DURGARAM R

Chengalpattu, Tamil Nadu | +91 93606 35207 | durgaramesh641@gmail.com
linkedin.com/in/durgaramr | github.com/Durgaram26 | Portfolio: www.durgaram.in

AI & ML Engineer

Professional Summary

AI & ML Engineer who designs, trains, and ships models and intelligent systems for production. Hands-on with **classical ML**, **deep learning** (**PyTorch**, **TensorFlow**), **computer vision**, **NLP**, and **LLM**/**RAG** (fine-tuning, retrieval, agents). Builds preprocessing and evaluation pipelines in Python (**Pandas**, **NumPy**), serves models with **FastAPI**/**Flask**, and optimizes **GPU** inference (**vLLM**, **CUDA**). Deploys end-to-end with **Docker** and **CI/CD** on **AWS**, **Azure**, and **GCP**; cares about accuracy, latency, drift, and reliability.

Education

Karpaga Vinayaga College of Engineering and Technology
*B.Tech — **Artificial Intelligence and Data Science*** *2022 – 2026*

- CGPA: **8.21 / 10.0**

Experience

AI & ML Engineer 01/2026 – Present
Melmaruvathur Adhiparasakthi Institute of Medical Sciences & Research (MAPIMS) *Melmaruvathur, Tamil Nadu*

- Shipped a **discharge summary** product: **fine-tuned LLM** generates structured clinical reports from patient notes, cutting doctor documentation time by **70%**.
- Delivered a hospital **Ticket Management System (TMS)** with **NLP**-based routing and prioritization, replacing manual triage and cutting response time by **50%**.
- Automated hospital reporting with **NLP**-based structuring and validation over **patient records**—cutting manual entry and stabilizing clinical fields for downstream analytics.

Technical Skills

Languages: Python, HTML5, CSS3, **SQL**

Frontend & Backend: React.js, **Node.js**, **Flask**, **FastAPI**, REST APIs, Nginx

ML Engineering & Modeling Workflow: Dataset prep, augmentation, labeling workflows, train/val/test splits, metrics & error analysis; **Pandas**/**NumPy**; iterative experimentation and production handoff (**Flask**/**FastAPI**)

AI / ML / DL: Scikit-learn, **TensorFlow**, **PyTorch**, Keras, **OpenCV**, Hugging Face Transformers, **YOLOv8**, **Whisper**, MediaPipe, Wav2Lip, Tesseract OCR

LLM & Generative AI: LangChain, **LangGraph**, **CrewAI**, **LlamaIndex**, **RAG**, LLM Agents, Prompt & Context Engineering, **OpenRouter**, **Ollama**, **vLLM**, Qwen, GPT APIs, Claude CLI, Claude Code

Vector DB & Search: FAISS, **ChromaDB**, Pinecone, Semantic Search, Embeddings, page-level citations

Databases: MySQL, MS SQL Server, **PostgreSQL**, **MongoDB**, SQLite

DevOps, Cloud & Deployment: Docker, Linux (Ubuntu/Debian), Git, GitHub, **CI/CD** (GitHub Actions, Azure DevOps), **AWS**, **Azure**, **Google Cloud** (GCP), Hostinger VPS, **NVIDIA CUDA**

Visualization: Matplotlib, Seaborn, **Streamlit** dashboards

Hardware & Edge: Raspberry Pi, Arduino, IoT sensor integration, edge inference optimization

Projects

Clinical ASR — Patient & Doctor Conversations (Vibe Voice) *— Python, Vibe Voice ASR, Streamlit / Web UI, Audio Pipeline* 2026

- Built a conversational **ASR** product for patient–doctor dialogue using **Vibe Voice**, aimed at clinical documentation rather than tabular symptom-to-disease prediction.
- **Audio** ingestion, **ASR** post-processing, review **UI** for transcript verification — tuned for **noisy clinic** environments.

Medical RAG Chatbot — Clinical Q&A (ChromaDB) — *Python, LangChain, LlamaIndex, **ChromaDB**, Embeddings, Page-Level Index* 2026

- RAG-powered clinical Q&A over documents; embeddings and retrieval in **ChromaDB**, responses tied to cited sources.
- Orchestrated with **LangChain** / **LlamaIndex**; page-aware chunks so answers reference the right sections.

Neural OCR Engine — Document Intelligence Platform — *Python, PyTorch, Flask, React, Tesseract OCR, XML Schema Parsing* 2026

- Trained and deployed **deep learning OCR** on scanned & digital **PDF**s with field parsing to custom **XML** schemas; **97%+** field accuracy.
- Productized via **Flask** API + **React** UI: document ingestion, layout-aware extraction, and sidebar navigation for review/export.
- Engineered local **GPU LLM** inference with **Ollama** and **vLLM** (**sub-200ms** latency, **zero** cloud API cost) for downstream document workflows.

Auto Image Trainer Studio — *Python, Flask, PyTorch, OpenCV, Docker* 03/2025

- Web platform for **annotation**, dataset management, and model training — **40%** faster workflows vs. manual pipelines (**Flask**, **PyTorch**, **OpenCV**).
- Live training metrics; **object detection** and **image classification** pipelines.

Vox Persona — 3D AI Interview Bot with Lip Sync — *Python, FastAPI, Qwen LLM, Three.js, Wav2Lip, Whisper, TTS* 2025

- Built a fully **offline** 3D conversational bot with lip-synced avatar using **Wav2Lip** — **zero** cloud dependency, runs on local hardware.
- Two modes: **Conversation Mode** and structured **Interview Mode** with role-specific questions and dynamic follow-ups. **FastAPI**; 3D in **Three.js**.

AI Feedback System — Voice Capture & LLM Summaries — *Python, FastAPI, Whisper STT, OpenRouter, React, MongoDB, Docker* 2025

- End-to-end feedback product: **STT** (**Whisper**) transcribes voice/text input; **OpenRouter** routes multi-model **LLM** calls for structured summaries, themes, and action items.
- Full-stack pipeline—audio ingestion, transcript storage, review UI, and **AI-generated summary** exports for stakeholders (**FastAPI**, **React**, **MongoDB**); deployed with **Docker** + **CI/CD**.

Human Fall Detection System — *Python, OpenCV, TensorFlow/Keras, MediaPipe* 2025

- Vision pipeline that detects falls from camera streams using **pose cues** + temporal logic, triggering alerts for safety monitoring (**OpenCV**, **MediaPipe**).
- **CNN**-based classifier for fall vs. normal activity; tuned for responsive inference on streamed frames.

Vehicle Monitoring & Real-Time Tracking System — *Python, YOLOv8, DeepSORT, OpenCV* 11/2024

- **Multi-vehicle** detection & tracking with **YOLOv8** + **DeepSORT**: **96%** tracking precision at **20+ FPS** (**OpenCV**).

Face Recognition Attendance System — *Python, OpenCV, face_recognition, SQLite, Raspberry Pi* 10/2023

- Deployed on Raspberry Pi, identifying **50 individuals** with **99.2% accuracy** and fully replacing manual attendance logging — **0 manual entries** needed post-deployment.
- **Published Research:** *Face Recognition Based Attendance System Using Anti-Spoofing Methods* — IIP Series (E-ISBN: 978-93-6252-750-9)

Certifications

- **Google Cloud Data Analytics** — Credly
- **Google Cloud Engineer** — Credly
- **Complete Data Science, ML, DL & NLP Bootcamp** — Udemy
- **Advanced Learning Algorithms** — Coursera / DeepLearning.AI
- **Supervised ML: Regression & Classification** — Coursera / DeepLearning.AI
- **Python for Machine Learning** — Great Learning

Languages

- **Tamil** — Native
- **English** — Professional Working Proficiency