# Saikiranmansa Sunnam

ssaikiranmansa2023@gmail.com 7164150744 Buffalo Linkedin Portfolio Hugging Face

## PROFESSIONAL EXPERIENCE

**Machine Learning Engineer,** *Rivach LLP* 07/2021 – 06/2023

- **Designed** and implemented a geo-personalized gig recommendation engine using spatial clustering and collaborative filtering, resulting in a **42% improvement in task-user match accuracy** across the Brekrr platform.
- **Engineered** a robust identity verification system using Siamese neural networks and OpenCV, **reducing fake profile creation by 35%** and enhancing overall user trust.
- **Developed** predictive maintenance algorithms leveraging XGBoost and Random Forest, achieving **60% forecast accuracy** and decreasing equipment downtime by **28%** within the Landmanager ecosystem.
- **Spearheaded** the deployment of a content-based recommender system powered by matrix factorization techniques, leading to a **50% increase in user engagement** on Guiding Young Minds.
- **Implemented** Transformer-based NLP moderation pipelines for real-time chat analysis and policy enforcement, **cutting manual moderation overhead by 60%** while ensuring regulatory compliance.
- **Optimized** ML deployment pipelines using FastAPI and Docker within a microservices architecture, maintaining **sub-200ms latency** for high-frequency API interactions.
- **Automated** scalable inference systems using AWS Lambda and Firebase Functions, **reducing cloud compute expenses by 20%** and improving deployment agility.
- **Established** comprehensive monitoring solutions via MLflow and Prometheus, delivering **99.9% model uptime** and real-time visibility into production model health.

**Data Scientist,** *Rivach LLP* 05/2020 – 06/2021

- Built and maintained **automated data pipelines** using **SQL** and **Apache Airflow**, implementing basic data modeling and validation to process ~50K records/month from user activity logs.
- Developed a fraud detection model using Logistic Regression and Decision Trees, improving suspicious activity flagging by ~35% based on real-time behavior patterns.
- Conducted **A/B testing** on task notification timing, leading to a 10–15% increase in daily engagement; tracked experiments and results using MLflow and Jupyter Notebooks.
- Collaborated with product and ops teams to define KPIs and delivered weekly performance dashboards using **Tableau**, streamlining insights for business decisions.

## TECHNICAL PROJECTS

**Domain-Specific QA System Using DeepSeek and RAG**

- Built a production-grade QA system using **DeepSeek**, **FAISS**, and **RAG**, improving response accuracy by 30%+ and boosting **BLEU/ROUGE** scores by 15–20% with Sentence-BERT embeddings.
- Indexed domain-specific documents using FAISS and Elasticsearch, enabling <250ms retrieval time and enhancing semantic relevance for context-aware queries.
- Deployed a scalable pipeline on AWS using **FastAPI**, **Docker**, and **MLflow**, reducing retrieval latency by 50% while ensuring robust monitoring and uptime.

**Advanced Anomaly Detection and Text Classification Using Deep Learning**

- Engineered a deep learning framework with **autoencoders** achieving $R^2 = 0.9916$, detecting 25–74 anomalies across 5,315 time-series records using PyTorch.
- Improved AG News classification accuracy from 90.08% to 90.53%, and reduced overfitting by 12% using L2 regularization, dropout, and learning rate tuning.
- Validated models using precision, recall, F1-score, confusion matrix, and ROC AUC (0.91) with data preprocessing and visualization via Matplotlib.

**LLaMA-Based Sentiment Analysis with LLaMA2 & LLaMA3**

- Fine-tuned LLaMA2 for 3-class sentiment classification with 92% accuracy and LLaMA3 for binary sentiment analysis on a 10,000-sample dataset, boosting F1-score by 10%.
- Applied 3 optimization techniques—learning rate scheduling, decoding strategies, and attention masking—to reduce overfitting by 12% and improve training stability.
- Deployed models with <200ms inference latency using Gradio and batching, achieving 30% reduction in response time and enabling real-time prediction.

## EDUCATION

**State University of New York at Buffalo,** *Master of Science in Computer Science* 12/2024 | Buffalo, NY

**Courses:** Machine learning, Deep learning, Computer Vision & Image Processing, Operating Systems, Algorithms Analysis and Design, Data Intensive Computing, Computer Security, Data Mining and Query Language, Software Engineering

## SKILLS

**Core ML & AI:** Generative AI, LLMs (LLaMA, GPT, BERT), RAG, NLP, Computer Vision, Anomaly Detection
**Frameworks & Libraries:** PyTorch, TensorFlow, Scikit-learn, Hugging Face, Keras
**Programming:** Python, SQL, R, Java, C
**Data & Visualization:** Pandas, NumPy, Matplotlib, Seaborn, Plotly
**DevOps & Cloud:** Docker, FastAPI, REST APIs, AWS (EC2, S3, Lambda, SageMaker, DynamoDB, CloudWatch), CI/CD
**Big Data & Databases:** Hadoop, Spark, MySQL, Oracle Database
**Tools & Collaboration:** Git, Jupyter Notebooks, Data Warehousing, Testing, Code Reviews, Documentation

## PUBLICATIONS

**Granite classification using machine learning and edge computing**
Published in **F1000Research**: Link