Vikas Reddy

Machine Learning Engineer — LLMs, Multimodal & Real-Time Inference

(443)-992-2554 | wh22215@umbc.edu | linkedin.com/in/vvikasreddy | github.com/vvikasreddy | Portfolio

Education

University of Maryland Baltimore County (UMBC) — Baltimore, MD
Master's in Data Science; GPA: 4.0 — *Aug 2023 – May 2025*

Institute of Aeronautical Engineering (IARE) — Hyderabad, India
Bachelor's of Technology in Computer Science and Engineering; GPA: 3.41 — *May 2019 – May 2023*

Research Publications

Deep fusion of neurophysiological and facial features for enhanced emotion detection. (n.d.). IEEE Journals & Magazine — IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/10945364

Experience

Brain-Machine Interface Lab, UMBC — Baltimore, USA
ML Researcher – Multi-Modal & Real-Time Systems — *Jan. 2024 – Present*

- Developed a scalable multi-modal learning pipeline combining Transformers and ConvLSTM for EEG + facial feature analysis, achieving 97% accuracy. Integrated optimized attention mechanisms for streaming inference.
- Built EmoFormer, a SegFormer-inspired deep learning model for facial emotion classification, leveraging low-level feature extraction and efficient preprocessing pipelines (Scikit-learn). Benchmarked 77.34% on FER2013 and 67.71% on AffectNet.
- Engineered a real-time EEG-based deception detection system with Apache Kafka and CNNs, achieving 81% streaming accuracy. Employed low-latency processing optimized for edge inference.
- Led the design of a structured EEG data acquisition protocol (25 participants) and applied augmentation strategies to address signal imbalance, laying groundwork for retrieval-based attention modeling.

Epam Systems — Hyderabad, India
SDE Intern – Backend & Cloud Distributed Systems — *Jan. 2023 – June 2023*

- Built scalable backend services using Python/Java and containerized microservices with Kubernetes and Docker, supporting high-throughput production environments.
- Designed CI/CD pipelines with Jenkins and Docker, automating deployment and reducing test/build times by 35%. Contributed to latency-aware system diagnostics.
- Developed ETL workflows with AWS Lambda, S3, and Athena to process 50GB+ data efficiently; incorporated distributed data transformation pipelines.
- Followed SOLID and design patterns to design fault-tolerant services.

IndicWiki Project — Hyderabad, India
Intern – Data Analyst for NLP Systems — *Mar. 2022 – June 2022*

- Built multilingual data pipelines to ingest, clean, and normalize 10,000+ Wikipedia articles, enabling downstream NLP models for machine translation and low-resource language tasks.
- Streamlined ETL processes using Python/SQL, reducing processing time by 30%, and supporting scalable retrieval-based translation models.
- Conducted analytical evaluation of translated outputs using visualization (Tableau, Matplotlib) and statistical metrics to improve data quality for inference.

Projects

Stock Price Forecasting System | *Python, TensorFlow, LSTM, NumPy, Matplotlib* — GitHub

- Developed a time-series forecasting model using **LSTM** to predict stock prices, achieving 84% of predictions within ±10% error. Applied data normalization, cross-validation, and hyperparameter tuning to boost model accuracy.

Lexically Constrained Neural Translation Engine | *Python, PyTorch, Hugging Face Transformers* — GitHub

- Built a custom **beam search decoder** to enforce lexical constraints during transformer-based neural translation. Improved **BLEU scores** on WMT TR-EN data by guiding model outputs across 300+ test scenarios.

Technical Skills

Programming Languages: Python, SQL, Java, C++, Bash
Deep Learning & LLMs: Transformers, PyTorch, TensorFlow, Hugging Face, Scikit-learn, ConvLSTM, SegFormer
Retrieval & Ranking Systems: Apache Kafka, Real-Time Streaming, RAG, Ranking Pipelines, Attention Mechanisms
Cloud & DevOps: AWS (Lambda, S3, Athena), Docker, Kubernetes, Jenkins, Git, CI/CD
Data Engineering & Visualization: ETL Pipelines, PySpark, Pandas, NumPy, Tableau, Matplotlib, SQL Databases