SAROJ SAH

☎ (+91) 900-341-0891 | M sarojsahaero@gmail.com | in saroj-shah

EXPERIENCE

SENIOR DATA SCIENTIST | ACCENTURE | MAY 2025 – PRESENT

Verisk Analytics (GenAI Production)

- Led end-to-end development of production **RAG chatbot**, using LlamaIndex, AWS Bedrock, Postgres etc. enabling **500+** users to query internal data with **Role Based Access Control**.
- **Optimized** solution architecture & database execution, **reducing response latency to 10s,** achieving **>98% accuracy** for real-time querying.
- Implemented **Root Cause Analysis, Business Narrative generation** and **Anomaly Detection,** transforming raw KPIs into actionable insights.
- Deployed strict **AI Guardrails** and **fallback logic** to eliminate hallucinations, ensuring the solution was strictly grounded in client data.

Hormel Foods (GenAI Innovation)

- Architected a **Text-to-SQL** solution for supply chain analytics, allowing non-technical users to query complex databases with **>90% accuracy**.
- Developed custom **Intent Detection** algorithms and interactive dashboards, demonstrating a **>90% reduction** in data retrieval time during client validation phase compared to previous manual approach.

AIML RESEARCHER | TCS RESEARCH | AUG 2021 – APR 2025

CLINICAL TRIAL INSIGHTS GENERATION

- Developed a system leveraging **LLM** on **AWS Bedrock** to automate the generation of **clinical trial insights** for medical monitoring.
- This resulted in **20%** improvement in medical monitoring efficiency by enabling faster data review**, 10%** improvement in early risk signal detection and **10%** reduction in patients dropout rates**.**

LABEL DETECTION ON DRUG PACKAGES

- **Fine-tuned YOLOv8** model and **AWS Rekognition** to perform image segmentation of **drug labels**, and **labelImg** for annotation and performed automated comparison between master and target labels.
- It detected the presence, absence, & positional changes of labels on drug packages with over **90% accuracy** to classify its categories.

SITE SELECTION THROUGH KNOWLEDGE GRAPH AND NLP

- Designed **knowledge graphs** from unstructured and structured clinical data leveraging **OpenAI, RAG**, **Langchain** and Ne**o4j graph database.**
- Leveraged query based, multi-level filtering for efficient **site selection and patient recruitment** during clinical trials.

PATIENT PROFILE SCORING SYSTEM IN CLINICAL TRIAL (PATENTS)

- Devised a **patient scoring system** for assessing trial risk in patients, resulting in **15%** improvement in trial risk identification accuracy.
- This novel method employed **Isolation Forest** to identify high-risk patients, combined with **statistical** and **normalization** techniques followed by **PCA** for patient risk stratification and ranking.

LITGPT - LITERATURE INSIGHTS

- Leveraged **fine-tuned LLM** and **NLP** to extract and analyze data from **medical databases**, generating insights based on user queries**.**
- Implemented **ICSR** classification system that **automates** the analysis of medical literature for categorizing as **valid**, **potential** or **non**-**ICSR**.
- Achieved **~24%** operational efficiency gain, reducing task completion time from **3.5 hours** to **50 minutes.**

EDUCATION

IIT, BOMBAY

M.Tech | CPI 8.23/10.0 | 2019-21

HITS, CHENNAI

BTech | CPI 9.0/10 | 2015 – 19

CERTIFICATIONS

- **AWS Certified Machine Learning Speciality** (July 2024)
- **AWS Certified Solution Architect** (June 2022)

PUBLICATION

Advancing Regulatory Intelligence with Conversational and Generative AI

Published in PharmaSUG 2024

Patent: Patient Profile Scoring System for Clinical Trial Risk Assessment (Patent Pending)

TECHNICAL SKILLS

LANGUAGES: Python, SQL, MATLAB, C

ML/DL: Scikit-learn, TensorFlow, PyTorch, Keras, XGBoost, Computer Vision, NLP

LLM & GENAI: OpenAI, AWS Bedrock, HuggingFace, LangChain, LlamaIndex, RAG, Fine-tuning, Prompting, Agentic AI, MCP

CLOUD & DATA: AWS (SageMaker, S3, Redshift), Azure, GCP, Postgres, Vector Databases (Pinecone, ChromaDB), Docker

OTHER PROJECTS

- Applied **LSTM model** for predictive maintenance of industrial motors by analyzing sensor data, predicting Remaining Useful Life to minimize downtime.
- Designed, fabricated, and piloted a full scale **Martian Rover** at **NASA Marshall Space Flight Center**, leading team in navigating Martian terrains to collect samples.

ACADEMIC

REDUCED-ORDER MODELLING OF THERMOACOUSTIC INSTABILITIES | IIT BOMBAY| MTECH PROJECT | 2020-21

- Applied data-driven DMD techniques to analyze **complex**, **high dimensional dynamical** systems in gas turbine engine.
- Developed ROM using **Dynamic Mode Decomposition** and **SVD** on thermoacoustic instabilities data, to extract dominant modes at corresponding frequency to pinpoint instability sources and contributed to **preventing system failures.**