Rohan Pandey

Email: rohan1706pandey@gmail.com
GitHub: github.com/rohan-pandeyy
LinkedIn: linkedin.com/in/rohanpandeyy
Portfolio: rohan-pandeyy.github.io

EDUCATION

Bennett University — Greater Noida, Uttar Pradesh
B.Tech Computer Science, CGPA - 8.25/10.0 (6th Sem)
2023 – 2027

- **Relevant Coursework:** Artificial Intelligence & Machine Learning, Generative AI, Computer Vision, Probability, Statistics, Linear Algebra, Database Management Systems, Computer Networks, Operating Systems.

EXPERIENCE

HERE Technologies — Mumbai
Agentic AI Intern
Jul 2026 – Present

- Expanded and optimized an agentic AI workflow for business operations by developing a new automation module and implementing prompt caching, reducing LLM workflow costs by **39%**, from **$33 to $20 per run.**
- **Automated a multi-source map-content freshness evaluation pipeline** to benchmark HERE's map data against licensed third-party and open-source sources, including **OpenStreetMap**.

Australian Open Source Software Innovation and Education (AOSSIE) — Remote
Google Summer of Code (GSoC) 2026 | Contributor
May 2026 – Nov 2026

- Selected as a contributor for GSoC 2026 under AOSSIE *(top ~4.88% of 23K+ applications)* to expand PictoPy, an on-device AI-powered desktop image gallery.
- Implemented my proposed work focused on Visual Semantic Intelligence, smart Memories, video handling, and secure sharing.
- Implemented Quality Gate checks and Adaptive Density-Aware Face Clustering pipeline from recent research papers enhancing face detection and clustering by 4x.

KeenHeads Semiconductors — Greater Noida
Machine Learning Intern
Jun 2025 – Jul 2025

- Compressed the original LeNet architecture in C++ achieving 94.84% model compression via pruning and quantization.
- Achieved a 1.7x reduction in LeNet-5 inference latency with a negligible accuracy trade-off of less than 3%.
- Achieved a peak throughput of 33 FPS on ultra-low-power FPGA simulation, enabling real-time inference.

TECHNICAL SKILLS

- **Languages:** Python, C++, SQL, JavaScript, TypeScript, Rust
- **Machine Learning:** PyTorch, OpenCV, TensorFlow, YOLO, ONNX, Pandas, NumPy, Matplotlib
- **Frontend & Backend:** React, Next.js, FastAPI, SQLite, Docker
- **Tools & Platforms:** Git, GitHub, AWS, Linux

NOTABLE GITHUB PROJECTS

Drive Guardian

Python, PyTorch, YOLO, ONNX, OpenCV, Pillow, NumPy, CustomTkinter

- Developed a real-time ABS pipeline, merging object detection and lane segmentation under 1.1GB of VRAM.
- Optimized GPU management by decoupling PyTorch and ONNX Runtime execution environments.
- Achieved <30ms inference speeds and 20 FPS throughput across dual state-of-the-art neural networks.

Arogya for Maternal Health

Python, FastAPI, Next.js, PostgreSQL, NeonDB, Ollama, RAG, LLMs, Telegram API

- Built an AI-powered infant care tracking application with text and voice support, delivering response times under 5 seconds, capable of reducing annual childbirth-related mortality rate by up to 90% worldwide.
- Implemented Retrieval-Augmented Generation (RAG) pipeline using local Gemma4:12b, reducing hallucinated responses by 60% and indexing 10,000+ WHO guidelines, pediatric care documents, and infant health records.

OPEN SOURCE

- **AOSSIE Org:** Contributed 60,000+ lines of code across 40+ merged PRs, to PictoPy, optimizing ML pipelines, CI/CD workflows and improving both frontend and backend systems. View PRs
- **Activist Org:** Fixed CI/CD bugs, dependency errors, making review process 12% faster. View PRs
- **Open Healthcare Network:** Contributed by completing three frontend tasks, adding support to display a new "Severity" attribute, and aligning two misaligned/overflowing components. View PRs