BART and transformer-based models, integrated with GPT-4 and Claude APIs for research recommendations, reducing response time by 25% via LangChain orchestration
- Built production-grade Django REST API with 10+ permission-restricted endpoints, real-time chatbot interface, and Lambda functions for automated paper processing and classification
- **Technologies:** Python, PyTorch, BART, FAISS, Vector DBs, LangChain, OpenAI API, Claude API, AWS (Lambda, RDS, EC2), Django REST, ML Optimization

Medical Image Classification with Deep Learning *Feb – Mar 2025*

- Developed and optimized WideResNet-based CNN model achieving 92% accuracy on bone fracture classification across 10,000+ X-ray images, implementing data augmentation pipelines to reduce overfitting by 20%
- Conducted ablation studies on architectural choices and hyperparameters, reducing misclassification by 15% through systematic experimentation and performance analysis using confusion matrices and ROC curves
- **Technologies:** PyTorch, TensorFlow, WideResNet, CNN Architectures, Transfer Learning, Computer Vision

Large-Scale ML System for Customer Behavior Prediction *Sep – Nov 2024*

- Designed distributed machine learning system using Random Forest ensemble achieving 99% accuracy on 100K+ records, implementing feature engineering and model optimization techniques reducing customer churn by 25%
- Performed comprehensive exploratory data analysis and preprocessing pipeline, reducing dataset noise by 15% through outlier detection, feature selection, and statistical validation methods
- **Technologies:** Python, Random Forest, scikit-learn, XGBoost, Pandas, Statistical Analysis, ML System Design

Certifications & Research Interests

AWS Certified Cloud Practitioner – Cloud infrastructure and deployment

Oracle Generative AI Professional – LLM architectures and applications

Oracle AI Vector Search Professional – Semantic search and embeddings

Research Interests: Reinforcement Learning, Scaling Laws, Code Generation, AI Safety & Alignment, Interpretability

Leadership Experience

Photon Club, VIT-AP University Andhra Pradesh, IN

Administrator – Event Management & Team Leadership *Feb 2024 – Apr 2025*

- Led cross-functional team of 15+ members in executing events and workshops, achieving 15% membership growth through strategic outreach and collaborative initiatives