Harsha Vardhan Yellela

United States | +1-248-497-9965 | harsha.yellela@gmail.com | LinkedIn | GitHub

SUMMARY

Machine Learning Engineer with hands-on experience in **LLM fine-tuning (QLoRA, PEFT)**, **deep learning (LSTM, CNN, GNN)**, and **MLOps pipelines**. Built production ML systems using **PyTorch, TensorFlow, HuggingFace Transformers**, and **AWS SageMaker**. Skilled in **NLP, computer vision, time-series forecasting**, and deploying **scalable inference services** with **auto-retraining feedback loops**.

EXPERIENCE

Graduate Research Assistant – Agentic AI **Jan 2025 – Present**
Lawrence Technological University *Southfield, MI*

- Built **multi-agent AI systems** using **CrewAI + LangChain** with **RAG pipelines**, integrating **OpenSearch Serverless Vector Engine** for **semantic search** and document retrieval.
- Deployed **persistent ML agent services** on **AWS Fargate** and **Amazon EKS**, combining **Bedrock-hosted LLMs** with custom inference tools.
- Designed **hybrid AI pipelines** achieving **70% reduction in manual process time** through intelligent workflow automation and model orchestration.

Infor India Pvt. Ltd. **Apr 2022 – Dec 2023**
LN Technical Consultant *Hyderabad, India*

- Developed **modular, production-ready tools** for global clients (**Ferrari, Boeing, Triumph**) by extending **Infor LN ERP** workflows for inventory and manufacturing.
- Integrated **AWS S3, Lambda, and API Gateway** with enterprise systems to enable **asynchronous data pipelines**, reducing **batch processing delays by ~40%**.
- **Containerized** business logic services using **Docker** and orchestrated deployments, improving **consistency** across client environments.

TECHNICAL SKILLS

- **ML Frameworks:** PyTorch, TensorFlow/Keras, PyTorch Geometric, HuggingFace Transformers, scikit-learn
- **Deep Learning:** CNN, LSTM, GRU, VGG19, Graph Neural Networks (GCN), Transfer Learning, Attention Mechanisms
- **LLM & Fine-tuning:** QLoRA, PEFT, TRL, 4-bit Quantization (NF4), Prompt Engineering, Instruction Tuning
- **NLP:** Sentiment Analysis (VADER, BERT), NER, Text Preprocessing, Embeddings, RAG Systems
- **Computer Vision:** Image Classification, Medical Imaging, OpenCV, Data Augmentation, YOLO, CLIP
- **MLOps & Cloud:** AWS SageMaker, Bedrock, ECS Fargate, Lambda, S3, Docker, Kubernetes, Terraform
- **Languages:** Python, SQL, Go, TypeScript | **Data:** Pandas, NumPy, DynamoDB, PostgreSQL, Qdrant

PROJECTS

Resume Optimizer – QLoRA Fine-tuned LLM for ATS Optimization **Oct 2024 – Present**
PyTorch, Transformers, PEFT, QLoRA, FastAPI, Ollama | GitHub

- Fine-tuned **Qwen3-4B** using **QLoRA (4-bit NF4 quantization)** with **LoRA rank 16, alpha 32**, reducing GPU memory to **18-22GB peak VRAM** for 4B parameter model.
- Processed **1,800+ resumes** and generated **1,304 training examples** with ATS scores; achieved **9.5/10 quality score** (GPT-5.1 eval) vs 9/10 baseline.
- Built **FastAPI inference endpoint** with **deterministic JSON output** (temperature=0.0), generating optimized resumes in **3-5 seconds** on RTX 3090.

ML Sentiment Feedback Loop – Production MLOps Platform **Nov 2024 – Present**
PyTorch, HuggingFace, SageMaker, ECS Fargate, Terraform, FastAPI | GitHub

- Designed **8-microservice MLOps architecture** with **auto-retraining pipeline**: API Gateway, Inference, Feedback, Model Registry, Evaluation, Retraining, Notification, and Model Init services.
- Deployed **twitter-roberta-base-sentiment** model on **AWS ECS Fargate** with **SageMaker training jobs**, automated **model versioning**, and **CI/CD via GitHub Actions + Terraform**.
- Implemented **complete feedback loop**: user corrections trigger **model evaluation → retraining → auto-deployment** with version tracking in Model Registry.

Sentiment-Driven Market Forecasting – LSTM for NVDA Returns **May 2025**
TensorFlow, LSTM, VADER, SageMaker, FinSpace, Pandas

- Engineered **multimodal pipeline** combining **Reddit sentiment signals** (VADER + virality scoring) with **financial indicators (OHLCV, VIX)** to predict **NVDA's next-day returns**.
- Trained **7-day rolling LSTM model** in **Amazon SageMaker**, achieving **measurable correlation against market baselines** while reducing **local experimentation time by 60%**.
- Used **Amazon FinSpace** for **time-series alignment** and **high-resolution feature aggregation** across sentiment and pricing data sources.

Pneumonia Detection – Deep Learning for Chest X-ray Classification — Dec 2024

TensorFlow, VGG19, AWS SageMaker, Flask, OpenCV | GitHub

- Built **pneumonia detection CNN** using **VGG19 with custom classifier layers**, applying **transfer learning** and **image augmentation** for medical image classification.
- Trained on **AWS SageMaker**, reducing **training time by 60%** compared to local GPU while scaling to larger datasets with improved validation accuracy.
- Deployed as **Flask web app** with **real-time inference** and interactive diagnosis interface; participated in **RSNA Pneumonia Detection Challenge (upper quartile)**.

Traffic Flow GNN – Anomaly Detection with Graph Neural Networks — Nov 2024

PyTorch Geometric, GCN, SUMO, TraCI, NetworkX | GitHub

- Built **end-to-end traffic analysis pipeline** from simulation to anomaly detection using **SUMO traffic simulator** with **TraCI interface** for real-time data collection.
- Designed **graph-based network representation** (nodes=intersections, edges=roads) and implemented **2-layer Graph Convolutional Network (GCN)** for anomaly scoring.
- Created **modular PyTorch Geometric architecture** extensible to real-world traffic data; demonstrated novel application of **GNNs to transportation systems**.

Food Image Classifier – EA-Optimized CNN Hyperparameter Tuning — Dec 2024

TensorFlow/Keras, DEAP, CMA-ES, Keras Tuner | GitHub

- Compared **6 hyperparameter optimization methods** for CNN: ES(1+1), CMA-ES, Keras Tuner Hyperband, DEAP, Random Search, and baseline.
- Achieved **79.80% accuracy with ES(1+1)** vs 72.47% baseline (**+10.1% improvement**); CMA-ES reached 77.07%, Hyperband 79.27%.
- Implemented **pure Python CMA-ES** with dual step-size adaptation; created **comprehensive benchmarking framework** for evolutionary optimization research.

EDUCATION

Lawrence Technological University — Expected Dec 2025

Master of Science in Computer Science · GPA: ***3.6/4.0*** — *Southfield, MI*

- **Relevant Coursework: Machine Learning**, **Deep Learning**, **Natural Language Processing**, **Artificial Intelligence**, **Intelligent Robotics (ROS)**, **Evolutionary Computation**

Geethanjali College of Engineering & Technology — Graduated: August 2022

Bachelor of Technology in Computer Science & Engineering · GPA: ***7.5/10 (~3.0/4.0)*** — *Hyderabad, Telangana*

- **Relevant Coursework: Deep Learning & Python**, **Machine Learning Foundations**, **Software Engineering**, **Internet of Things**

ACHIEVEMENTS

- **Selected for Amazon Nova AI Challenge: Trusted AI Track** (2025)
- **RSNA Pneumonia Detection Challenge** – Ranked in upper quartile using VGG19 transfer learning (2024)
- **Gold Medalist in Indian National Mathematical Olympiad (INMO)** (2012)