

Amodh Herath

Associate AI & ML Engineer

076-875-8499 | amodhwork@gmail.com | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)

Summary

AI and ML Engineer with hands-on experience in machine learning, deep learning, NLP, computer vision, and production AI systems. Experienced in developing and evaluating models using Python, PyTorch, TensorFlow, and Scikit-learn, with additional experience integrating ML capabilities through FastAPI and scalable backend architectures.

EDUCATION

University of Jaffna, Faculty of Engineering

Apr 2022 - Aug 2026

Bachelor of Science in Computer Engineering Honours

Final Year Research: Sinhala Emotion Classification | Python, PyTorch, BERT, Scikit-Learn

- Architected a hybrid deep learning model combining a pre-trained SinBERT transformer with a custom Convolutional Neural Network (CNN) for multi-class emotion (Ekman's Basic Emotions) classification of Sinhala text.
- Engineered a robust fine-tuning pipeline featuring 5-fold stratified cross-validation, dynamic layer freezing, and automated hyperparameter tuning to optimize performance scores.
- Published and presented research findings at the Undergraduate Symposium "Computing Beyond Boundaries", Department of Computer Engineering, University of Jaffna.

EXPERIENCE

AI Fullstack Engineering Intern

Dec 2025 - May 2026

Dilexus Infotech, Colombo, Sri Lanka

- Created a local document data extraction pipeline using Docling, Small Language Model (SLM), Ollama and FastAPI for an air-gapped system and subsequently published a technical deep dive on the system.
- Optimized resource management across the extraction service to enable controlled LLM loading and mitigate memory-related issues in production, making it suitable for resource-constrained environments,
- Developed responsive Next.js UIs, specifically tuning LCP and TTFB metrics to achieve improvements in web vitals under simulated production loads.
- Built a high-performance e-commerce storefront in Next.js, leveraging Supabase edge functions for server-side business logic and pgvector for ML-driven product similarity search, improving semantic discovery.

OPEN-SOURCE CONTRIBUTIONS

Lanka Data Foundation | [RTI-Tracker](#)

- Contributed to core repository stability by engineering fail-fast database connection logic using bounded retry, reducing startup failures in unstable DB conditions.
- Designed comprehensive unit test suites to simulate key issues to validate bounded retry thresholds.

PERSONAL PROJECTS

South Carolina Real Estate Price Prediction | Python, XGBoost, Scikit-Learn, Pandas

- Developed an end-to-end machine learning pipeline for real estate price prediction, using Random Forest-based imputation to reconstruct missing historical data.
- Engineered data preprocessing and feature selection workflows, leveraging exploratory data analysis (EDA) and correlation analysis to identify high-value features in noisy datasets.
- Trained and optimized an XGBoost regression model using cross-validation and hyperparameter grid search, evaluating performance with RMSE, MAE, and R².

Centurion (In Progress) | Java 21, Spring Boot, Python, TypeScript, LangGraph, Kafka, Temporal, Next.js

- Building a polyglot AIOps platform to automate incident response, leveraging Event-Driven Architecture (Kafka) for real-time error ingestion and Durable Execution (Temporal) for fault-tolerant remediation workflows.
- Prototyping a Retrieval-Augmented Generation (RAG) investigation pipeline using Python, LangGraph, and ChromaDB to analyze stack traces, governed by a human-in-the-loop Next.js approval dashboard.
- Integrating Llama 3 to dynamically compare incoming error logs against vectorized internal runbooks, autonomously generating actionable remediation plans to minimize Mean Time To Recovery (MTTR).

Janus (In Progress) | Go, Python, gRPC, Protocol Buffers, Docker

- Engineering a low-overhead AI Security Gateway and reverse proxy in Go, designed to secure LLM workflows with real-time prompt injection defence and streaming Personally identifiable information (PII) detection and redaction.
- Architecting a hybrid inter-process inspection loop via gRPC and Protocol Buffers, bridging a high-throughput Go networking plane with a Python machine learning engine for semantic security enforcement.
- Integrating Small Language Models (SLMs) within the Python engine to rapidly evaluate adversarial jailbreak attempts, deliberately optimizing model size to prevent bottlenecks in networking throughput.

Sortify (Computer Vision Platform) | Python, FastAPI, React, TensorFlow/Keras

- Developed an end-to-end computer vision platform classifying waste into 12 categories, achieving a 92% macro-averaged F1-score through a soft-voting ensemble of fine-tuned MobileNetV2 models.
- Engineered a high-performance Python FastAPI backend for real-time synchronous inference, integrated with a dynamic React frontend for image upload and classification history visualization.
- Implemented advanced digital image processing (CLAHE, YCrCb color space conversion) and a rigorous 5-fold cross-validation training pipeline to maximize feature extraction robustness.

Sazzler (E-Commerce Microservices) | Java, Spring Boot, Kafka, PostgreSQL, MongoDB

- Architected a highly scalable, event-driven microservices backend utilizing Spring Boot, Spring Cloud Gateway, and Eureka for dynamic API routing and service discovery.
- Implemented a polyglot persistence strategy, strategically decoupling unstructured product catalogs in MongoDB from strictly relational user and order states in PostgreSQL.
- Engineered an asynchronous data synchronization loop via Apache Kafka and utilized Spring State Machine to safely orchestrate complex distributed transactions and order lifecycles.

Leadership & Activities

Team Yarl Racing – Formula Student University of Jaffna

Low Voltage Sub-Team Lead 2024 – 2025

- Designed the Brake System Plausibility Device (BSPD) and the low-voltage electrical system architecture, collaborating cross-functionally to integrate the overall system

Financial Manager 2025 – 2026

- Collaborated with US-based technical advisors and controlled budget allocation across powertrain, aerodynamics, and electrical sub-teams to maximize performance within strict competition regulations.
- Organized fund-raising events for the project.

TECHNICAL SKILLS

Programming: Python, Java, Go, TypeScript	Backend Frameworks: FastAPI, Spring Boot, REST, gRPC
Frameworks & Libraries (AI/ML): PyTorch, TensorFlow/Keras, Scikit-learn,	Distributed Systems: Kafka, Temporal, Microservices
ML/AI Domains: NLP, Computer Vision, Transformers, RAG, LLMs, Agentic AI	Cloud/DevOps: AWS, OCI, Docker, GitHub Actions
Databases: PostgreSQL, MongoDB, pgvector, ChromaDB	Data & Evaluation: Pandas, NumPy, EDA, Cross-Validation, Hyperparameter Tuning, RMSE, MAE, R², Precision, Recall, F1-score, Confusion Matrix