Mohith Sai Yandra

Email: mohithsaiyandra828@gmail.com | **Phone:** (940) 843-5754 | GitHub

SUMMARY

AI/ML Engineer with **4+ years** of experience building and deploying ML solutions across fraud detection, insurance automation, NLP-driven use cases. Skilled in Python, PyTorch, TensorFlow**,** MLOps tools like MLflow**,** Docker**,** Kubeflow. Proven ability to turn complex datasets into production-ready models with high business impact. Hands-on with cloud platforms like AWS**,** Azure ML Studio**,** GCP Vertex AI. Experienced in NLP**,** LLMs**,** CV**,** model interpretability using SHAP/LIME. Passionate about building scalable, ethical, human-centric AI systems.

TECHNICAL SKILLS

- **Artificial Intelligence & ML:** Machine Learning (ML), Deep Learning, Generative AI, Large Language Models (LLMs), NLP, Computer Vision
- **Programming Languages:** Python, R, Scala, SQL
- **ML Frameworks & Libraries:** TensorFlow, PyTorch, Scikit-learn, XGBoost, Keras, LangChain, OpenCV, SciPy, NLTK, spaCy, Hugging Face
- **MLOps & Deployment:** MLflow, Kubeflow, FastAPI, Docker, Kubernetes, CI/CD, Linux
- **Cloud Platforms:** AWS (EC2, RDS, SageMaker, S3, Redshift, Glue, QuickSight), Azure (ML Studio, Databricks), GCP (Vertex AI, Dataproc)
- **Big Data & Data Engineering:** PySpark, Apache Spark, Kafka, ETL Pipelines, Pandas, NumPy, Hadoop, Airflow
- **Infrastructure & DevOps:** Terraform, Git, GitHub Actions, Jenkins, Kubernetes, IAM, BigQuery
- **Databases & Warehousing:** PostgreSQL, MongoDB, MySQL, Vector Database (Pinecone, FAISS, Chroma DB), DynamoDB, Cloud SQL
- **AI/ML Specializations:** LLM Fine-tuning, Prompt Engineering, SHAP, LIME, Model Optimization (Quantization, CUDA)
- **Data Visualization:** Tableau, Power BI, Matplotlib, Seaborn

EXPERIENCE

Machine Learning Engineer | Verizon Wireless Systems **Jan 2025 – Current**

- Partnered with risk & compliance teams to understand fraud typologies and translate them into model features and labeling strategies.
- Engineered custom features such as transaction velocity, geolocation mismatches, unusual login times across millions of user records.
- Built & compared multiple supervised ML models (Random Forest, XGBoost, Logistic Regression), achieving high precision in pilot phase.
- Implemented SMOTE and class-weight tuning to address high class imbalance (~0.2% fraud cases) in training data.
- Developed real-time scoring pipeline using Kafka, Spark Streaming, and FastAPI, enabling sub-second detection in production.
- Integrated ML models into Power BI for fraud monitoring, enabling analysts to review edge cases using SHAP-based explainability.
- Deployed trained models on AWS SageMaker, orchestrated CI/CD with GitHub Actions and containerized with Docker.

Research Assistant | University of North Texas, USA **Aug 2023 – Dec 2024**

- Led the development of Altriva, a personalized medical chatbot, by fine-tuning Llama 3.1 8B using PEFT, improving diagnostic accuracy.
- Engineered a RAG pipeline using FAISS and LangChain, enhancing document retrieval accuracy and contextual response relevance.
- Integrated multi-turn conversation support using LangChain and NeMo Guardrails for safe, context-aware AI interactions.
- Reduced model size by 40% via quantization and LoRA, enabling low-latency deployment with minimal diagnostic accuracy loss.
- Conducted data-driven model evaluations, leading to iterative improvements and measurable gains in system performance.

Machine Learning Engineer | Coforge **Aug 2021 - Dec 2022**

- Built end-to-end OCR pipeline using Tesseract and AWS Textract with custom heuristics to extract data from scanned claim forms.
- Applied NLP techniques to parse unstructured medical notes, diagnosis codes, and treatment details using spaCy & transformers (BERT).
- Designed rule-based and ML-based claim classification models (approved/denied/flagged), achieving an 87% F1 score on validation data.
- Integrated workflows with insurance CRM systems and automated email communication for accepted/rejected claims using FastAPI.
- Implemented feedback loop via claim outcome tags (e.g., fraud, appeal, payout revised), improving model adaptation over time.
- Worked with DevOps team to containerize and deploy on Azure ML Studio, with weekly model retraining pipelines using Airflow.

Artificial Intelligence Engineer | SEDS – VIT **Aug 2019 – Dec 2020**

- Developed and deployed deep learning models for robotic perception using PyTorch and OpenCV, improving perception accuracy.
- Generated depth maps and point clouds from monocular cameras using neural networks and 3D-KNN for improved terrain analysis.
- Developed 3D scene understanding pipelines with 3D-CNN and PointNet++ for spatial localization using LiDAR and camera data.
- Implemented SLAM with ROS for real-time mapping and localization, enabling autonomous navigation through multi-sensor fusion.
- Optimized neural networks with ONNX Runtime and quantization, achieving 3× faster inference and low-latency edge deployment.

RESEARCH

- **Altriva:** An AI-Powered Chatbot for Personalized Alternative Medicine and Holistic Health Guidance
- Design and Economic Analysis of a Standalone Hybrid Renewable Energy System using Grey Wolf Optimizer

EDUCATION

- **Master of Science: Data Science |** University of North Texas (UNT), Denton, Texas
- **Bachelor of Technology: Electrical and Electronics Engineering |** Vellore Institute of Technology (VIT), Vellore, India

CERTIFICATIONS & ACHIEVEMENTS

- **Large Language Model Operations (LLM Ops) |** Duke University, 2024
- **NASSCOM IoT Domain Specialist |** National Association of Software and Service Companies (NASSCOM), 2021
- **Deep Learning Specialization |** Coursera (Andrew Ng), 2020
- **Machine Learning with Python |** Cognitive Class, 2020
- **University Rover Challenge** (Finalist, University Rover Challenge 2020 (Led 12-member team to build a Martian Rover, selected among top international teams in Utah, USA).