Tirumala Teja Yegineni

3235 Heritage Trl, Denton, Texas 76201

940-977-0152 yegineni2002@gmail.com LinkedIn GitHub Portfolio

Summary

- AI/ML Data Engineer with 3+ years of hands-on experience designing, deploying, and maintaining production-grade machine learning systems across real-world use cases.
- Proven expertise in end-to-end ML pipelines, covering data extraction,data ingestion, feature engineering, model training, evaluation, explainability, and deployment with a strong focus on reproducibility and reliability.
- Strong background in deep learning and large-scale data processing, leveraging PyTorch, TensorFlow, PySpark, Azure Databricks, Data Lake, ETL Pipelines and modern MLOps workflows using MLflow, DVC, Docker, and CI/CD.
- Hands-on experience deploying low-latency models on Azure Machine Learning and AWS SageMaker, with emphasis on model interpretability using SHAP and explainable AI practices.
- Actively building Generative AI and Agentic AI solutions, including LLMs, RAG pipelines, and AI agents using LangChain, LlamaIndex, and CrewAI, with a strong interest in scalable ML platforms and MLOps.

Education

University of North Texas

Master of Science in Data Science *TX, USA*

Relevant Coursework

- Machine Learning
- NLP
- Deep Learning
- Database Management
- Artificial Intelligence
- Generative AI
- Systems Programming
- MLOPs

Experience

University of North Texas **September 2024 – Present**

AI/ML Engineer *Dallas, Texas*

- Designed and deployed a production-grade, domain-specific Generative AI platform to enable intelligent querying across university web pages and documents, reducing information retrieval time by ≈50%
- Built scalable FastAPI microservices for retrieval, inference, and orchestration, achieving sub-second semantic search latency on datasets exceeding 10K+ documents.
- Applied machine learning models for query intent classification and confidence scoring, reducing off-topic and low-confidence responses by 25% and improving system reliability.
- Implemented model explainability using SHAP to analyze feature contributions and decision behavior for intent classification and scoring models, improving transparency and explainable AI
- Established end-to-end MLOps workflows using MLflow, DVC, Docker, and CI/CD pipelines, improving experiment reproducibility and significantly reducing deployment effort.
- Collaborated with faculty and research teams to validate system accuracy, usability, and real-world applicability, delivering a deployable AI feature suitable for direct integration into a university website.
- Created model monitoring dashboards to track data drift, feature stability, and prediction confidence thresholds.

Velagapudi Ramakrishna Siddhartha Engineering College **September 2022 – October 2023**

Deep Learning Researcher *Vijayawada, India*

- Designed and developed an end-to-end short-range rainfall forecasting system using LSTM time-series models, achieving 18–25% RMSE reduction over statistical and traditional ML baselines.
- Built a reproducible ML pipeline covering data ingestion, preprocessing, feature engineering, training, and evaluation, leveraging DVC for dataset versioning to reduce data inconsistencies across experiments.
- Tracked and managed 20+ model experiments using MLflow, logging hyperparameters, metrics (MAE, RMSE), and artifacts to improve model comparison and selection efficiency.
- Applied model explainability techniques to analyze lag features, rolling statistics, and weather signals, improving transparency and reducing debugging time by 25%.
- Containerized training, inference, and monitoring components using Docker, and implemented CI/CD pipelines with GitHub Actions to automate testing, validation, image builds, and deployments.
- Deployed models via a FastAPI inference service with a Streamlit dashboard, integrating Prometheus and Grafana to monitor latency, throughput, and performance trends in production-like environments.

Velagapudi Ramakrishna Siddhartha Engineering College **November 2021 – August 2022**
Machine Learning Engineer *India*

- Contributed to applied machine learning projects focused on real-world agricultural decision-making as part of an academic AI/ML research lab.
- Designed and deployed a sugarcane price and yield forecasting system, leveraging historical crop prices and rainfall data to support data-driven planning for farmers.
- Performed end-to-end data preprocessing and feature engineering, training a Decision Tree Regressor to capture non-linear relationships and improve forecast accuracy by 18–22% over baselines.
- Applied SHAP to explain model predictions, enabling clear interpretation of key drivers and improving trust among non-technical stakeholders.
- Built an interactive Streamlit dashboard for visualizing forecasts and trends, and exposed predictions through a FastAPI-based inference service.
- Containerized the complete ML pipeline using Docker, gaining hands-on experience across the full ML lifecycle from modeling and evaluation to explainability and deployment.

Projects

Multi-Agent Collaboration for Research Tasks |*Agentic AI, LLMs, RAG, LangChain, Azure OpenAI, LlamaIndex* **2025**

- Designed and implemented a multi-agent research assistant using LLMs to automate information retrieval, synthesis, and validation through agent-based workflows.
- Built a 3-agent pipeline (Research → Summarize → Fact-check) using CrewAI, improving response reliability and reducing hallucinations via structured agent collaboration.
- Integrated Retrieval-Augmented Generation (RAG) using LangChain and LlamaIndex for document ingestion, indexing, and retrieval, enabling grounded, source-aware research outputs.
- Developed modular agent components for planning, retrieval, execution, and response generation, enabling reusable and extensible agent orchestration across research tasks.
- Applied agentic AI principles such as task decomposition, iterative reasoning, prompt design, and tool usage to evaluate and improve relevance, consistency, and factual accuracy.

AI-Driven Plant Disease Detection System | *Python, PyTorch, TensorFlow, CNN, Deep Learning, FastAPI* **2024**

- Developed an end-to-end deep learning solution to classify potato plant diseases using Convolutional Neural Networks (CNNs), supporting early disease detection for agricultural use cases.
- Collected, cleaned, and preprocessed plant image data, applying data augmentation and normalization techniques to improve model robustness and generalization across disease classes.
- Trained and evaluated CNN models using PyTorch/TensorFlow, iterating on network architecture and hyperparameters to improve classification performance.
- Evaluated CNN model performance using classification accuracy and training/validation loss, and deployed the optimized model as a FastAPI-based inference service to enable low-latency, API-driven disease detection for real-world usage.

Spotify Analytics Platform for User and Artist Insights
| *PySpark, Databricks, Delta Lake, AWS Cloud, Snowflake, PowerBI* **2020**

- Designed and built a scalable Apache Spark / PySpark data pipeline on Azure Databricks, processing large-scale Spotify streaming data stored in AWS S3 to analyze user listening behavior, track popularity, and artist performance.
- Implemented end-to-end ETL workflows using AWS Glue and PySpark, applying filtering, joins, aggregations, schema optimization, and data quality checks to produce analytics-ready datasets.
- Orchestrated and monitored data processing jobs using AWS Lambda and Amazon CloudWatch, enabling automated execution, logging, and operational visibility across the pipeline.
- Loaded curated datasets into Snowflake and enabled downstream analytics using SQL and Amazon Athena, with insights visualized through Power BI dashboards for trend and performance analysis.

Technical Skills

Programming: Python, SQL, Bash
Machine Learning & Deep Learning: PyTorch, TensorFlow, Scikit-Learn, NumPy, Pandas, Matplotlib, Seaborn
Generative AI/NLP: LLMs (e.g., GPT, Llama), RAG, Agentic AI Workflows, LlamaIndex, LangChain, Hugging Face, Transformers, MOdel Context Protocol(MCP)
Data Engineering & MLOps: PySpark, Databricks, Delta Lake, MLflow, DVC, Airflow
Cloud & Model Deployment: Azure (Azure ML, Azure Databricks, Azure OpenAI, Blob Storage), AWS (SageMaker, S3, EC2, Lambda, IAM, Glue, Athena, Cloudwatch), Docker, FastAPI, Streamlit, REST APIs
Model Monitoring & Version Control: Git, GitHub, Azure DevOps, CI/CD Pipelines(Github Actions),Prometheus, Grafana, Snowflake, Power BI
Tools & Environments: Jupyter Notebook, VS Code, Cursor AI IDE, claude desktop