Dehit T
San Jose, CA
https://www.linkedin.com/in/dehit-trivedi/
+1 (628) 400-4013 | dehit.ml@gmail.com

Summary

Machine Learning Engineer and Gen AI with over 7+ years of experience delivering end-to-end AI/ML solutions across AWS, GCP, and Azure platforms. Strong background in designing, deploying, and optimizing machine learning models, including deep learning, Generative AI, and LLMs (Azure OpenAI, LLaMA 2/3, Hugging Face). Demonstrated success in deploying real-time, scalable AI systems across domains such as autonomous driving, voice-controlled agents, predictive analytics, and intelligent document processing.

Experienced in designing AI/ML solutions for high-precision decision systems, including model validation, confidence scoring, and error analysis. Skilled in analyzing false positives, edge cases, and improving model accuracy for critical workflows. Familiar with audit-ready AI systems aligned with compliance and responsible AI.
Vector DB & Search: Pinecone, FAISS, Semantic Search
Strong expertise in deep learning, model evaluation, and large-scale data pipelines using PyTorch and TensorFlow.
Proven experience in developing production-grade ML pipelines, optimizing model performance, and working with cloud and MLOps tools.
Hands-on experience with Generative AI, LLMs, and Hugging Face frameworks, including chatbot development and model inference pipelines.
Adept at solving complex real-world problems and collaborating across cross-functional teams in fast-paced environments.
Cloud: AWS (SageMaker, Bedrock, Lambda), GCP (Vertex AI), Azure (OpenAI, Databricks)
Data Engineering: PySpark, Spark, Kafka, Airflow, DBT, BigQuery, Dataflow
Machine Learning: PyTorch, TensorFlow, Scikit-learn, Diffusion Models, Transformers, Time-series
MLOps & Production: MLflow, Kubeflow, TFX, Docker, Kubernetes, CI/CD, Model Monitoring, A/B Testing
Generative AI & LLMs: GPT-4/4o, Azure OpenAI, LLaMA 3, Mixtral, Claude, RAG, LangChain, LlamaIndex, Agents (Bedrock/OpenAI)
Healthcare & AI Validation: Model Validation, Confidence Scoring, Precision Optimization, Error Analysis, Model Explainability, Responsible AI, Data Governance

Unstructured Data Processing: Document AI, OCR Pipelines, NLP on Unstructured Data, Multimodal Processing

Recent contributions include developing advanced diffusion models for autonomous vehicle trajectory prediction and leading the development of Toyota's VIA (Vehicle Intelligence Agent), a real-time, voice-driven diagnostic AI assistant integrated with CAN/OBD-II data and Bedrock-based RAG systems.

Core Competencies:

- **Machine Learning & AI Solutions:** Expert in developing and deploying advanced ML models including diffusion policy models, deep neural networks (CNNs, RNNs), and ensemble methods. Skilled in trajectory prediction, classification, clustering, and real-time inference for edge and cloud environments.
- **Natural Language Processing & Generative AI:** Experienced in fine-tuning and deploying LLMs (Azure OpenAI, LLaMA 2/3, Hugging Face Transformers), and building NLP applications using DistilBERT,

LangChain, and RAG architectures. Developed voice-enabled assistants and ASR pipelines for real-time conversational AI.

- **MLOps & Model Deployment:** Proficient in end-to-end ML lifecycle management with CI/CD, Docker, Kubernetes, MLflow, TFX, and Kubeflow. Automated model retraining, monitoring, and deployment across cloud and edge devices.
- **Cloud Platforms & Infrastructure:** Hands-on experience with AWS (SageMaker, Lambda, S3, CloudFormation), GCP (Vertex AI, BigQuery, Dataflow), and Azure (OpenAI, Databricks, Data Factory) for scalable AI/ML system development and infrastructure automation.
- **Data Engineering & Pipelines:** Skilled in building robust data pipelines using PySpark, Apache Spark, Kafka, Airflow, and DBT (Data Build Tool). Experienced in implementing data modeling techniques including Star Schema and Dimensional Modeling for scalable analytics platforms. Managed large-scale ETL/ELT processes, data transformations, and data quality validation using DBT tests and SQL-based data validation frameworks.
- **Data Preparation & Feature Engineering:** Expert in data cleaning, transformation, augmentation, and feature extraction using Python (Pandas, NumPy). Addressed long-tail distributions and rare event modeling using advanced loss functions and Bayesian optimization.
- **AI for Embedded & Edge Systems:** Developed and optimized models for deployment on edge devices, integrating multimodal sensor data (LiDAR, radar, camera) for real-time inference in autonomous systems.
- **Visualization & Insights:** Created performance dashboards and training visualizations using Tableau, Power BI, and Matplotlib. Delivered actionable insights to business and engineering stakeholders.
- **Collaboration & Communication:** Led cross-functional projects and technical workshops on topics like Generative AI, diffusion models, and computer vision. Strong collaborator across product, data, and engineering teams.

Core Expertise

AI/ML Solutions	Azure Open AI, Llama 2/3, Mixtral
Cloud Platforms	AWS, GCP, Azure AI services (Azure AI Search, Azure Open AI)
Cloud Resources	Azure Databricks, AWS Glue, GCP BigQuery, Dataflow, DBT (Data Build Tool)
Programming Languages	Python, PySpark, C++, SQL, JavaScript, HTML, CSS
ML Framework	TensorFlow, Keras, DistilBert, PyTorch, Langchain & Llama Index, sci-kit-learn, NLTK, Spacy, Pandas, NumPy, Hugging Face Transformers, NLTK, OpenCV, CUDA
ML Algorithms	Regression (Linear, Polynomial, Ridge, Lasso, Decision Tree Regressor, MLP, ANN), Classification (Logistic Regression, SVM, Decision Tree, Random Forest, Naïve Bayes, KNN, ANN, Ensembling techniques), Clustering (K-means, K-median, K-mode, Agglomerative Clustering.
Frameworks	Flask, Django, Express, EJS
Data Management	Data Modeling (Star Schema, Dimensional Modeling), DBT (Data Build Tool), Vector Databases, SQL, NoSQL, Data Warehousing, Data Transformation
AI/ML Ops Practices	Model Monitoring, optimization & deployment, fine-tuning
Software Version Control & Documentation	Git, JIRA, Confluence
Containerization & Orchestration Tools	Docker, Kubernetes, AirFlow & MLFlow
Monitoring Tools	Power BI & Tableau

Soft Skills	Excellent communication, leadership, collaboration, and Project management

Experience

Toyota Research Institute **Aug 2024 – Present**
Sr AI/ML Engineer/Gen AI Los Altos, CA

Project Objective: Developed and refined advanced diffusion policy models to improve Toyota's autonomous driving capabilities. This work focused on enabling the car to accurately predict future paths and actions in complex driving environments, contributing to safer and more reliable self-driving technology.

Key Responsibilities:

- Led the development and optimization of diffusion models using PyTorch, Hugging face for future path prediction, ensuring high accuracy and reliability in autonomous driving scenarios, including handling complex multi-agent interactions and diverse road conditions.
- Fine-tuned diffusion models to capture temporal dependencies in driving data, leveraging Generative AI and probabilistic modeling to anticipate car actions and paths in dynamic environments.
- Integrated multimodal sensor data (e.g., LiDAR, radar, camera inputs) into the diffusion models, enhancing situational awareness and enabling accurate trajectory predictions.
- Optimized model performance using advanced hyperparameter tuning strategies, such as experimenting with noise schedules, diffusion timesteps, and model architectures to ensure rapid convergence and reduced latency.
- Designed and deployed end-to-end ML pipelines handling users / requests per day
- Built deep learning models (NLP/CV) improving accuracy.
- Developed scalable ML evaluation pipelines for autonomous driving systems using Waymo and NavSim datasets.
- Explored agentic workflows by integrating LLM-based decision pipelines to automate multi-step reasoning tasks.
- Built automated inference and tracking pipelines integrated with CI/CD workflows for model validation.
- Deployed and tested AI/LLM pipelines in AWS cloud environments.
- Contributed to development of internal GenAI chatbot leveraging LLMs and Hugging Face frameworks.
- Designed LLM orchestration pipelines integrating prompts, inference layers, and backend services for scalable AI workflows.
- Applied data augmentation and synthetic data generation techniques to increase model robustness in handling rare and edge-case driving scenarios.
- Developed and maintained a scalable training pipeline on high-performance computing infrastructure, deploying models on cloud platforms with distributed GPU training to accelerate development cycles.
- Collaborated with cross-functional teams to ensure real-time implementation feasibility, providing insights into model improvements based on real-world autonomous driving metrics.
- Optimized the model for the deployment on the edge devices which is scalable for the massive production for Toyota car manufacturing.
- Streamlined deployment pipelines using CI/CD practices with Docker and Kubernetes, ensuring seamless model updates and high system reliability in real-world test environments.
- Conducted in-depth model evaluation using simulation tools and real-world testing to monitor the accuracy and robustness of predicted trajectories across varying conditions.

- Led technical knowledge-sharing sessions on Generative AI for trajectory prediction, diffusion policy modeling, and autonomous driving strategies, fostering team-wide advancements in predictive modeling.

Additional Contributions – VIA (Vehicle Intelligence Agent) Project:

- Developed a voice-controlled AI assistant (VIA) capable of interpreting vehicle CAN and OBD-II data to diagnose real-time vehicle issues and provide intelligent, contextual responses to drivers.
- Integrated VIA with a Bedrock agent architecture for retrieval-augmented generation (RAG), combining internal Toyota knowledge bases with live data for accurate and vehicle-specific answers.
- Designed and deployed a robust voice-to-text pipeline optimized for in-vehicle environments, leveraging ASR models adapted for noisy conditions and constrained edge devices.
- Built a scalable backend for VIA using AWS Lambda, DynamoDB, and API Gateway to support both mobile (One Toyota app) and embedded in-vehicle applications.
- Worked closely with UX and HMI teams to design intuitive, safe, and voice-first interaction patterns tailored for in-car usage.
- Contributed to the integration of VIA into Toyota's production ecosystem, ensuring security, data privacy, and OTA update capabilities.
- Spearheaded internal demos and workshops showcasing VIA's capabilities in predictive maintenance, onboard diagnostics, and user education via multimodal interfaces.

Technical Tools and Environment: Python, PyTorch, TensorFlow, Pandas, NumPy, Matplotlib, diffusion policy models, Generative AI techniques, AWS/GCP, Docker, Kubernetes, distributed training setups, CI/CD

Intuit, Mountain View, CA **Feb 2021 - Aug 2024**
Machine Learning Engineer

Project Objective: Led the development of a high-accuracy HS tax code classification system, achieving over 90% accuracy. This significantly enhanced product classification for customs and taxation in cross-border contexts, utilizing advanced deep learning techniques and overseeing end-to-end model training and deployment on AWS.

Key Responsibilities:

- Developed a high-accuracy classification system for ~100,000 HS codes, improving accuracy from 70% (for 20% of codes) to over 90% (for 60% of codes).
- Fine-tuned DistilBERT and implemented hierarchical tree-based models (e.g., Decision Trees, Logistic Regression) for 6/10-digit tax code prediction.
- Applied NLP techniques using Hugging Face Transformers to process and analyze complex product descriptions.
- Built and optimized deep neural network architectures in PyTorch, experimenting with hyperparameters (e.g., learning rate, batch size, activation functions) for better convergence and accuracy.
- Utilized Pandas, NumPy, and Scikit-learn for data preprocessing, analysis, and model evaluation.
- Created scalable data pipelines with PySpark, Apache Spark, Kafka, and SQL to automate ingestion, transformation, and feature engineering for training.
- Designed dimensional data models and Star Schema structures in BigQuery to support scalable analytics and reporting pipelines.
- Implemented DBT (Data Build Tool) for ELT transformations, modular SQL transformations, and version-controlled data workflows.

- Built data quality and validation tests using DBT testing frameworks to ensure integrity, consistency, and reliability of analytical datasets.
- Streamlined deployment using Docker, Kubernetes, and CI/CD pipelines, ensuring robust model delivery and updates.
- Trained and deployed models on AWS using SageMaker, S3, Lambda, and CloudFormation for scalable, serverless infrastructure.
- Integrated GPU acceleration with CUDA and cuDNN, significantly reducing training times.
- Built MLOps pipelines with MLflow, Kubeflow, and TFX to automate model monitoring, retraining, and deployment.
- Tackled long-tail data challenges using advanced loss functions and Bayesian optimization for hyperparameter tuning.
- Automated end-to-end workflows for training, evaluation, and deployment, boosting operational efficiency.
- Designed OCR-powered document management solutions leveraging deep learning for enterprise automation.
- Employed Pinecone for similarity search and nearest-neighbor retrieval in high-dimensional vector spaces.
- Created rich visualizations of model performance and training metrics using Matplotlib and developed analytics dashboards in Tableau for strategic insights.
- Built testing frameworks using Pytest and conducted regular A/B tests and statistical analysis to measure business impact.
- Led technical workshops on Generative AI and Computer Vision, fostering team innovation and cross-functional collaboration.
- Partnered with engineers and data scientists to deliver ML solutions for customer segmentation, marketing optimization, and financial forecasting.

Technical Tools and Environment: Python, PyTorch, Pandas, NumPy, Scikit-learn, Matplotlib, DistilBert, Transformers, Hugging Face, Git, GitHub, Jupyter Notebook, Anaconda, Amazon Web Services (AWS), Pyspark, CI/CD (Docker, Kubernetes)

JP Morgan Chase , Bangalore, India **Dec 2019 – Jan 2021**
Machine Learning Engineer

Responsibilities:

- Developed a machine learning model to automate the identification of root causes for failed events on self-checkout machines (POS systems).
- Reduced enterprise service tickets by 15% by implementing "Backup as a Service" using AWS Backup, allowing customers to initiate server backups and restores.
- Led a data migration project, transitioning ETL processes from SAS to Python on Azure Databricks.
- Constructed a machine learning model for capacity planning by analyzing historical CPU and disk usage data.
- Implemented automated data ingestion pipelines using AWS Glue and Azure Data Factory, streamlining ETL processes for dashboarding.
- Built data pipelines with Apache Airflow to process data from store checkout devices into BigQuery on Google Cloud.

- Developed a classification model using Random Forest and Logistic Regression to predict the likelihood of customer referrals.
- Productionized machine learning pipelines on GCP using Cloud Composer, BigQuery, and GCP storage buckets.
- Designed an automation process using Docker to manage common configurations and detect drifts across virtual infrastructure.
- Applied mean-variance optimization algorithms, including Markowitz portfolio theory, using Python's scipy.optimize library to construct efficient investment portfolios.
- Implemented anomaly detection algorithms, such as isolation forests and autoencoders in Python, to detect fraudulent financial transactions.
- Developed credit risk models using XGBoost and LightGBM to predict default probabilities for loan applicants.
- Created interactive visualizations using Tableau, matplotlib, ggplot2, and Seaborn to present data analysis results effectively.
- Optimized model serving infrastructure on Databricks for low-latency inference through techniques like model caching, distributed serving, and parallel processing.
- Built data pipelines using Python and Kafka to aggregate data from multiple sources (vCenters, databases, store devices) into Google BigQuery.
- Designed and deployed end-to-end machine learning pipelines on GCP Vertex AI, emphasizing security and compliance.
- Conducted quality analysis testing and validation using Django, ensuring model accuracy and thorough API testing before product launch.
- Leveraged machine learning algorithms, including logistic regression, K-means, and recommendation systems, to extract actionable insights from data.
- Developed machine learning models using Python libraries (Pandas, NumPy, Scikit-learn) and algorithms such as Linear Regression, Logistic Regression, Gradient Boosting, SVM, and KNN.
- Created REST APIs to serve data from BigQuery and Cloud SQL, facilitating seamless data access and integration.
- Provided Agile coaching and training to teams, fostering a shared understanding of Agile principles, practices, and ceremonies for efficient project delivery.

Technical Tools and Environment: Python, R, SQL, NoSQL (BigQuery, Cloud SQL), Pandas, NumPy, Scikit-learn, TensorFlow, XGBoost, LightGBM, Docker, Apache Airflow, Kafka, AWS (Glue, Backup, SageMaker), Azure (Databricks, Data Factory), Google Cloud Platform (GCP) (BigQuery, Vertex AI, Cloud Composer, GCP Storage), Tableau, matplotlib, ggplot2, Seaborn, Django, Flask, REST APIs, Agile Methodologies, SCRUM Process.

Risk Span Tech , **Bangalore, India** **May 2017 – Nov 2019**
Data Scientist

Responsibilities:

- Performed data profiling to analyze traffic patterns, locations, dates, and times using advanced analytics platforms.

- Extracted data from distributed storage systems by writing optimized queries in Apache Spark SQL.
- Conducted preliminary data analysis using Python with Pandas and NumPy for descriptive statistics, data cleaning, and missing value imputation.
- Prepared datasets for predictive models in Google Cloud ML Engine, enabling robust predictive modeling in cloud environments.
- Managed and monitored data pipelines using Apache Airflow to automate workflows and ensure data processing reliability.
- Cleaned data and selected relevant features using Databricks' MLlib in a PySpark environment.
- Modeled complex data structures with deep learning frameworks like TensorFlow and Keras.
- Conducted customer segmentation using hierarchical and K-means clustering in Python to enhance targeted strategies.
- Developed data processing scripts using Python, Scala, and R in cloud-based Hadoop environments such as Amazon EMR.
- Evaluated model performance using metrics like Cross-Validation, Log Loss, ROC Curves, and AUC in Jupyter Notebooks.
- Analyzed traffic data patterns through time-series analysis in R, calculating autocorrelations with various time lags.
- Applied Principal Component Analysis (PCA) for feature reduction, enhancing the efficiency of high-dimensional data analysis.
- Designed business intelligence reports using Power BI and Tableau to predict future trends and inform decision-making.
- Retrieved and transformed data from SQL Server and Oracle databases using ETL tools like Apache NiFi.
- Ensured data integrity and quality after migrations and integrations by creating SQL scripts.
- Collaborated with cross-functional teams to communicate analytical results and support data-driven decision-making.

Technical Tools and Environment: Apache Spark SQL, Python (Pandas, NumPy, NLTK, spaCy), TensorFlow, PyTorch, Apache Airflow, Databricks, MLlib, Google Cloud ML Engine, Keras, Apache Hive, Pig, Apache NiFi, R, Power BI, Tableau, SQL Server, Oracle, Jupyter Notebooks, Scala, Amazon EMR, Principal Component Analysis (PCA).

Education

Nirma University — Gujarat, India
B. Tech, Electronic & Communication Engineering