Dinesh Kothandaraman

West Haven, CT | dineshkothand@gmail.com | +1 475-227-5994 | LinkedIn | Github

Summary

Experienced AI/ML Engineer with over 6 years of software delivery experience, specializing in designing, implementing, and deploying production-grade AI/ML solutions. Proven expertise in developing scalable MLOps supportive data pipelines, end-to-end Machine Learning workflows, and advanced Gen AI applications. Highly skilled in Python programming, SQL, TypeScript, Go(Basic familiarity), machine learning frameworks (TensorFlow, PyTorch), containerization (Docker, Kubernetes), CI/CD practices with strong foundation in data structures, algorithms, and object-oriented design(OOP). Adept at leveraging large language models (LLMs), Retrieval Augmented Generation (RAG), and advanced cloud platforms like Google Cloud Platform (GCP), Vertex AI, and Google AutoML. Strong background in Big Data technologies including Spark, Hadoop, and Hive for distributed data processing. Committed to delivering robust, scalable, and high-performance AI/ML products.

Technical Skills

AI/ML & Gen AI: Machine Learning, Large Language Models (LLMs - Gemini, OpenAI, Claude, open-source LLMs), Retrieval Augmented Generation (RAG), Supervised Tuning, Neural Networks, Transformers (T5, BART), Attention Mechanisms, NLP (NLTK, KeyBERT, Sentence Transformers), Agentic AI Workflows, Intent Recognition, Text Summarization, Automated Speech Recognition (ASR), SSML Tagging for TTS.
MLOps & DevOps: Docker, Kubernetes, Git, GitHub Actions, DVC, MLflow, Experiment Tracking, Model Versioning, Deployment Automation, Workflow Orchestration (Apache Airflow), Kubeflow, CI/CD Pipelines, Agile Methodologies, Unit Testing.
Big Data Technologies: Apache Spark, Hadoop, Hive.
Programming Languages: Python, JavaScript/TypeScript, SQL, Java, HTML, Go, OOP
ML Frameworks & Libraries: TensorFlow, PyTorch, Scikit-learn, Apache Spark ML, FastAPI, RESTAPI, Flask, LangChain.
Cloud Platforms & Tools: Google Cloud Platform (GCP), Vertex AI, Google AutoML, AWS (S3, Lambda, Glue, QuickSight), Azure.
Databases: Vector DBs,,Snowflake, PostgreSQL, SQLiteStudio, SQL Server.
Other: Distributed Systems, Data Science, Data Ingestion, Data Processing, Data Storage.

Experience

Agentic AI Developer | Kamali Inc**,** Remote-United States| Sep 2024 – Jun 2025

- Designed and deployed conversational chatbots for document-based QA using LangChain, vector search (HNSWLIB), and Large Language Models (LLMs), demonstrating expertise in RAG techniques.
- Engineered multi-agent workflows with LangChain to simulate user intent planning and response routing, enhancing conversational continuity and user experience.
- Implemented and deployed AI/ML applications using FastAPI and Docker on AWS, following CI/CD best practices, ensuring scalable and robust production deployment.
- Customized prompt templates and agent memory for contextual continuity across chatbot sessions.
- Applied OOP principles and secure coding practices to ensure maintainability and safety in production.
- Leveraged automatic speech recognition (ASR) systems and SSML tagging for text-to-speech (TTS) optimization to support voice-to-text input in chatbot interactions.

Machine Learning Engineer | Cognizant Technology Solutions, India | Jan 2021 – July 2023

- Designed and implemented ML pipelines for automating underwriting and risk assessment processes, enabling faster and data-driven policy issuance.
- Used PySpark to process large-scale health and claims data (~TBs), performing ETL, feature engineering, and aggregations across policyholder and claims histories.
- Developed classification models (XGBoost, Logistic Regression) to predict individual risk scores, improving underwriting precision and reducing turnaround time by 35%.
- Exposed model predictions via Flask-based REST APIs, integrated into internal dashboards and agent-facing tools.
- Containerized and deployed services on AWS EC2 and Lambda, ensuring high availability and low-latency responses across regions.
- Enhanced model interpretability using SHAP and LIME, enabling underwriters to understand key drivers behind risk decisions.
- Collaborated cross-functionally in Agile sprints with actuarial analysts, data engineers, and QA teams to deliver iterative model improvements and API enhancements.

Machine Learning Engineer | Jubliee Information Technology Pvt Ltd, India | July 2017 – Dec 2020

- Developed ML-based product recommendation system to suggest personalized insurance plans based on demographic, behavioral, and claim history data, improving upsell conversions by 20%.
- Built customer segmentation pipelines using K-Means and hierarchical clustering to group policyholders by health risk, lifestyle indicators, and purchasing behavior.
- Processed large-scale policyholder and claim data using PySpark, performing feature engineering, outlier detection, and aggregation at household and individual levels.
- Automated daily batch scoring workflows using PySpark on AWS EMR, pushing predictions to downstream CRM systems via Kafka and PostgreSQL.

Publication

Hybrid CNN-LSTM Model for Jail Activity Recognition, Oct 2024

Published in **2024 IEEE 15th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)**
https://ieeexplore.ieee.org/abstract/document/10754670

Projects

Document Based QA System (Conversational AI Chatbot)

- Developed a chatbot using Python and Retrieval-Augmented Generation (RAG) with large language models (LLMs) to answer questions based on document content, leveraging vector embeddings for efficient context retrieval and intent recognition.
- Orchestrated data processing and model inference using Apache Airflow to automate pipeline execution, ensuring seamless integration of document ingestion and question-answering tasks.
- Implemented CI/CD workflows with GitHub Actions to automate model updates and monitoring, achieving robust production-grade deployment for a conversational agent.

Remote Patient Care with AWS and IoT (Healthcare Monitoring System)

- Designed a scalable health monitoring system integrating IoT devices with AWS tools, demonstrating experience relevant to the healthcare domain.

- Utilized AWS Glue for processing health data and Python for anomaly detection in real-time metrics, supporting proactive patient care.
- Developed dashboards in AWS QuickSight to visualize and monitor health indicators, providing actionable insights for healthcare providers.

Text Summarization - Natural Language Processing

- Developed three text summarization models: an LSTM with attention mechanism and two pre-trained transformer models (T5 and BART), focusing on both abstractive and extractive techniques.
- Fine-tuned T5 and BART models on domain-specific datasets, leveraging their encoder-decoder architecture for high-quality summarization with improved ROUGE and BLEU scores.
- Implemented an LSTM with attention to enhance focus on relevant text segments, visualizing attention weights for interpretability and improving model performance on custom datasets.
- Evaluated and compared models on benchmark datasets, demonstrating the superiority of transformer-based methods while showcasing expertise in both traditional and cutting-edge NLP techniques.

Car Logo Detection - Deep Learning

- Developed a car logo detection model using PyTorch and transfer learning, leveraging pre-trained CNN architec- tures (ResNet, VGG) to achieve high accuracy on custom datasets.
- Implemented transfer learning by fine-tuning a pre-trained model on a labeled car logo dataset, reducing training time and improving generalization.
- Integrated ClearML for streamlined experiment tracking, performance monitoring, and efficient model versioning during the development lifecycle.
- Optimized model performance through data augmentation, learning rate scheduling, and loss function tuning, achieving significant improvement in precision and recall metrics.

Jail Activity Recognition Using CNN and LSTM

- Built a real-time activity recognition model combining CNN and LSTM architectures.
- Processed and analyzed time-series data, achieving high model accuracy.

Gender and Age Detection

- Developed a Python-based deep learning project to detect faces and classify age and gender from images.
- Leveraged pre-trained models to achieve high accuracy and real-time performance.
- Utilized advanced computer vision techniques for optimized analysis.

Education

MS, University of New Haven, Data Science | Aug 2023 – May 2025

- GPA: 3.89/4.0
- **Coursework:** Artificial Intelligence, Machine Learning, Distributed and Scalable Data Engineering, Programming for Data Science, Deep Learning, Natural Language Processing, Power BI and Bayesian Data Analysis.

BE, Chennai Institute of Technology, Mechatronics Engineering | June 2013 – April 2017

- GPA: 7.35/10
- **Coursework:** Computer Programming, Transforms and Partial Differential Equations, Statistics and Numerical Methods, Object Oriented Programming in C++.