Madhumitha Sekamuri

madhutrish06@gmail.com | +1-614-483-7876 | LinkedIn | GitHub | Ashburn, VA(open to relocation)

Summary

Data Scientist with proven experience delivering scalable ML, anomaly detection, and LLM driven solutions across healthcare, insurance, and enterprise data systems. Adept in predictive modeling, generative AI, and cloud platforms (AWS, Azure, GCP). Skilled in the full ML lifecycle from EDA and feature engineering to deployment and optimization. Passionate about combining GenAI, anomaly detection, and knowledge graphs to deliver explainable insights in high-stakes domains including audit, healthcare, and risk analysis.

Experience

Moxie Ashburn, VA (remote)
Data Scientist Feb 2025 – Present

- Developed anomaly detection models on financial and operational datasets using **scikit-learn, PyTorch, and XGBoost, deployed on AWS SageMaker,** to surface irregular transactions and reporting inconsistencies; improved detection of audit-relevant anomalies by 35% compared to rule-based baselines.
- Built an **LLM-powered** audit document review assistant leveraging **LangChain, FAISS, and OpenAI embeddings, deployed on AWS Lambda and API Gateway,** to summarize contracts, invoices, and compliance reports; reduced auditor review time by 60% while maintaining source-grounded transparency.
- Engineered a knowledge graph pipeline using **Neo4j, Python, and entity extraction with GPT-4** to link clients, accounts, and transactions across disparate systems, enabling auditors to visualize high-risk relationships and improving investigative efficiency during assurance reviews.

Behavioral Neuroscience – The Ohio State University Columbus, OH
Data Scientist May 2024 – Dec 2024

- Implemented a SLEAP-based multi-animal tracking system using **Python, TensorFlow, and OpenCV** on **Google Cloud Compute Engine**, enabling scalable processing of high-frame-rate rodent behavior videos and improving annotation efficiency by **70%**.
- Built a deep learning pipeline integrating **CNNs and LSTM architectures in PyTorch**, trained on **Google Cloud AI Platform**, for pose estimation and sequential behavior classification, achieving **92% accuracy** on labeled mouse interaction data.
- Applied transfer learning with a **U-Net model using Keras and TensorFlow**, trained on **GCP TPUs**, to detect fine-grained social behaviors (Ano-genital sniffing, nose-to-nose contact), achieving an **F1-score of 0.92** on custom datasets.
- Built baseline vs. fine-tuned model comparisons for pose estimation pipelines, helping quantify the added value of deep architectures (CNNs, LSTMs) over traditional models.
- Designed an automated behavioral analysis workflow using **scikit-learn, ffmpeg, and NumPy**, executed via **Cloud Composer** to streamline preprocessing, inference, and post-processing in a reproducible and scalable environment.

Pelotonia Research Center – Cancer Research Columbus, OH
Data Science Intern Jan 2023 – May 2024

- Built and automated end-to-end **ETL pipelines** using **Python, SQL, Apache Airflow, and Pandas** on **Azure Data Factory**, to ingest, clean, and integrate large-scale structured and unstructured **EHR and genomic datasets**, improving data accessibility and reporting efficiency by **50%**.
- Developed **medical image registration pipelines** using **MONAI, PyTorch, and SimpleITK**, deployed on **Azure Machine Learning Compute**, applying **diffusion models, affine, and deformable transformations** to achieve **97% alignment accuracy** across histopathology slides.
- Developed **classification and survival analysis models** using **scikit-learn, XGBoost, and Lifelines** on **Azure ML Studio** to predict patient outcomes and treatment response; performed **Kaplan-Meier analysis** and **log-rank tests** to identify survival patterns.
- Created interactive **Tableau dashboards** hosted via **Azure App Service**, to visualize survival curves, treatment efficacy, and patient stratification, enabling researchers and clinicians to explore model outputs and key clinical insights.

Cognizant Technology Services Chennai, India
Machine Learning Engineer Mar 2021- Dec 2022

- Built and deployed **risk prediction models** using **LightGBM and pandas** on **Amazon SageMaker** to classify high-risk insurance policies, enhancing fraud detection and underwriting decisions with a **30% improvement in accuracy** across 100K+ policy records.
- Deployed ML models as **REST APIs using FastAPI**, containerized with **Docker**, and hosted on **AWS ECS**, while managing experiment tracking and version control via **MLflow on S3**, streamlining the MLOps lifecycle in a regulated insurance environment.
- Designed and orchestrated **scalable data pipelines** using **Databricks (PySpark)** and **dbt**, integrated with **AWS Glue and Redshift**, automating ingestion and transformation of **1M+ insurance records** and reducing model retraining time by **40%**, supporting real-time policy risk scoring workflows.
- Conducted comparative benchmarking of machine learning algorithms (**LightGBM, XGBoost, Logistic Regression**) using cross-validation and **AUC/F1** metrics to determine optimal model for policy risk classification.
- Built and automated end-to-end **ETL pipelines** using **Python, SQL, Apache Airflow, and Pandas** on **Azure Data Factory**, to ingest, clean, and integrate large-scale structured and unstructured **EHR and genomic datasets**, improving data accessibility and reporting efficiency by **50%**.

Skills

- **Languages:** Python, R, Java, JavaScript, C, D3.js, HTML/CSS
- **ML & DL:** Scikit-learn, XGBoost, LightGBM, Lifelines, H2O, PyTorch, TensorFlow, Keras, MONAI, CNNs, LSTMs, U-Net, LBPH, Transfer Learning, PyTorch Geometric
- **LLM & NLP:** OpenAI (GPT-4, ChatGPT APIs), Google PaLM, LLaMA 2, Mistral, LangChain, Retrieval-Augmented Generation (RAG), Prompt Engineering, ChromaDB, FAISS, Pydantic, TextBlob
- **Data Analytics & Visualization:** Pandas, NumPy, Matplotlib, Seaborn, Plotly, Tableau, Metabase, Exploratory Data Analysis (EDA), Kaplan-Meier Analysis, Log-rank Test
- **Data Engineering & Pipelines:** Apache Airflow, dbt, SQL, MySQL, Azure Data Factory, AWS Glue, ETL Pipelines, Data Cleaning, Data Integration, Data Modeling
- **MLOps & Model Deployment:** FastAPI, Flask, Docker, MLflow, REST APIs, AWS Lambda, AWS ECS, AWS Fargate, Amazon SageMaker, Azure ML Studio, Azure ML Compute, Google Cloud AI Platform, Cloud Composer
- **Computer Vision & Image Processing:** OpenCV, SimpleITK, Dlib, Haar Cascades, FaceNet, Image Registration (Affine, B-spline, Deformable, Diffusion Models), High-Frame-Rate Video Analysis, SLEAP
- **Cloud Platforms:** AWS (Lambda, ECS, Fargate, S3, SageMaker, Glue, Redshift), Azure (App Service, ML Studio, Data Factory, ML Compute), GCP (AI Platform, Composer, Compute Engine, TPU)
- **Databases & Storage:** MySQL, PostgreSQL, Oracle SQL, SQL Server, NoSQL, DynamoDB, Redshift, Amazon S3, Elasticsearch, Neo4j
- **Testing & Optimization:** Cross-Validation, Hyperparameter Tuning (Grid Search, Random Search), Model Benchmarking, F1-Score, Accuracy, Precision
- **Frameworks/Libraries:** Flask, FastAPI, Node.js, Express.js, React.js, Angular.js, Apache Spark, MLlib, Airflow, Hadoop
- **Data Governance & Privacy:** HIPAA compliance, PII handling, regulated data workflows (Insurance & Healthcare)
- **Certifications:** Microsoft Azure Cloud Fundamentals, Azure AI Fundamentals, Google – Python for Data Science Certification, CyberArk Trustee Certification, NPTEL – Java, Machine Learning

Projects

AI-Powered Business Idea Evaluation System:
Built an end-to-end pipeline to evaluate 500+ business ideas using LLMs (LLaMA 2, Mistral, ChatGPT APIs) and classical ML models; performed text preprocessing, EDA, and model benchmarking with Logistic Regression and Decision Trees, achieving over 85% agreement with expert-labeled criteria through cross-validation and hyperparameter tuning.

Text-to-SQL App:
Built a **Streamlit application** that translates natural language into SQL queries using **LangChain and Google PaLM**, enabling seamless interaction with **MySQL databases**. Integrated prompt chaining, query execution, and result visualization to support real-time data exploration for business users.

Smart Mobility - Smart-Mobility-GNN
Built a GNN-based model using **PyTorch Geometric** on the **Open Traffic Dataset**, modeling road networks as graphs for route optimization and traffic prediction; achieved a **15% improvement in path prediction accuracy** over baseline models.

Face-X - Face Recognition
Developed a high-accuracy facial recognition attendance system using OpenCV, Haar Cascade, and LBPH, achieving 98% accuracy in real-time identification enhanced with Dlib and FaceNet for robust feature extraction. **Presented and Published at ICIVC 2022**.

Education

Master of Science in Computer Science and Engineering – Ohio State University – CGPA: 3.53 Dec 2024
Coursework: Neural Networks, Data Mining, Fairness in Artificial Intelligence and Databases, Data Visualization, Parallel Computing, Advanced OS, Cybersecurity

Bachelor of Technology in Computer Science and Engineering – Anna University – CGPA: 4 June 2021
Coursework: Problem-Solving, Object-Oriented Design, Data Structures & Algorithms, Databases, Operating Systems, System Design, Networking, Machine Learning