RAMA PAVAN NAGA SAI SANTHOSH KATRAGADDA

Dallas, Tx, USA • +1 (573) 514-4759 • [rk7rd@umsystem.edu](mailto:rk7rd@umsystem.edu%20) • [LinkedIn](https://www.linkedin.com/in/santhosh-katragadda/)

# PROFESSIONAL SUMMARY

Machine Learning/AI Engineer with 5 years of experience designing and delivering robust, user-centric applications. Skilled in transforming complex requirements into high-performance solutions and developing LLM models. Committed to innovation and excellence, with a track record of driving project success.

# EDUCATION

University of Missouri**,** Columbia, MO, USA Jan 2023 – May 2024

**Master of Computer Science**

Vignan’s Lara Institute of Technology & Sciences, Guntur, India Aug 2018 – Jun 2022

**Bachelor of Technology in Computer Science Engineering**

# SKILLS & OTHER

**Programming Languages:** Python, Java, C, C++, SQL

**Web Technologies:** HTML, CSS, JavaScript, Angular, React

**Frameworks & Libraries:** Django, FastAPI, Node.js, Express.js

**Big Data Technologies:** Apache Spark, Apache Hadoop, Apache Airflow, AWS Glue, Azure Data Factory, Databricks

**Operating Systems:** Windows, Unix/Linux, Ubuntu

**Cloud & Containerization:** Google Cloud Platform (GCP), Amazon Web Services (AWS), Microsoft Azure, Jenkins, Kubernetes, Docker, Terraform, CI/CD

**Databases & Storage:** MySQL, PostgreSQL, Oracle, MongoDB (NoSQL), Snowflake, Neon DB

**Tools/IDE/Application Servers/ Skills:** Git, GitHub, GitLab, Bitbucket

**AI & ML Frameworks:** TensorFlow, PyTorch, Hugging Face Transformers, keras, Vertex AI, AWS SageMaker, RAG, LLM’s AND Lcm’s

**Data Visualization & Analytics:** Power BI, Tableau, Looker, SQL Reporting Services (SSRS), Prompt Engineering (ChatGPT, Claude, Gemini)

# PROFESSIONAL EXPERIENCE

## AT&T, Dallas, Tx June 2024 – Present

**AI/Machine Learning Engineer**

* Designed and implemented a scalable vector database leveraging NeonDB, Pinecone, Weaviate, and pgvector to enable high-performance semantic search and efficient similarity retrieval across diverse data sources.
* Developed a Retrieval-Augmented Generation (RAG) pipeline powered by DeepSeek-R1-Distill-Llama-70B, deployed on AWS SageMaker and Vertex AI, delivering scalable, context-aware AI responses through optimized inference workflows.
* Integrated OpenAI’s GPT suite, including hands-on experience with GPT-4V and GPT-4o, to support advanced multimodal tasks such as visual reasoning, conversational agents, and document understanding within GenAI-powered pipelines.
* Integrated OLAMA, Kubernetes, and AWS SageMaker for containerized, scalable deployment, ensuring high availability and resource-efficient orchestration across hybrid cloud environments including GCP GKE.
* Implemented HNSW indexing and hybrid search (BM25 + dense embeddings) on AWS OpenSearch and Vertex AI Matching Engine to boost retrieval accuracy and search relevance in both structured and unstructured datasets.
* Built robust ETL pipelines for automated data ingestion, transformation, and embedding storage using AWS Glue, S3, DynamoDB, and GCP Cloud Functions with BigQuery and Cloud Storage, ensuring high data integrity and accelerated access.
* Enhanced LLM inference performance via query compression, response caching, and indexed retrieval, using AWS Inferentia, SageMaker Neo, Amazon Elastic Inference, and GCP TPU, reducing API latency and backend load by 40%.
* Deployed and fine-tuned LLMs on both AWS SageMaker and Vertex AI using distributed training with EC2, SageMaker Training Jobs, and Vertex AI Pipelines, accelerating model adaptation to domain-specific tasks.
* Implemented real-time observability with AWS CloudWatch, Prometheus, Grafana, and GCP Cloud Monitoring, enabling proactive system health checks, performance tracking, and query optimization.
* Advanced multi-model inference by integrating Gemini, GPT, DeepSeek, and Qwen APIs, enabling AI-driven decision-making, comparative response generation, and cross-model orchestration across AWS, GCP, and on-premises environments.

Life Sciences Center, University of Missouri, Columbia, MO, USA Aug2023 – May 2024

**Gen AI/Machine Learning Engineer**

* Created an innovative solution for automating gene pathway analysis from documents, leveraging the GPT-4 API to streamline extraction accuracy and enhance analytical efficiency.
* Designed and deployed a specialized machine learning model to accurately detect and extract gene pathway images from diverse sources, ensuring high-quality data retrieval critical for in-depth pathway analysis.
* Seamlessly integrated GPT-4, Claude 3, Gemini API’s and advanced analytical prompts, enabling precise extraction and interpretation of complex gene relationships and interactions within pathway images.
* Built and deployed LLM-powered Retrieval-Augmented Generation (RAG) pipeline specifically for gene data, integrating with gene databases.
* Leveraged advanced OCR and computer vision models including MMOCR, SAM, Google OCR, and Detectron and Siamese networks enhancing processing, entity extraction, and multimodal biological data interpretation.
* Built a robust verification system to enhance reliability, significantly reducing error rates and ensuring consistent, high-quality outputs in gene pathway research.
* Developed a hybrid analytical model, combining the capabilities of Claude 3 opus with my custom pipeline, achieving a notable improvement in the precision of biological data interpretation.
* Successfully published a research paper and presented a poster at the AMIA conference, demonstrating the project’s impact on cancer research and bioinformatics, and showcasing expertise in integrating cutting-edge computational approaches to support therapeutic discovery.

## Life Sciences Center, University of Missouri, Columbia, MO, USA Jan 2023 – Aug 2023

**Full Stack Developer**

* Led an innovative gene detection project, integrating object detection and text recognition models (SAM by Meta-AI, Google OCR, and MMOCR) to enhance genetic research and bioinformatics analysis.
* Designed and developed an intuitive, researcher-friendly UI, enabling seamless selection from multiple detection methods and optimizing usability, accessibility, and workflow efficiency.
* Automated analytical workflows, facilitating comprehensive gene pathway analysis, uncovering critical insights into gene activation and inhibition mechanisms.
* Published groundbreaking research in a leading international journal, showcasing expertise in AI-driven bioinformatics, machine learning, and genetic data processing.
* Engineered scalable web applications using React.js, Node.js, and MySQL, ensuring efficient data flow, robust performance, and seamless front-end/back-end interactions.
* Integrated and optimized API interactions, implementing RESTful services and CI/CD pipelines to enhance development efficiency, deployment automation, and system reliability.
* Developed reusable React components, improving code modularity, maintainability, and UI/UX responsiveness for enhanced user experience.
* Built high-performance server-side applications with Node.js, optimizing data processing, API performance, and backend scalability.
* Executed full Software Development Life Cycle (SDLC), including unit testing, integration testing, Agile methodologies, and DevOps best practices.

## Bank of Baroda, India Jan 2020 – Dec 2022

**Machine Learning Engineer**

* Designed and deployed scalable ML pipelines for fraud detection, credit risk assessment, and customer segmentation, leveraging Azure ML, Azure Databricks, AKS, and Azure Functions for end-to-end automation.
* Implemented real-time predictive analytics using Python (scikit-learn, TensorFlow, PyTorch) and Azure Synapse Analytics, reducing fraudulent transactions by 30% through early detection models.
* Developed ETL workflows for processing multi-terabyte financial data, integrating Azure Data Factory, Azure Data Lake, and SQL Data Warehouse to ensure low-latency data processing and regulatory compliance.
* Optimized ML feature engineering using Azure Logic Apps and Function Apps, improving fraud detection and credit scoring efficiency by 40%.
* Built AI-powered personalized financial recommendation systems using Azure Cognitive Services and MLflow, increasing customer engagement and cross-selling by 20%.
* Enhanced credit risk prediction models using LSTMs and time-series forecasting, improving loan default prediction accuracy by 15%.
* Deployed and monitored ML models in production, utilizing Azure ML Ops, Azure Monitor, and Application Insights to ensure SLA compliance and model drift detection.
* Implemented big data analytics with Apache Spark on Azure Databricks, reducing query processing time for financial transactions by 30%.
* Developed classification and clustering models for customer segmentation and risk analysis, improving fraud detection and loan approval processes by 25%.
* Collaborated in Agile teams, driving sprint planning, iterative model improvements, and AI-driven financial solutions aligned with business objectives.

**PROJECTS**

**Assessment and Integration of Emerging AI Chatbots for Tumor Signaling Pathway Curation**

* Led a groundbreaking project titled "Assessment and Integration of Emerging AI Chatbots for Curation of Tumor Signaling Pathways" as part of my role in the Department of Electrical Engineering and Computer Science at the University of Missouri. This initiative focused on integrating advanced AI chatbots—GPT-4, Claude3 Pro, and Gemini—with a custom pipeline I developed, utilizing Large Language Models (LLMs) for the detailed analysis of gene interactions in tumor signaling pathway images. Designed and implemented a hybrid model that combined the powerful analytical capabilities of Claude3 Pro with the precision of my pipeline, significantly improving the accuracy of biological data analysis. This project resulted in a research paper publication and a poster presentation at the AMIA conference, highlighting its pivotal contributions to the field of bioinformatics. My work demonstrated proficiency in deploying advanced computational strategies to enhance cancer research and therapeutic discovery efforts.

**Multimodal Expression Recognition**

* Led the development of a multi-modal expression recognition project, focusing on detecting expressions through image-based and relation-based inputs. The image-based component involved processing three types of inputs: images, videos, and webcam feeds. I implemented two models: the MTCNN model for face extraction (from both static images and video frames), followed by the VGG-19 model for feature extraction and expression classification. The relation-based component used text and audio inputs, utilizing a Bag of Words model for text analysis and Librosa for audio-based expression detection.
* This project culminated in the publication of a paper, which holds an impact score of 3.74, further validating the innovation and significance of the work in multi-modal expression recognition.

# PUBLICATIONS

* Presented research and poster at AMIA Conference, highlighting effective communication and engagement with professionals.
* Published paper in IEEE conference(bibm) under Bio-medical and AI advancements

(<https://www.computer.org/csdl/proceedings-article/bibm/2024/10822181/23onpGBssrm>)