Sai Dhiren Musaloji

musalojidhiren@gmail.com | +1 862-423-8830 | LinkedIn | GitHub | Portfolio

Education

New Jersey Institute of Technology - MS Data Science May 2025 | GPA: 3.85/4.0

Mahatma Gandhi Institute of Technology - B. Tech Electronics Engineering Jun 2023 | GPA: 3.00/4.0

Technical Skills

Machine Learning: TensorFlow, PyTorch, Scikit-learn, XGBoost, Neural Networks, Deep Learning, Reinforcement Learning
ML Frameworks: Transformer Models, LSTM, CNN, Graph Neural Networks, Actor-Critic Methods, Policy Gradients
Programming: Python, SQL, REST APIs, Flask, NumPy, Pandas, Matplotlib, Seaborn
Cloud & MLOps: AWS (EC2, S3, SageMaker), Azure ML, Docker, MLflow, Model Deployment, A/B Testing
Data Processing: Feature Engineering, Data Validation, ETL Pipelines, Apache Spark, Hadoop
Specialized: NLP, Computer Vision, Audio Processing, Retrieval-Augmented Generation, Vector Databases

Professional Experience

AI Engineering Intern | Tech Mahindra, Makers Lab | Pune, India Oct 2023 – Jan 2024

- NLP Pipeline Development: Built text classification system for multilingual content, implementing comprehensive preprocessing workflows including tokenization, cleaning, and regex-based filtering techniques
- Data Quality Engineering: Designed validation frameworks to handle noisy text data, achieving 20% improvement in content quality through systematic outlier detection and statistical analysis
- Production Integration: Collaborated with research team to document and deploy models for enterprise integration, gaining experience in production ML workflows and cross-functional communication

Projects & Research

AI Research Assistant with RAG - End-to-End ML System

- **API Integration:** Developed research assistant using Google Gemini 1.5 Flash API processing 10,000+ queries daily, achieving 2.3-second average response time through optimized connection pooling and caching strategies
- **Data Pipeline:** Created multi-source processing system extracting content from 25,000+ articles using newspaper3k and serper.dev APIs, maintaining 99.2% success rate with 100 requests/minute rate limiting
- **Vector Search:** Deployed semantic similarity search indexing 500K+ document embeddings using scikit-learn NearestNeighbors, optimizing retrieval to sub-50ms through dimensionality reduction techniques
- **Transformer Workflow:** Orchestrated RAG pipeline with BART, BERT, and DistilBERT models processing 10,000+ documents/hour, enabling batch embedding generation with automated model versioning

Deep Reinforcement Learning for Lunar Lander - Autonomous Control Systems

- **Actor-Critic Training:** Trained spacecraft landing agent using policy gradient methods in OpenAI Gym environment, achieving consistent 250+ point scores (target: 200) across 1,000+ landing simulations
- **Neural Architecture:** Constructed dual-network system with experience replay buffer (10K samples) and target networks, stabilizing training convergence within 500 episodes through gradient clipping and learning rate scheduling
- **Performance Optimization:** Delivered 95% successful landing rate through reward shaping and trajectory optimization, reducing training time by 40% via vectorized environment processing across 16 parallel instances
- **Continuous Control:** Mastered precise landing maneuvers with 2.5-meter average landing accuracy, demonstrating robust performance across varying wind conditions and fuel constraints

Transformer-Based Speaker Classification - Audio ML System

- **Speaker Identification:** Developed transformer encoder classifying 600 unique speakers from variable-length audio sequences, achieving 94.2% accuracy on test set of 50,000+ audio samples
- **Audio Processing Pipeline:** Established efficient data loading system processing MFCC features with custom positional encoding, handling 10GB+ audio datasets with 3x faster training through mixed precision and gradient accumulation
- **Model Architecture:** Constructed multi-head self-attention mechanism (8 heads, 512 dimensions) optimized for audio processing, surpassing RNN baselines by 18% accuracy through dropout (0.3) and label smoothing techniques
- **Production Deployment:** Scaled inference system processing 5,000+ audio files/hour with 150ms average classification time, utilizing batch processing and model quantization for real-time speaker recognition

BERT-Based Reading Comprehension - Transfer Learning, Model Fine-tuning

- **Model Fine-tuning:** Fine-tuned BERT model for extractive question answering on SQuAD 2.0 dataset (130K+ examples), implementing span prediction with custom loss functions that achieved 89.3% F1 score
- **Performance Enhancement:** Surpassed baseline approaches by 15% through bidirectional context understanding and gradient accumulation techniques across 48 training epochs

Certifications

- Microsoft: Azure AI Engineer Associate
- Microsoft: Azure Data Scientist Associate