ANIRUD MOHAN

College Park, MD | +1-703-705-8375 | amohan26@umd.edu | LinkedIn | Github

SUMMARY

Junior Machine Learning Engineer with experience in deploying LLM-powered chatbots and optimizing model accuracy by up to 96%. Developed end-to-end systems, data pipelines, and robust validation workflows using Python, Java, and related tools. Currently pursuing an MS in Applied Machine Learning, demonstrating strong technical and collaborative skills in agile settings.

SKILLS

- **Languages:** Python, R, SQL, Java, C, C++
- **ML/AI:** Scikit-learn, TensorFlow, Pytorch, HuggingFace, LangChain, Llamalndex, Guardrails, Data/ETL pipelines, Fine-tuning Generative AI Models, MLOps workflow
- **Data Analytics:** Pandas, NumPy, Seaborn, Matplotlib, Plotly, ggplot2, PowerBI, Microsoft Excel, PySpark, Jupyter, SciPy
- **Cloud:** AWS, GCP, IBM Cloud, DataBricks, HuggingFace Spaces
- **Tools:** Docker, Git, LangGraph, ClearML, Flask, Gradio

EDUCATION

University of Maryland **May 2026**

Master of Science, Applied Machine Learning (GPA: 3.76 / 4.00)

- **Coursework:** Principles of Machine Learning, Principles of Data Science, Probability and Statistics, Optimization, Algorithms and Data Structures for ML

Misrimal Navajee Munoth Jain Engineering College **Jul 2019 - May 2023**

Bachelor of Engineering, Computer Science & Engineering (GPA: 8.82 / 10.00)

- **Coursework:** Database Management System, Artificial Intelligence, Software Engineering, Operating Systems

WORK EXPERIENCE

Thapovan Info Systems | *Junior Machine Learning Engineer* **Oct 2023 - Jun 2024**

- Designed and deployed end-to-end LLM-powered chatbot systems using HuggingFace Transformers and Python.
- Integrated Guardrails for LLMs by creating structured input/output validation pipelines and employing Retrieval-Augmented Generation (RAG), improving interaction accuracy by 30%.
- Enhanced model performance with advanced Retrieval Augmented Generation techniques, achieving accuracy improvements up to 96%.
- Collaborated with cross-functional teams using Agile methodologies to support the fine-tuning and deployment of AI models.

Azentio Software | *Software Developer Intern* **Feb 2023 - Sep 2023**

- Refactored a monolithic architecture into a multi-server structure to improve system scalability and performance.
- Restructured secure and scalable APIs by segregating UI and API functionalities, focusing on exposing only the logistics API.
- Implemented Python-based data extraction and transformation methods using Scrapy and BeautifulSoup4, identifying and rectifying three critical system performance bottlenecks.

COURSEWORK & CAPSTONE PROJECTS

AI Powered Code Reviewer - MLops Driven Solution **Feb 2025 - May 2025**

- Fine-tuned Microsoft’s CodeReviewer using LoRA and DDP-based training, leveraging parameter-efficient strategies on 150k+ GitHub pull requests to generate automated, context-aware code suggestions.
- Fine-tuned model achieved BLEU 80.53 and Exact Match 26%, outperforming state-of-the-art baselines.
- Led model deployment and API development using FastAPI, containerized the inference service with Docker and deployed on Hugging Face Spaces to enable scalable, low-latency serving.
- Orchestrated a GitHub Actions CI/CD pipeline, embedding model predictions into pull requests, accelerating developer feedback cycles and slashing integration time by 15% due to faster iteration.

Document Classification using SVM **Aug 2024 - Dec 2024**

- Composed an end-to-end document classification pipeline, implementing SVM from scratch and optimizing TF-IDF feature extraction for high-dimensional text data.
- Engineered an optimized sparse matrix representation, reducing memory usage significantly and accelerating computational speed by 3x in large-scale text processing tasks.
- Implemented 5-fold cross-validation and hyperparameter tuning, achieving 90% accuracy on the 20 Newsgroups dataset through feature scaling and normalization.

Early Diagnosis of Alzheimer's Disease **Jan 2023 - May 2023**

- Designed and implemented deep learning architectures, including CNNs, to analyze MRI scans for early Alzheimer's detection, optimizing model performance and scalability.
- Engineered an end-to-end pipeline for data preprocessing, model training, and real-time inference using TensorFlow and PyTorch, improving computational efficiency.
- Leveraged ensemble learning techniques and hyperparameter tuning to enhance model robustness, achieving over 97% accuracy on the ADNI dataset.