

SIVA PRASAD MANDALAPU

Machine Learning Engineer | Production ML, MLOps, Fraud & Healthcare AI

Irving, TX. | +1 9433008560 | [Email](#) | [LinkedIn](#) | [Github](#) | [Portfolio](#)

PROFESSIONAL SUMMARY

Machine learning engineer who gets models into production and keeps them healthy there. 5 years running real-time ML in regulated industries: 5M+ fraud scores/day at 99.9% uptime at HDFC Bank and 13x faster medical-imaging inference at Philips Healthcare. AWS Certified Machine Learning – Specialty.

TECHNICAL SKILLS

- **MLOps**: CI/CD for ML, model registry, experiment tracking, automated retraining, drift monitoring, rollback, MLflow, Kubeflow, Airflow, GitHub Actions, Jenkins
- **Model Serving**: real-time inference, GPU optimization, TensorRT, FastAPI, Docker, Kubernetes, AWS SageMaker, Azure ML
- **Observability**: Prometheus, Grafana, LangSmith, Weights & Biases
- **Data & Cloud**: Spark, Kafka, Pandas, Dask, PostgreSQL, AWS (S3, EC2, Lambda, EKS), Terraform
- **ML & Languages**: Python, SQL, Bash, PyTorch, TensorFlow, scikit-learn, XGBoost, LightGBM, LangGraph, SHAP, Flower

PROFESSIONAL EXPERIENCE

Philips Healthcare

Dec 2023 - Present

AI/ML Engineer (Contract)

Remote.

- Cut inference latency 13x (2s to 150ms per scan) at 10K+ scans/day with no loss in precision by converting PyTorch models to TensorRT and rebuilding model serving as containerized FastAPI services on on-prem GPU edge devices.
- Cut model release-candidate turnaround from 4 weeks to 1 by building ML CI/CD with GitHub Actions, Jenkins, Docker and Kubernetes, an MLflow model registry, and regression tests on a locked validation set.
- Monitored drift and model health for the 10K+ scans/day service with Prometheus and Grafana, tracking data drift, service health, and per-subgroup model performance to provide early visibility into model and service issues
- Made every training run reproducible over 500GB+ of imaging data with Spark and Kafka pipelines, schema validation and automated data-quality gates feeding AWS SageMaker training jobs.
- Trained models across 5 hospitals with zero patient images moved by building the client-side training and evaluation pipeline for a federated learning pilot (Flower).
- Helped cut radiologist review time 35% in an LLM pilot by building the LLMops layer for a LangGraph report-drafting pipeline on Azure OpenAI: LangSmith tracing and automated checks for unsupported statements.

HDFC Bank

Dec 2019 - Nov 2022

AI/ML Engineer, Fraud & Risk

Bengaluru, India.

- Served 5M+ real-time fraud scores/day at under 100 ms and 99.9% uptime by deploying containerized model endpoints on Azure ML and Kubernetes.
- Turned model updates into scheduled, automated releases by building Kubeflow retraining pipelines with MLflow experiment tracking and model versioning, feature and score drift detection, and automated rollback.

- Cut false-positive alerts 30% at 95% recall and fraud losses 35% by partnering with data scientists and fraud analysts to productionize an LSTM, Transformer and graph neural network ensemble.
- Cut fraud investigation time 40% by attaching SHAP explanations to every flagged transaction and auto-generating case summaries in the format required by RBI auditors
- Improved model F1 22% with Pandas and Dask feature engineering pipelines over 1M+ labeled transactions.

EDUCATION

• M.S., Information Systems Marist College	Jan 2023 - Dec 2024 Poughkeepsie, NY
• B.Tech, Mechanical Engineering Sir C.R.R. College of Engineering	Jun 2015 – May 2019 India

CERTIFICATIONS

- AWS Certified Machine Learning – Specialty
- TensorFlow Developer Certificate (Google)
- Generative AI with Large Language Models (DeepLearning.AI & AWS)