

RANBIR KUMAR

+91 83760 86612 · rk26.ftw@gmail.com · Portfolio: [ranbirkumar26.github.io](https://github.com/ranbirkumar26)
github.com/Ranbirkumar26 · linkedin.com/in/ranbirkumar-ai

EDUCATION

B.Tech, Computer Science (AI & ML) — Vellore Institute of Technology, Chennai | CGPA: 8.28/10 | Aug 2023 – May 2027
Coursework: Data Structures & Algorithms, Machine Learning, Deep Learning, Computer Vision, Explainable AI, DBMS & SQL

EXPERIENCE

Software Testing & Product Development Intern | **IAMAI** (Internet & Mobile Association of India) | May 2026 – Jul 2026

- Automated manual pre-release QA with an **LLM-planned browser agent** (Playwright, Chromium) that reads DOM and accessibility snapshots, executes safe UI actions, and emits machine-readable defect reports, cutting click-through effort an estimated 70% per release, alongside k6 and Burp Suite hardening before launch.

Technical Integration Lead, AI Systems & Embedded Products | **IORA**, Noida | Nov 2025 – Apr 2026

- Drove architecture reviews across PCB, firmware (MQTT over TLS, GNSS, flash-buffered telemetry), and power subsystems for an nRF9151 LTE-M/NB-IoT wearable, aligning 3 teams on one interface contract.
- Architected the company's agentic RAG platform on FastAPI and Next.js, adding response caching and failover across 3 LLM providers against vendor outages.

Deep Learning & Computer Vision Intern | **Proeffico Solutions Pvt. Ltd.**, Noida | May 2025 – Oct 2025

- Shipped 4 production vision pipelines (speed, pothole, fire, workplace safety) on Jetson Xavier NX over live RTSP feeds at sub-40 ms latency, roughly 25 FPS per stream, replacing continuous manual review.

Rover Lead, Autonomous Systems & Computer Vision | **Technocrats Robotics**, VIT Chennai | Aug 2025 – Aug 2026

- Led a 60-member Rover Department, owning subsystem architecture, code review, and the build timeline for a perception stack spanning soil classification, arrow-marker detection, and an end-effector camera pipeline.

PROJECTS

IORA DocQA — Production Agentic RAG Document Platform | *FastAPI · Next.js · pgvector · DuckDB · Multi-LLM*

- Engineered ingestion (parsing, chunking, embedding into Supabase pgvector) behind a FastAPI retrieval service with a hard coverage gate that refuses unsupported answers, extended with DuckDB text-to-SQL, per-user memory, and a semantic cache resolving an estimated 25% of repeat queries without an LLM call.

GeoCaptionVLM — Vision-Language Foundation Model Fine-Tuning | *BLIP-2 · LoRA · PyTorch · HuggingFace*

- Fine-tuned BLIP-2 with LoRA for satellite captioning and retrieval on RSICD, raising CIDEr 0.36→0.68 and Recall@1 0.10→0.16; basis of a first-author paper under review at CVPR EarthVision.
- Extended it with a Siamese change-detection head on bi-temporal scene pairs, reaching 0.53 F1 and 0.36 IoU.

HiRAGe — Production RAG Candidate-Shortlisting Platform | *Next.js · Supabase · pgvector · Background Workers*

- Architected a multi-tenant RAG platform parsing and chunking resumes before retrieval, ranking the pool by a blended pgvector similarity score, serving real recruiters and applicants in production.
- Built a Postgres-backed job queue (FOR UPDATE SKIP LOCKED, partial unique index) so parsing and shortlisting run exactly once per job, with role checks enforced by both server and row-level security.

IAMAI Publishing Platform — Full-Stack CMS for a Live Public Website | *Next.js · Supabase · Postgres · GitHub Actions*

- Modelled eleven content collections on a normalized Postgres schema with CMS-managed page blocks, guarded by RBAC and released through GitHub Actions to Netlify under a smoke-test monitor.

TerraView — Indoor Venue Navigation & Routing Engine | *React · Vite · Graph Pathfinding* jio-map.netlify.app

- Represented a convention centre as a single routing graph, resolving cross-hall and cross-floor journeys across 4 floor views and indexing 393 stalls and 225 companies at Global Fintech Fest.

RESEARCH & PATENTS

Parameter-Efficient Fine-Tuning of Vision-Language Models for Satellite Image Captioning and Cross-Modal Retrieval | Under review, *CVPR EarthVision Workshop*. First author with two advisor co-authors; the GeoCaptionVLM work above.

Few-Shot Learning with Explainable AI | *Peer-reviewed journal, under review* — Prototypical Networks implemented from scratch ([RetinaProto](#)), 93% accuracy with Grad-CAM attribution.

Patrolling Bot — Autonomous Campus Surveillance Robot | *Patent filed* — YOLO detection and action recognition on Jetson Xavier NX with LiDAR/ultrasonic navigation. **Vision-Based Terrain Analysis** | *Accepted* — YOLOv8 fused with MiDaS depth for terrain classification.

SKILLS

Languages: Python, TypeScript, JavaScript, Java, SQL

AI & Machine Learning: PyTorch, TensorFlow, scikit-learn, HuggingFace Transformers, YOLOv8, ResNet, Prototypical Networks, LoRA/PEFT, vision-language models, Grad-CAM, SHAP, LIME, Jetson edge inference

Generative & Agentic AI: Agentic RAG, LangGraph, Tool Calling, Model Context Protocol, Vector Search (pgvector, FAISS), LangChain, Semantic Caching, Text-to-SQL

Backend & Data: FastAPI, Flask, Next.js, React, REST API Design, PostgreSQL, Supabase, DuckDB, RBAC & RLS, System Design

Cloud & Tooling: Docker, Git, GitHub Actions, Linux, Vercel, Netlify, Playwright, k6

ACHIEVEMENTS

Patent Filed · 2-time Finalist, DD Robocon & IRC · IRC Top 20 Worldwide 2025 · National Finalist, SIH 2025 & Insightify 6.0