Vaishnavi K

AI/ML Engineer

USA | +1 (571) 637-9044 | vaishnavi.k111101@gmail.com | LinkedIn | Portfolio

SUMMARY

AI/ML Engineer with 4+ years of experience designing and deploying end-to-end machine learning solutions across healthcare and manufacturing domains. Skilled in building production-grade ML pipelines using Python, PySpark, and cloud platforms (AWS, Azure, GCP), with hands-on experience in LLMs, RAG architectures, and GenAI applications that drive measurable business impact. Proven track record of architecting scalable microservices, MLOps frameworks, and real-time anomaly detection systems while maintaining compliance standards and achieving 99%+ uptime in mission-critical environments.

TECHNICAL SKILLS

Languages & Frameworks: Python(Numpy, Pandas, Matplotlib), R, SQL, Bash, Scikit-learn, PyTorch, Tensorflow, Keras

Databases: MySQL, PostgreSQL, MongoDB, Vector Databases (Pinecone, Chroma DB)

ML & Statistical Modeling: Supervised & Unsupervised Learning, Reinforcement Learning, Feature Engineering, Model Selection, Hyperparameter Tuning, Ensembling, Recommender Systems, Anomaly Detection, A/B Testing

ML Algorithms: Linear & Logistic Regression, Decision Trees, Random Forest, Gradient Boosting (XGBoost, LightGBM), SVM

Deep Learning & NLP: Langchain, Langgraph, Neural Networks (CNN, RNN, LSTM), Hugging Face Transformers, Embeddings, Tokenization, Text Preprocessing, NER, Semantic Analysis, Prompt Engineering

LLMs & Architectures: LLM Fine-Tuning (LoRA, QLoRA, PEFT), Transformers(BERT, RoBERTa, BART, GPT), RAG, GAN

Cloud, Big data & ETL: AWS(S3, EC2, Lambda, Sagemaker), Azure(ML Studio, Blob Storage), GCP(BigQuery, AI Platform), Apache Spark (PySpark), Apache Airflow, Databricks, Kafka, Hadoop, ETL/ELT Pipelines

Data Analytics & BI: Exploratory Data Analysis (EDA), Data Visualization, Tableau, Power BI

MLOps & DevOps: MLflow, CI/CD, Docker, Jenkins, Git, Model Versioning, Fast API

PROFESSIONAL EXPERIENCE

AI/ML Engineer | TCS, USA **| Jan 2025 – Present**

- Designed and deployed end-to-end ML pipelines leveraging AWS SageMaker, Lambda, and S3 with Python and PySpark, accelerating model deployment cycles by 70% while establishing MLflow version control and CI/CD automation through Jenkins for continuous model updates.
- Developed LLM applications using LangChain and RAG architecture to analyze technical documentation with GPT and BERT transformers, reducing manual review efforts by 45% through intelligent semantic search and prompt engineering on vector databases like Pinecone.
- Built GenAI troubleshooting applications using LangChain and Langgraph that integrate RAG with maintenance manuals and repair histories stored in MongoDB, enabling knowledge-driven diagnostics that cut troubleshooting time by 30% while maintaining Git-based version control.
- Engineered predictive maintenance models combining XGBoost gradient boosting, CNN, and fine-tuned transformers to process time-series sensor data from 200+ factory assets with MySQL and PostgreSQL backends, achieving 94% prediction accuracy with embedded LLMOps monitoring.
- Developed multimodal AI inspection systems integrating YOLOv5 computer vision with semantic text analysis using transformers, coupled with A/B testing frameworks to validate model improvements, reducing manual inspection workload by 60% while sustaining 99.8% defect detection.
- Implemented cloud-native ML infrastructure using Kubernetes, MLflow, and Docker containers for automated model retraining and zero-downtime deployments across production microservices, with comprehensive monitoring dashboards enabling seamless scaling.
- Built real-time anomaly detection systems deployed as FastAPI microservices that consume Kafka event streams and identify production anomalies within seconds, supported by comprehensive logging, Git branching, and automated alerts that reduced quality incidents by 25%.

AI/ML Engineer | HCL Tech, India **| Jun 2020 – Jul 2023**

- Developed ensemble-based anomaly detection models using Random Forest and SVM on medical insurance claims data, successfully identifying
- $2.5M in annual claim anomalies with 96% precision and 91% recall, leveraging MySQL and Oracle databases for HIPAA-compliant data storage.
- Designed hospital capacity forecasting systems using ARIMA, Prophet time-series models, and LSTM neural networks to predict ER visits and optimal bed utilization across facilities, reducing average patient wait times by up to 4 minutes through data-driven resource allocation.
- Automated clinical document processing using BERT transformers and custom NLP pipelines for discharge summaries and prescriptions with feature engineering techniques, cutting manual audit efforts by 60% while maintaining end-to-end HIPAA compliance and data privacy standards.
- Built personalized patient engagement systems using collaborative filtering and recommendation algorithms with Keras neural networks, driving 20% increase in patient engagement rates and reducing 15% of preventable readmissions through targeted intervention suggestions.
- Engineered HIPAA-compliant ETL pipelines using PySpark, Apache Spark, Apache Airflow, and Hadoop ecosystems to ingest and process multi-source EHR data from distributed healthcare systems, enabling real-time clinical analytics while ensuring data lineage and governance compliance.
- Deployed scalable AI microservices architecture with FastAPI, Docker containers, and Azure Kubernetes Service across 6+ healthcare facilities, achieving 99.9% uptime with Redis caching for performance optimization and PostgreSQL/MongoDB database integration.
- Implemented Explainable AI using SHAP and LIME interpretability tools to enhance clinical decision-making and physician trust, with Power BI dashboards visualizing model predictions and feature importance, supporting FDA audit approvals and accelerating physician adoption of outputs.

EDUCATION

Master of Science in Data Science | University of New Haven, West Haven, USA **| May 2025**

CERTIFICATION

- Tata - GenAI Powered Data Analytics Job Simulation by Forage **| Jul 2025**
- Python for Data Science by IBM **| May 2023**
- Microsoft Certified: Azure Fundamentals (AZ-900) – Microsoft **| Oct 2022**

PROJECTS

- Plagiarism Detection using Transformers **| Dec 2024**
- Object Detection using yolov5s **| Dec 2024**

PUBLICATIONS

- Integrating MobileNetV3 and SqueezeNet for Multi-class Brain Tumour Classification **|** Link
- Miniaturised Planar Dual Band Monopole UWB Antenna using Capacitively Loaded Loop Resonator with Notch Characteristics **|** Link