Ruthvik Mannem

mannemruthvik@gmail.com | +1-314-906-1731 | linkedin.com/in/mruthvik

SUMMARY

- Experienced Machine Learning Engineer with 5+ years of expertise in deep learning, computer vision, NLP, and GenAI delivering AI-driven solutions that enhance business performance.
- Proven ability to design, train, and deploy high-performance models across various domains for tasks such as object detection, image classification, regression, sentiment analysis and Generative AI.
- Expert in cloud-based AI deployments and MLOps practices, utilizing AWS SageMaker to create scalable, automated workflows with CI/CD pipelines, Micro Services, Docker, and Kubernetes.

SKILLS

- **Programming Languages:** Python, Java, C++, R
- **Areas of Expertise:** Deep Learning, Computer Vision, NLP, AI Algorithms, LLMs & GenAI
- **Cloud & DevOps**: AWS SageMaker, Azure Machine Learning, GCP, Docker, Kubernetes, MLOps, CI/CD
- **Frameworks & Tools**: TensorFlow, PyTorch, JAX, Scikit-Learn, Keras, SpaCy, OpenCV, NLTK, Pandas, NumPy, PySpark, Kubeflow, MLFlow, SQL, Postgres, FastAPI, Open AI, RAG, Langchain, Streamlit, JavaScript, ReactJS, NodeJS

WORK HISTORY

Software Engineer | Saint Louis University, St. Louis ***Jun 2023 – Present***

- Tech-Lead for a team of four, actively contributing to development while driving AI-driven solutions by conducting code reviews, enforcing best practices, and fostering collaboration.

Machine Learning Engineer | IBM INDIA Pvt Ltd, Bangalore ***Mar 2020 – Aug 2022***

- Spearheaded end-to-end development of machine learning models using **Python** and **C++**, including data collection, preprocessing, model training, evaluation, performance optimization, inference, and deployment to cloud.
- Collaborated with clients, stakeholders, and cross-functional teams to align project goals and gather requirements, ensuring the solutions effectively met business objectives.

PROJECTS

Chatbot using RAG

- Developed an AI-powered chatbot using Streamlit, integrating Langchain for dynamic dialogue management and Llama 2 for LLM-based conversational responses.
- Implemented RAG techniques to enhance the chatbot's ability to provide accurate, context-aware answers by leveraging OpenAI APIs and Pinecone for vector database.
- Built a scalable architecture with a focus on performance, leveraging Langchain workflows to optimize agentic interactions and improve system responsiveness.

Video Transcript Generator Using LLM

- Developed a YouTube transcript generator using Streamlit, Gemini API, and yt-dlp, automating large-scale transcript extraction and processing, reducing manual effort by 80%.
- Integrated Langchain and RAG to ensure contextually accurate and efficient transcript generation, leveraging real-time APIs to improve workflow efficiency for researchers and content creators by 40%.

BubbleScan-AI

- Innovated an Open-Source AI application, achieving 100% accuracy in identifying and analyzing bubbled responses from scantron sheets, reducing manual grading effort by 95%.
- Designed and implemented AI modules in Python, utilizing OpenCV for image preprocessing, TensorFlow and PyTorch for deep learning, and Scikit-Learn for model evaluation.
- Optimized neural networks with advanced image preprocessing, achieving 98% accuracy in image classification using CNNs and developed a Transformer-based NLP model for handwritten text analysis, improving text extraction accuracy by 85%.

- Introduced Agile and DevOps methodologies, streamlining the CI/CD pipeline with GitHub Actions, which increased project completion speed by 30% and reduced manual deployment intervention by 70%, ensuring timely release of sprints.

Predictive Maintenance with Anomaly Detection

- Engineered an end-to-end AI based predictive maintenance application for industrial equipment, encompassing data collection, preprocessing, model training, evaluation, and deployment.
- Employed machine learning algorithms to predict equipment failures with 85% accuracy, scheduling maintenance proactively that reduced downtime by 40% and lowered maintenance costs by 25%.
- Architected micro services for real-time anomaly detection and predictive analytics systems using Python, Scikit-Learn, and PySpark, efficiently processing large volumes of industrial data on Azure Databricks.
- Refined model performance through extensive validation and tuning in multi-node environments, utilizing GPU acceleration with CUDA, cutting training time by 40%.
- Optimized deployment using MLOps practices including model KPI, integrating CI/CD pipelines, Docker, Kubernetes, and Azure Machine Learning for scalable, robust operations, boosting system uptime by 20%.

Automated Warehouse Management System

- Spearheaded the development of machine learning models in Python to automatically identify labels and key information from drone-captured frames, improving operational efficiency by 65%.
- Utilized TensorFlow, PyTorch, and Scikit-Learn for data wrangling, model training, and management in a distributed AWS environment, achieving a 30% reduction in model training time.
- Implemented ResNet and YOLO for real-time object detection, enhancing application performance by 35% and enabling more accurate analysis and handled time series data, signal, image, and video processing.
- Optimized GPU tasks with CUDA, accelerating computational performance by 40%, and developed REST APIs integrated with AWS Lambda for scalable execution and seamless integration.
- Evaluated model performance using metrics like accuracy, precision, recall, and F1-score, and conducted cross-validation to ensure robustness, improving reliability by 25%.
- Streamlined CI/CD pipelines with Git, AWS CodePipeline, Docker, and Kubernetes, ensuring robust and scalable deployments in a distributed environment.

Attention-Based Deep Driving Model for Autonomous Vehicles

- Re-engineered an advanced deep learning model for autonomous vehicle navigation utilizing multi-view camera data, leveraging Python and TensorFlow.
- Synchronized front, rear, and side-view camera inputs, employing attention mechanisms to enhance semantic understanding of environmental cues.
- Designed and validated the Multi-View Attention module, improving model performance in vehicle steering and speed prediction by 20%.
- Conducted in-depth model training with the Drive360 dataset, surpassing existing models in Mean Square Error metrics.

End-User Sentimental Analysis

- Developed a sophisticated sentiment analysis platform to assess customer feedback and reviews, utilizing advanced NLP techniques to derive actionable insights.
- Implemented data preprocessing, feature extraction, and model development using Python, NLTK, and Scikit-Learn, achieving 92% accuracy in sentiment classification.
- Applied tokenization, stop-word removal, and TF-IDF vectorization to enhance the quality of textual data, improving model performance.
- Conducted model validation and tuning, optimizing hyperparameters to ensure robustness and reliability across diverse datasets.

EDUCATION

MS in Artificial Intelligence | Saint Louis University, St. Louis, MO. ***Aug 2022 – May 2024***

MS in Computer Applications | Visvesvaraya Technological University ***Aug 2016 – Jun 2019***

Bachelors in Computer Science | Krishna University ***Jul 2013 – Mar 2016***