SUPUN LAKSHAN

077 0891666 supun.ud@outlook.com LinkedIn GitHub

AI ENGINEER

Results-driven AI Engineer with a proven ability to develop and deploy innovative AI solutions that drive business success. Experienced in creating advanced models that enhance efficiency, optimize processes, and solve complex problems across industries. Demonstrated leadership in AI research and development, improving system scalability, reducing processing times by 40%, and increasing forecast accuracy by 25%. Adept at mentoring teams and streamlining workflows, contributing to significant operational improvements and cost savings.

PROFESSIONAL EXPERIENCE

IT Pathfinders (PVT) Ltd — Pannipitiya

Associate AI Engineer — ***Dec 2023 - Present***

- **AI Research and Development:**Spearheaded the design, training, and deployment of cutting-edge AI models in computer vision and generative AI. Delivered scalable, real-time detection and tracking systems with precision optimization, fine-tuning LLMs with Hugging Face, TensorFlow and PyTorch to automate content creation, reducing manual work by 50%. Deployed solutions via AWS and GCP to enhance efficiency and accuracy.
- **ML Pipeline Architecture:** Built cloud-based ML pipelines with Docker, AWS SageMaker, and Kubernetes, reducing processing time by 40% while improving scalability and system reliability. Enhanced predictive models using ensemble methods, boosting business forecast accuracy by 25%.
- **Team Leadership and Mentorship:** Led and mentored a team of junior engineers, driving innovation through AI/ML best practices, DevOps, and Agile methodologies, fostering collaboration and continuous development.

AI & ML Engineer - Intern — ***June 2023 - Nov 2023***

- **Developed Advanced Computer Vision Systems:** Designed and deployed sophisticated computer vision projects using state-of-the-art models such as YOLOv8, YOLOv10, and Faster R-CNN, combined with multi-object tracking algorithms (ByteTrack, Deep SORT, FairMOT) to achieve high-precision object detection, tracking, and counting.
- **Developed Advanced Generative AI and NLP Solutions:** Engineered and deployed Retrieval-Augmented Generation (RAG) chatbots leveraging LLMs like GPT, LLaMA 2, and Falcon. Utilized LangChain, Hugging Face, and vector databases (e.g., ChromaDB, Faiss, Pinecone) to enhance conversational AI and real-time NLP applications, delivering high-impact, industry-specific solutions.
- **Optimized Workflows with Agile Practices:** Streamlined development processes by implementing Agile methodologies, continuous integration/continuous deployment (CI/CD) pipelines, and version control using Git, accelerating project delivery and improving team coordination.

Freelancing-Upwork — USA

AI & ML Engineer — ***2023-PRESENT***

- **Implemented Intelligent RAG Chatbot:** Designed and deployed a Retrieval Augmented Generation (RAG) chatbot for the gcore.com platform, enhancing user engagement through AI-driven, knowledge-based interactions. Leveraged the OpenAI API for high-performance embeddings and integrated Flask for seamless API delivery.
- **Optimized Content Generation Workflows:** Developed and fine-tuned workflows for vid2vid, txt2img, and txt2vid using ComfyUI, automating and streamlining content creation processes, thereby increasing productivity and reducing manual effort.
- **Led Project Management and Collaboration:** Directed team collaboration efforts using Jira for project management and Git for version control, improving project timelines, communication, and overall efficiency across distributed teams.

TECHNICAL SKILLS

- **Programming Languages:** Python, C++, SQL, HTML/CSS
- **Artificial Intelligence:** Machine Learning, Deep Learning, Natural Language Processing, Computer vision, Generative AI & LLMs (including GPT, Gemini, LLaMA 2, Falcon LLM)
- **Core AI Concepts:** Neural Network Architectures, Algorithm Optimization, Statistical Analysis, Data Structures and Algorithms, Model Evaluation and Validation
- **Technologies & Frameworks:** TensorFlow, Keras, PyTorch, scikit-learn, Hugging Face, LangChain, OpenCV, CNN, RNN, Transformer Networks, RAG techniques, YOLO, Deep SORT, ByteTrack, FairMOT, Ollama
- **MLOps & Deployment:** MLflow, Kubeflow, Docker, AWS, CI/CD, FastAPI, Flask, AWS SageMaker
- **Tools & Platforms:** Linux/Ubuntu, Git & GitHub, Jira, Streamlit, Faiss, Pinecone, ChromaDB, Midjourney

ACADEMIC QUALIFICATION

Kingston University — **United Kingdom**
BSc (Hons) Computer Science in Software Engineering (TOP UP) — *2024-PRESENT*

University of Moratuwa (ITUM) — **Sri Lanka**
Information Technology — *2019-PRESENT*

Dharmadutha National School — **Badulla, Sri Lanka**
Passed the GCE Advanced Level Examination (Maths Stream) — *Feb. 2014 – Dec. 2016*

PROJECT EXPERIENCE

Smart Surveillance System - ***Python, Computer Vision, YOLOv8, OpenCV, PyTorch, Tracking Algorithms (Deep SORT, ByteTrack, FairMOT)*** **July 2024**

- Developed an advanced surveillance system for real-time human detection and tracking using Python, custom-trained YOLOv8, OpenCV, and PyTorch.
- Implemented Deep SORT, ByteTrack, and FairMOT as tracking algorithms, and used ResNet-50 for feature extraction, cosine similarity, and Kalman filtering to enhance tracking accuracy and Re-Identification across multiple camera views.
- Designed region-based analysis features, including object counting and waiting time calculation, to provide insights into crowd density and movement patterns.
- Integrated walk flow detection to visualize and analyze movement within surveillance areas, improving public space management.

End-to-End-Medical-Chatbot-Using-Llama2 - ***Python, Generative AI, Llama2, Langchain , Pinecone, sentence-transformers, Flask, AWS*** **April 2024**

- Developed a high-precision medical chatbot powered by Llama2 and Langchain, offering accurate, real-time medical advice through NLP and AI-driven interactions.
- Deployed the chatbot with Flask and AWS, ensuring a scalable, secure platform capable of handling high traffic while maintaining low-latency responses.
- Enhanced user interaction and data management with advanced NLP capabilities, leveraging Flask for the web framework and AWS for scalable deployment.

End-to-End-Chest-Cancer-Classification-using-MLflow-DVC - ***Python, TensorFlow, Keras, MLflow, VGG16, Docker, AWS, DVC, CI/CD Pipeline, Flask, HTML*** **Mar 2024**

- Developed a robust solution for classifying chest cancer cases with 92% accuracy.
- Utilized MLflow for experiment tracking, Docker for containerization, and deployed on AWS with Flask web app integration.

Fake-News-Classification-Using-RNN - ***Tensorflow, keras, LSTM, nltk, One Hot Encoding*** **Jun 2024**

- Developed a high-performance fake news detection system using LSTM networks, achieving 91% classification accuracy by processing large volumes of textual data.
- advanced natural language processing techniques, including tokenization with NLTK and one-hot encoding, to efficiently handle and analyze unstructured news articles.
- Optimized model performance through hyperparameter tuning and deployed the solution in a scalable environment, providing a robust tool for detecting misinformation in real time.