Manan Shah

Jersey City, NJ | mananshah1703@gmail.com | +1 (201) 936-7141 |

Professional Experience

Symphony AI, California, United States | *Capstone Project, Data Scientist* **Sep 2024 - Present**

- Designed and deployed an end-to-end multi-stage AI pipeline for P&ID digitization, integrating YOLOv11 for symbol detection (92% mAP), Azure document intelligence for text extraction, and Hough Transforms via OpenCV for line detection—achieving an 80% reduction in manual effort and significant cost savings
- Constructed graph representations of diagrams using GraphSAGE-based GNNs and implemented a Siamese ResNet trained with contrastive loss to classify unseen symbols based on latent space similarity.
- Boosted graph reconstruction F1-score by 15% over the baseline through iterative model tuning, pipeline modularization, and noise-tolerant edge prediction strategies.

New York University, New York, United States | *Graduate Research Assistant* **May 2024 - Present**

- Fine-tuned LLaMA-7B using QLoRA on NYU's HPC cluster using slurm with six custom corpora using Hugging Face Transformers, PEFT, and bitsandbytes, optimizing for low-rank adaptation and efficient GPU memory usage.
- Conducted large-scale benchmarking of ChatGPT, Claude, Mistral, Gemini, Perplexity, DeepSeek via their APIs, orchestrating multi-model prompt pipelines with LangChain, and managing data flow using Pandas and SQLite.
- Analyzed model stability using sentence-transformers to compute embedding similarity across runs, and performed topic modeling on generated outputs using NLTK, spaCy, and LDA via Gensim.
- Co-authored and visualized findings in "Ethical Logic in Six Large Language Models", accepted at IEEE Conference on Artificial Intelligence (CAI 2025, Santa Clara), including cross-model trends, value alignment, and output variability.

Morgan Stanley Capital International (MSCI), Mumbai, India | *Data Science Research Intern* **Jan 2022 - Jun 2022**

- Designed and implemented scalable ETL pipelines using Apache Airflow and Snowflake SQL, processing financial data across investment strategies, reducing data processing time by 30% and improving data quality through automated data validation.
- Developed complex data extraction and transformation workflows in Snowflake SQL, creating modular data models that automated index methodology calculations, resulting in 22% reduction in tracking error and enabling near-real-time index performance reporting for 15+ investment product lines.
- Engineered end-to-end data integration solution leveraging Snowflake's data sharing and zero-copy clone capabilities, enabling real-time index verification for 20+ equity and fixed income indexes, enhancing process efficiency by 10% and reducing manual validation time from 5 days to 3.5 days through automated data reconciliation and advanced statistical validation techniques.
- Optimized machine learning data preprocessing pipelines, implementing advanced feature engineering and data cleaning strategies that refined 5 core algorithms, enhancing model accuracy by 18% and reducing false positives by 25%, ultimately improving risk assessment precision from 88% to 95%.

Projects

Indian Classical Raga Identification System

- Built a deep learning pipeline to classify audio into seven Indian ragas (Bageshree, Bhairavi, Sarang, Malkauns, Yaman, Todi, Bhoopali) with 93.6% accuracy, using MFCCs extracted from live concert recordings and an LSTM-RNN architecture implemented in Keras.
- Deployed the trained model via a Django REST API, enabling real-time inference and integration with external applications for automated raga recognition.

Driver Alertness and Awareness Monitoring System

- Developed a real-time computer vision system using YOLOv3, PINet, CNN, and Random Forest for driver behavior classification across nine distraction categories. Achieved 95% accuracy in drowsiness detection using eye-mouth ratio analysis, and 98% accuracy in lane detection with a Lab Color Space–based segmentation algorithm using OpenCV and NumPy.
- Co-authored a peer-reviewed research paper on the system, accepted at the International Conference on Computing, Communication, and Network Technologies (ICCNT), showcasing applied deep learning, image processing, and behavioral analysis for real-world safety applications.

Skills

Programming Languages: Python, SQL, Java, R, C
Machine Learning & Deep Learning: Scikit-learn, PyTorch, TensorFlow, Keras, Decision Trees, Linear Regression, SVM
Natural Language Processing & Computer Vision: HuggingFace Transformers, LangChain, OpenCV, CNN, YOLO, NLTK, spaCy, Gensim, Pillow, Azure OCR
Version Control & Collaboration: Git, Github, Jira, Confluence
Big Data, Data Engineering & Data Visualization: Hadoop, HPC (Slurm), Apache Spark, Apache Airflow, Pandas, NumPy, MySQL, Matplotlib, Seaborn, Tableau, Power BI, GCP, Snowflake
Statistical Analysis & Tools: Hypothesis Testing, Regression Analysis, A/B Testing, Exploratory Data Analysis, Correlation Analysis, Predictive Modeling, Clustering, Data Modeling

Education

New York University Center For Data Science, New York, NY **Aug 2023 - May 2025**
Masters of Science in Data Science *GPA: 3.78/4*
Courses: Probability & Statistics, Big Data, Intro to Data Science, NLP, Machine Learning, Reinforcement Learning

Sardar Patel Institute of Technology, Mumbai, India **Aug 2019 - May 2023**
Bachelor of Technology in Computer Engineering *GPA: 8.95/10*
Courses: Algorithms, Data Structures, Software Engineering, Linear Algebra, Database Management Systems