VIHAS ADI

AI Research Engineer — Machine Learning Engineer — Generative AI Engineer

347-665-6612 — adivihas@gmail.com — LinkedIn — Portfolio — New York, NY, USA

Summary

AI Engineer and Master's student in Artificial Intelligence with hands-on experience building deep learning and generative AI systems. Strong background in PyTorch, diffusion models, transformers, and LLM-based pipelines including RAG and prompt engineering. Experienced across the end-to-end ML lifecycle, including data preprocessing, model training, evaluation, production inference, deployment, and experiment tracking using Docker, Git, MLflow, and FastAPI. Actively seeking AI Engineer or Machine Learning Engineer roles. Hands-on experience designing and deploying LLM-powered applications using retrieval-augmented generation workflows.

Experience

AI Research Engineer (Academic & Independent Projects)
Yeshiva University, New York, USA — Jan 2024 – Present

- Built and trained diffusion-based deep learning models for radiotherapy dose prediction using PyTorch and DoseDiff
- Integrated MambaVision-based architectures and evaluated model performance using clinical dose metrics and DVH analysis
- Implemented LLM-powered data ingestion and RAG pipelines for automated research paper retrieval and summarization using LangChain
- Executed end-to-end model training, validation, debugging, and ablation studies to assess architectural and loss-function trade-offs
- Designed reproducible ML pipelines with experiment tracking using Git, Docker, and MLflow
- Deployed trained models for production-style inference using FastAPI and Docker on cloud infrastructure, enabling scalable and low-latency model serving

Machine Learning Intern
OctaZen Software Solutions Pvt. Ltd., Hyderabad, India — May 2022 – Jun 2022

- Built and optimized supervised machine learning models using Python and Scikit-learn on real-world datasets
- Designed robust data preprocessing pipelines, including missing-value handling, feature scaling, and categorical encoding
- Evaluated and compared models using multiple performance metrics to assess generalization and stability
- Applied end-to-end ML workflows covering analysis, training, validation, and interpretation
- Performed systematic model and feature comparisons to select configurations with the best validation performance

Education

M.S. in Artificial Intelligence
Yeshiva University, New York, NY, USA — Jan 2024 – Dec 2025

B.Tech in Computer Science Engineering
TKR College of Engineering & Technology, Hyderabad, India — Aug 2019 – 2023

Technical Skills

Programming	Python, SQL
Machine Learning & AI	Machine Learning, Deep Learning, Generative AI, LLMs, NLP, Computer Vision, RAG
Model Development	PyTorch, TensorFlow, Scikit-learn, CUDA, hyperparameter tuning
Data Pipelines	Data preprocessing, normalization, augmentation, feature engineering
Evaluation	RMSE, MAE, Dice, IoU, DVH analysis, benchmarking, ablation studies
MLOps & Deployment	Git, Docker, FastAPI, MLflow, model versioning
Cloud & Databases	AWS (EC2, S3), MySQL, MongoDB

Projects

AI Research Paper Summarizer (LangChain, RAG, LLMs)

- Designed an end-to-end LLM-powered pipeline to retrieve research papers from arXiv, parse PDFs, and generate structured summaries and peer-style reviews
- Built a RAG-based summarization system using recursive text chunking and map-reduce chains for long-form documents (20–40+ pages)
- Implemented section-aware summarization (Abstract, Methods, Results, Conclusion) improving factual alignment
- Developed peer-review generation modules producing structured JSON outputs for downstream evaluation
- Exposed inference via FastAPI and containerized the application using Docker
- Deployed trained models for production inference using FastAPI and Docker on AWS EC2, enabling scalable, low-latency model serving

Medical Dose Diffusion Model (PyTorch, DoseDiff, MambaVision)

- Developed and benchmarked DoseDiff, CT-Mamba, and hybrid architectures for radiotherapy dose prediction using the Open-KBP dataset (50 patients) under identical experimental settings
- Demonstrated that CT-Mamba achieved the best clinical performance after extended training, reducing Dose Score from 12.38 to 1.56 and DVH error from 7.37 to 0.71 over 670 epochs
- Showed that extended training (670 epochs) had a larger impact than dataset scaling, outperforming models trained on more data but fewer epochs
- Identified that naive hybrid integration (DoseDiff + CT-Mamba) degraded performance, revealing optimization instability and feature redundancy in complex architectural fusion

Brain Tumor Segmentation (Deep Learning, Computer Vision)

- Evaluated U-Net, Attention U-Net, FCN, and SegNet for brain tumor MRI segmentation, with FCN achieving the best generalization (Dice = 0.75, IoU = 0.66) on limited data.
- Found that attention mechanisms increased pixel accuracy to 99%, while Dice/IoU improvements were inconsistent, indicating a tradeoff between localization precision and region overlap.
- Observed overfitting in deeper architectures under data-constrained settings, leading to weaker validation performance compared to simpler models.
- Concluded that data characteristics and metric selection had greater impact than architectural complexity, guiding future model design for medical imaging tasks.

Customer Churn Prediction (Performance Improvement)

- Improved baseline customer churn prediction performance by 20–25% (MAE/RMSE) by transitioning from Linear/Ridge models to Gradient Boosting.
- Engineered behavioral features (total spend, average spend, purchase frequency), leading to an additional 10–15% reduction in prediction error.
- Applied grid search and cross-validation to optimize model hyperparameters, improving stability and generalization across validation folds.

Ride Demand Prediction System (Machine Learning, Time Series)

- Developed forecasting pipelines using temporal and spatial features across 12+ months of data
- Achieved RMSE reduction of 12–18% compared to naive seasonal baselines
- Designed the system to support operational planning and resource allocation, enabling identification of peak-demand windows and high-utilization zones
- Translated model outputs into actionable insights, helping prioritize driver allocation and capacity planning for high-demand periods

Certifications

- Google Cloud Generative AI Certificate — Google Cloud Skills Boost
- AWS Machine Learning Foundations — AWS & Udacity
- Microsoft Azure AI Fundamentals (AI-900)