Samhith Raj Saneboina

NJ , 07307 | +1 862-800-5135 | samhithraj.s@mycvscout.com | LinkedIn

Professional Summary

- AI/ML Engineer with 5+ years of experience in designing, deploying, and scaling end-to-end machine learning solutions across NLP, Computer Vision, Generative AI (LLMs), Retrieval-Augmented Generation (RAG), and advanced analytics using deep learning and statistical modeling.
- Expert in developing supervised and unsupervised models using Logistic Regression, Random Forest, XGBoost, SVM, KMeans, DBSCAN, and ensemble techniques, optimizing performance with ROC-AUC metrics and fine-tuning of models for high scalability.
- Expert in fine-tuning and deploying deep learning architectures including CNNs, LSTMs, Transformers, and large language models (LLMs) such as BERT, GPT, and LLaMA, utilizing frameworks like TensorFlow, PyTorch, and Hugging Face Transformers.
- Proficient in Python programming and data engineering with Pandas, NumPy, Scikit-learn, OpenCV, NLTK, SpaCy, Matplotlib, Seaborn, and tools for feature engineering, data visualization, NLP pipeline development, and GPU-accelerated training.
- Skilled in ML infrastructure management and deployment using AWS (SageMaker, EC2), Google Cloud (Vertex AI, BigQuery), Azure DevOps, Docker, Kubernetes, CI/CD pipelines, and MLOps best practices for production-grade AI systems, including RAG frameworks and LLM hosting.

Education

Master of Science in Computer Science | Stevens institute of Technology, Hoboken, USA

Skills

Language/ IDE's: Python, MATLAB, Jupyter Notebook, Google Colab, VS Code, SSMS
Machine Learning: Linear & Logistic Regression, Decision Trees, Random Forests, NumPy, SVM, ROC-AUC, RAG, Fine Tuning of models
Deep Learning: CNN, RNN, LSTM, NLP, Large Language Model (LLM), LangChain, Hugging Face Transformers (BERT, GPT-3)
Cloud/Visualizations: AWS (EC2, SQS, SNS, Code Deploy, CloudWatch, API Gateway), GCP (Vertex AI, Google Cloud Storage), Tableau, Power BI
Statistical Techniques: Hypothesis Testing, Data Visualization, Data Modelling, A/B testing, Model Evaluation
Packages and Frameworks: NumPy, Pandas, Matplotlib, Scikit-learn, Seaborn, TensorFlow, Keras, NLTK, XGBoost, PyTorch, Hugging Face
Database and Tools: SQL Server, MySQL, PostgreSQL, Redis, Neo4j
Certifications: AWS Certified Associate

Work Experience

Johnson& Johnson, New Brunswick, NJ **Jun2024- Present**

AI/ML Developer

- Designed and implemented advanced deep learning models using CNN and sophisticated computer vision techniques on GPU clusters with CUDA optimization to analyze complex medical imaging data, achieving 94% validation accuracy and reducing diagnostic processing time by 40%.
- Engineered and optimized over 15 critical behavioral and physiological features using cosine similarity, complemented by NLP-based feature extraction and RAG-augmented clinical context retrieval via Hugging Face BERT embeddings, improving model precision by 38%.
- Developed highly efficient real-time gaze-tracking algorithms with linear transformation, normalization techniques, and fine-tuning of models to ensure consistent, high-quality data input across diverse devices, boosting model reliability by 25%.
- Built scalable and automated machine learning pipelines on AWS Lambda, Glue, and S3 using PyTorch and TensorFlow, with GPU-accelerated training and hyperparameter tuning of models, streamlining training and retraining workflows and reducing data engineering overhead by 60%.
- Collaborated closely with leading healthcare researchers to translate complex clinical diagnostic parameters into actionable machine learning features using LLM-based embeddings and NLP techniques, contributing to the development of AI systems with FDA regulatory guidelines.
- Created Power BI dashboards integrated with SQL databases to continuously monitor key model performance metrics, including inference speed, ROC-AUC, attention drift, and threshold sensitivity, providing actionable insights for product and research teams.

Goldman Sachs, Jersey City, NJ **Oct 2023– May 2024**

AI & Deep Learning Developer

- Architected cutting-edge deep learning models for financial time-series forecasting using GPU-accelerated training, significantly improving asset price prediction accuracy by 22%, enhancing algorithmic trading strategies and portfolio optimization.
- Implemented NLP pipelines and fine-tuned BERT transformers via Hugging Face to analyze financial news and social media sentiment, and integrated RAG retrieval mechanisms for enriched context, enhancing trading signal precision by 18% and improving market responsiveness.
- Executed scalable, cloud-native machine learning workflows on Google Cloud Platform BigQuery and Vertex AI, deploying LLM-based applications with fine-tuning, reducing model retraining time and risk evaluation latency by 50% while ensuring high availability.
- Applied unsupervised learning and autoencoder techniques using PyTorch and TensorFlow for transaction anomaly detection, optimizing ROC-AUC thresholds to reduce false positives by 30% and strengthening fraud detection capabilities.
- Streamlined end-to-end CI/CD pipelines for machine learning lifecycle management with Docker, Kubernetes, and GitHub Actions, incorporating model versioning and GPU resource management, reducing deployment times by 40% and ensuring seamless production rollouts.
- Partnered with quantitative analysts, traders, and compliance teams to ensure models adhered to regulatory standards and business KPIs, improving transparency, reliability, and stakeholder trust through LLM-driven interpretability frameworks.

Accenture, Noida, India **Oct 2019– Aug 2022**

Machine Learning Engineer

- Devised deep learning models including convolutional and recurrent neural networks for automated document classification and object detection, increasing operational throughput by 40% and accelerating enterprise AI adoption.
- Optimized and fine-tuned NLP models using BERT and GPT-3 via Hugging Face Transformers for sentiment analysis and intent detection, improving classification accuracy by 20% and reducing customer escalation times.
- Formed reusable PyTorch and TensorFlow pipelines with automated hyperparameter tuning and batch inference, reducing model training time by 60% and ensuring reproducibility and scalability across deployments.
- Formulated and deployed real-time AI services using OpenCV and LSTM networks to detect abnormal behaviors in video streams, reducing manual fraud monitoring efforts by 55% and strengthening compliance controls.
- Transformed legacy machine learning models to AWS SageMaker using containerized deployments and established API Gateway and Lambda endpoints, reducing deployment time by 50% and enabling seamless model scaling.
- Orchestrated model evaluation, logging, and drift detection with MLflow and AWS CloudWatch, enhancing governance, reducing production incidents by 40%, and ensuring continuous model performance.
- Crafted and maintained Tableau dashboards integrated with prediction confidence intervals and model impact KPIs, boosting stakeholder decision-making efficiency by 30% and accelerating data-driven strategy adoption across departments.