Anshuman Singh

+91 6201445053 | anshumansingh0802@gmail.com | singularityanshuman | Anshuman-cs50 | Portfolio

Summary

B.Tech CS undergraduate specializing in AI/ML engineering with hands-on experience building agentic LLM systems, RAG pipelines, on-device ML systems, and multi-model computer vision pipelines. Demonstrated expertise through participation in a Google-partnered AI/ML program, multiple hackathon awards, and independent development and delivery of RAG and computer vision systems.

Education

Haridwar University Sept. 2023 – June 2027
B.Tech in Computer Science and Engineering, 3rd Year *CGPA: 8.0 / 10.0*
- Coursework: ML, Deep Learning, Computer Vision, NLP, DSA, DBMS, OS, Networks

Experience

AI/ML Engineer (Virtual) | *EduSkills Foundation x Google Developers — AICTE Program, Cohort 15* Jan. 2026 – Mar. 2026
- Built and deployed an **on-device visual search and object detection app** on Android using **Google ML Kit** — fully local inference pipeline with <100ms latency, zero server round-trips, and no user data leaving the device.
- Applied **INT8 post-training quantization** via TFLite to compress model size by 69.5%, cutting cold-start latency below 40ms — validated on mid-range Android hardware (Snapdragon 6xx series).

Projects

DocAI — Personalized Healthcare Assistant | *Python, PostgreSQL, pgvector, RAG* Oct. 2025 – Feb. 2026
- Engineered a **stateful RAG pipeline** using PostgreSQL + **pgvector** for sub-200ms semantic retrieval across 10,000+ patient health records — retrieved context injected mid-conversation to ground `[SEARCH]` action responses, eliminating free-generation hallucinations by design.
- Designed a **ReAct-style agentic reasoning loop** on **MedGemma-4B** (self-hosted via Kaggle-backed Gradio endpoint) where the model dynamically selects between `[SEARCH]`, `[ASK]`, and `[ANSWER]` actions per turn — enabling structured multi-turn clinical reasoning without fine-tuning.
- Built a **prompt-driven clinical triage system** with emergency detection across 4 pattern categories, validated on **18 adversarial test cases** (including silent MI masking and atypical high-risk presentations) — achieving **17/18 pass rate** on a 4B parameter model purely through prompt architecture; NLP pipeline via **spaCy** extracts vitals, symptoms, and medication history into structured PostgreSQL records.

Civic Behaviour Monitoring System | *Python, YOLOv8, LSTM, OpenCV, PyTorch, PostgreSQL* Jan. 2026 – Apr. 2026
- Built a **two-stage detection pipeline**: YOLOv8n for person/object localization (mAP@50: 0.91, precision: 0.93) feeding into a **2-layer LSTM** for temporal activity classification — trained from scratch on a self-collected dataset of 600+ annotated video clips (80/20 split, post-augmentation).
- **LSTM activity classifier** achieved **95.7% accuracy** on temporal behaviour sequences, with classified activities feeding into a **rule-based civic scoring engine** that flags violations and cross-references a registered offender database in real time.

AI Timetable Generator | *Python, Flask, Google OR-Tools, CSP, PostgreSQL* Dec. 2025 – Jan. 2026
- Engineered a **CP-SAT scheduling engine** solving NP-hard timetabling for 50+ courses across multiple departments in under 10 seconds.
- Managed 3-dimensional constraint conflicts — room capacity, faculty availability, and student group limits — with **zero hard-constraint violations** in all generated schedules.
- Deployed a **full-stack web app** (Flask + Vercel) with CSV ingestion, real-time constraint validation UI, and exportable PDF/CSV timetable output — open deployment supporting multi-department scheduling configurations.

Technical Skills

Core Languages: Python (NumPy, Pandas, Scikit-learn), SQL (PostgreSQL + pgvector, SQLite3), C/C++
AI/ML Frameworks: PyTorch, TensorFlow/Keras, Hugging Face Transformers, OpenCV, LangChain
Generative & Deep Learning: CNNs, Transformers, Anomaly Detection, Retrieval-Augmented Generation (RAG)
LLM Engineering: Prompt Engineering, Model Quantization (TF-Lite / GGUF), RAG Architectures, Context-Injection Pipelines
MLOps & Infrastructure: Flask, Docker, Firebase, REST APIs, Git/GitHub, AWS S3/EC2 (basic)

Awards & Achievements

1st Place — Techwizard Hackathon | *Haridwar University* Oct. 2024
- Built **SaarthiHub** — an AI-powered personal assistant web app — in 24 hours, winning **1st place**. Stack: Flask backend, vanilla JS frontend, OpenAI API for conversational intelligence, deployed live during judging.

Top 33 / 150 Teams — National Level Hackathon | *Shivalik College of Engineering* Oct. 2025
- Built **VitalWatch** — a real-time health monitoring app ingesting smartwatch sensor data (BP, SpO2, sleep cycles) via Kotlin, with a Python/Firebase backend for pattern analysis using rule-based thresholds and an ML anomaly detection layer, triggering live push alerts on critical vital spikes — ranked **33rd out of 150 teams**.

Certifications & Coursework

- **Oracle OCI** — AI Associate Professional | **CS50AI** — Artificial Intelligence with Python, Harvard
- **Google Gen AI Learning Path** — In Progress | **DeepLearning.AI** ML & DL Specializations (Coursera, audited)