Harikrishna Marampelly

AI/ML Engineer

Tampa, FL | +1 (656)204 5806 | harikrishnamarampelly4@gmail.com | LinkedIn

SUMMARY

AI/ML Engineer with 4+ years of experience in designing, building, deploying, and optimizing machine learning models and AI-driven solutions. Proficient in end-to-end ML pipelines, including data preprocessing, feature engineering, model development, deployment, and monitoring, leveraging advanced frameworks such as TensorFlow, PyTorch, Scikit-learn, and Keras. Skilled in Natural Language Processing (NLP), Computer Vision, and time-series analysis, with expertise in cutting-edge techniques like transformers, BERT, GPT models, convolutional neural networks (CNNs), and generative AI approaches such as Stable Diffusion, GANs, and DALL·E. Experienced in implementing AI systems for predictive analytics, personalization, recommendation engines, and Retrieval-Augmented Generation (RAG) solutions using tools like LangChain. Proficient in cloud platforms, including AWS SageMaker, Azure Machine Learning, Google AI, and Databricks, for scalable deployment and Big Data technologies like PySpark and Hadoop for distributed data processing. Adept in MLOps practices with tools like MLflow, Kubeflow, Airflow, Docker, and Kubernetes for automation, model versioning, and monitoring, as well as integrating AI models with APIs and microservices to deliver impactful enterprise solutions. Skilled in Python, R, JavaScript, SQL, and NoSQL databases, and well-versed in emerging areas such as Explainable AI (XAI), AutoML, and reinforcement learning.

WORK EXPERIENCE

Cigna Healthcare, USA | AI/ML Engineer — May 2024 - Current

- Designed and deployed AI agents that served as virtual therapists, utilizing Generative AI (Gen AI) models to assist patients with aphasia and speech disorders, incorporating cutting-edge AI techniques.
- Built a custom lightweight Large Language Model (LLM) tailored to patient data, enabling contextually aware and empathetic agent responses.
- Fine-tuned existing LLMs, including GPT-4 and Gemini, to guide the behavior and tone of the AI agent, ensuring responses aligned with therapeutic goals.
- Implemented speech-to-text and text-to-speech APIs to enable a seamless voice interface, improving accessibility for patients with speech impairments.
- Utilized Hugging Face Transformers for NLP tasks such as text generation, entity recognition, and sentiment analysis to enhance conversational capabilities.
- Integrated OpenCV to analyze and interpret visual data, allowing the AI agent to process non-verbal cues for a more holistic interaction.
- Applied reinforcement learning techniques to improve agent adaptability and interaction outcomes based on patient feedback and engagement patterns.
- Implemented a RAG system to retrieve relevant patient data and provide personalized and informed responses from the AI agent.
- Combined speech, text, and visual data in a multi-modal learning framework to enable comprehensive understanding and interaction with patients.
- Built a user-friendly frontend using React Native, ensuring accessibility and a responsive design tailored to patient needs.
- Designed and maintained scalable backend APIs with Node.js, enabling secure and efficient data flow between the AI agent, databases, and user interfaces.
- Deployed AI models on cloud platforms such as AWS and Azure, ensuring scalability, reliability, and HIPAA-compliant data security for healthcare applications.
- Utilized Pandas and NLTK to preprocess patient datasets, including cleaning, tokenization, and feature extraction, ensuring high-quality inputs for AI models.

Byteworks Solutions, India | AI/ML Engineer — Aug 2019 - Dec 2022

- Designed and implemented a recommendation system to provide personalized financial recommendations by analyzing individual spending patterns and investment preferences, boosting user engagement by 30%.
- Utilized Pandas and NumPy for data preprocessing and statistical analysis of transactional and investment data to uncover actionable insights.
- Applied reinforcement learning techniques to create adaptive portfolio optimization models, improving long-term financial performance for users.

- Designed an intuitive financial management dashboard using React, enhancing user experience and accessibility across multiple devices.
- Built and deployed secure and efficient backend APIs using Node.js to handle real-time data communication and integration with external financial services.
- Leveraged BERT and word embeddings for natural language understanding, enabling accurate sentiment analysis and text classification in user interactions.
- Developed Transformer-based architectures, such as BERT and attention mechanisms, to enhance the contextual accuracy of user-specific recommendations.
- Designed sentiment analysis workflows using DistilBERT-SST to assess user sentiments, tailoring financial advice based on emotional context.
- Utilized Graph Sage and Graph Neural Networks (GNNs) to analyze relationships between financial assets and provide network-based recommendations.
- Employed Bi-LSTM networks to process time-series financial data, enabling predictive analytics for spending trends and investment forecasts.
- Applied Convolutional Neural Networks (CNN) to identify and classify recurring patterns in financial data for fraud detection and anomaly detection.
- Created data visualizations to present key financial metrics and insights using Python libraries like Matplotlib and Seaborn.
- Ensured high-quality data inputs by implementing advanced data preprocessing techniques, including normalization, outlier detection, and feature engineering.
- Enhanced model performance and interpretability by integrating attention mechanisms into neural network architectures for financial data analysis.
- Deployed AI models in cloud environments for real-time inference, ensuring scalability and low-latency performance in production.
- Designed and maintained secure APIs and data pipelines, adhering to financial regulations and data protection standards like GDPR and RBI guidelines.
- Worked with data engineers, UX designers, and financial experts to deliver an end-to-end solution that addressed business requirements and enhanced customer satisfaction.

TECHNICAL SKILLS

- **Methodologies**: SDLC, Agile, Waterfall
- **Languages**: Python, C++, Scala, GoLang.
- **AI/ML Domain**: Natural Language Processing, Computer Vision, Image Processing, Deep Learning, Machine Learning, Reinforcement Learning, Expert Systems.
- **AI/ML Technologies**: Large language models (LLMs), Retrieval Augmented Generation (RAG), Generative AI, Pytorch, TensorFlow, Keras, OpenNN, TorchVision, OpenCV, NLTK, SpaCy, Hugging Face, LangChain, Model Interpretability, Explainable AI, scikit-learn, Tensorboard, OpenAI Gym, Pandas, Matplotlib, Statistical Analysis, ChatGPT, BERT, Fine Tuning, Prompt Engineering, Conversational AI, Ranking, query classification, Named Entity Recognition (NER), Gradient Boosting Trees, PCA, Information Retrieval, Seq2Seq, Knowledge Graph, Linear Regression, Logistic Regression, Random Forest, Association Rules, Support Vector Machine, ggplot2, Pandas.
- **Software Technologies**: SQL, MySQL, MongoDB, RESTful API, Scala Play, FastAPI, Swagger, Flask, AWS, Sage Maker, ElasticSearch, Agile, CI/CD, Design Patterns, Concurrency, Git, GitLab, Linux, Windows, A/B Testing.
- **Operating System**: Windows, Linux

EDUCATION

Master's in Computer Science | University of South Florida, Tampa, Florida

Bachelor's in Computer Science | Osmania University, India

CERTIFICATION

ServiceNow Certified Application Developer.