M O N I S H G A L L A

AI & ML Engineer

Mobile: +1 (940) 843-8147 **Gmail:** monishgalla81@gmail.com **Location:** United States
LinkedIn: https://www.linkedin.com/in/monish-galla-0763a1195/

Professional Summary

AI/ML Engineer with 3+ years of experience designing and deploying scalable machine learning pipelines using AWS. Proficient in deep learning, NLP, and cloud-native AI systems, including Retrieval-Augmented Generation (RAG), FastAPI-based ML APIs, and CI/CD automation. Proven impact in real-time inference, predictive analytics, and cross-functional collaboration.

Core Competencies

- **Languages:** Python, SQL
- **Frameworks:** TensorFlow, PyTorch, Scikit-learn, Keras, OpenCV
- **AI/ML:** Generative AI, Deep Learning, YOLO, Transformers, A/B Testing, NLP, RAG
- **Tools:** Power BI, Tableau, Matplotlib, Postman, JIRA, Jenkins, AWS, MySQL, OpenCV
- **Cloud & MLOps:** AWS (SageMaker, EC2, Lambda, ECS), Docker, Jenkins, Terraform, FastAPI, CI/CD, CloudFormation.
- **Soft Skills:** Cross-functional collaboration, problem-solving, analytical thinking, agile team participation

Work Experience

CloudConsulTech (Jan 2024 - May 2025)
AI Cloud Engineer

- Designed end-to-end AI/ML pipelines on AWS (SageMaker, Lambda) with CI/CD automation (CodePipeline, CloudFormation).
- Built Retrieval-Augmented Generation (RAG) systems using LangChain, FAISS, and GPT APIs.
- Deployed containerized ML APIs via FastAPI, Docker, and ECS for real-time inference.
- Managed real-time data pipelines using AWS S3, Lambda, and CloudWatch for event-driven AI workflows.

Quantiphi
AI Engineer (July 2022 – June 2023)

- Designed and deployed deep learning models (CNNs, RNNs, Transformers) for computer vision and NLP tasks, including medical image classification and sentiment analysis.
- Applied self-supervised learning on medical datasets, boosting performance by 15%.
- Deployed models to production using Flask, Docker, and SageMaker.

Data Scientist Intern (Nov 2021 – May 2022)

- Collaborated with senior data scientists to clean, analyze, and model structured and unstructured datasets using Python (Pandas, NumPy) and SQL.
- Built churn prediction models (XGBoost, Random Forest); achieved 78% accuracy.
- Automated reporting with CRON and Python; created dashboards (Tableau, Power BI).
- Conducted A/B testing and hypothesis-driven analysis to optimize marketing campaigns, resulting in a 12% increase in CTR.

Projects

Automated Claim Processing with AI (Jan 2025 - May 2025)

- Built a generative NLP system for medical claim classification; reduced manual processing by 40%.
- Integrated MySQL pipeline for real-time classification of structured and unstructured healthcare data.
- **Tech Stack:** Flask, NLP, MySQL, TensorFlow, Python, and Transformers

AI-Based Traffic Sign Detection (T.R.A.C.) (Sep 2024 – Dec 2024)

- Built YOLOv8-based object detection model (95.7% mAP@0.5); deployed via Jenkins and AWS.
- Jenkins was used to automate training and testing tasks, and models were deployed utilizing cloud technologies (AWS).
- **Tech Stack:** PyTorch, OpenCV, Flask, TensorFlow, AWS, Jenkins

---

Certifications

- Completed AWS Machine Learning Speciality certification - AWS (May 2025)
- Completed Python for Computer Vision with OpenCV and Deep Learning course - Udemy (Jan 2024)

Education

M.S. in Artificial Intelligence

University of North Texas(Aug 2023 – May 2025)

- Relevant Coursework: Deep Learning, Natural Language Processing, Data Science, Machine Learning.
- **GPA:** 3.7

B.Tech in Mechanical Engineering

S.V. College of Engineering (June 2019 – Apr 2023)

- Academic Honors: Top 5% in class, AI Research Assistant.