Vinay Saunshi

Contact: 8884626824 • vinayrsaunshi@gmail.com • linkedin.com/in/vinayrsaunshi • github.com/Vinay-R-S

Professional Summary

AI/ML-focused Computer Science undergraduate with hands-on experience building production-oriented machine learning systems, computer-vision pipelines, and LLM-integrated automation. Strong foundation in model development, full-stack engineering and deployment (FastAPI/Flask, Docker), and blockchain-based audit trails. Proven track record of reducing manual effort and improving accuracy in real-world systems through automation and multi-modal AI.

Technical Skills

- **Languages:** Python, Java, JavaScript, TypeScript, C
- **ML & CV:** PyTorch, TensorFlow, YOLOv8, ResNet, MediaPipe, OpenCV, scikit-learn, NumPy, Pandas
- **LLMs & Orchestration:** LangChain, LangGraph, LLMs, Prompting, Groq MCP
- **Web & APIs:** React, Node.js, Flask, FastAPI, Next.js, REST, JWT
- **Infrastructure:** Docker, Git, GitHub Actions (CI/CD), Firebase, Supabase
- **Databases:** MongoDB, PostgreSQL, SQL
- **Tools & Dev:** Selenium, Selenium, HTML/CSS, Slack API, Solidity (Sepolia), n8n

---

Education

B.Tech (Hons.) - Computer Science (AI & ML specialization) RV University, Bangalore, Karnataka | Sep 2023 – Jun 2027 (Expected)

- Current CGPA: 9.33 / 10

Professional Experience

AI/ML Intern - Zodopt Technologies (Zoho Premium Partner) - Remote / Bangalore Jun 2025 – Jul 2025

- Designed AI-agent workflows and automated sales-research tasks using **n8n** and Selenium, reducing LinkedIn lead profiling time from **\~15 minutes to \~2 minutes per lead**.
- Structured and cleaned scraped lead datasets for downstream LLM-based enrichment and workflow integration, enabling contextual responses and automated follow-ups.
- Integrated automation with company processes to accelerate lead qualification and reduce manual data-entry errors.

Selected Projects (GitHub links included)

ReClaim-AI - Real-time CCTV Analysis & Recovery Platform

Tech: YOLOv8, LangGraph, LLMs, Flask, React, Solidity, Firebase

- Built a real-time YOLOv8 computer-vision pipeline for CCTV analysis achieving **\~85%+ accuracy** across 80+ object classes (COCO-based), cutting manual review time by **\~75%**.
- Designed a multi-modal matching system combining semantic NLP and visual-similarity to improve recovery-match accuracy by **\~60%**, reducing average recovery time to **2–3 days**.
- Implemented Ethereum smart contracts on Sepolia to create a tamper-proof handover flow and an auditable 3-tier microservice architecture.
- GitHub: https://github.com/Vinay-R-S/ReClaim-AI

AI-Powered Complaint Triage System

Tech: Groq MCP, LLMs, FastAPI, Supabase, Slack API, Docker

- Built an LLM-powered complaint classification and routing system automating intake across **7+ industries**, decreasing manual triage time by **70–80%** at an **85%** confidence threshold.
- Implemented a severity & risk scoring pipeline (1–5) with SLA multipliers to prioritize critical cases; improved high-severity response times by **75–90%**.
- Architected a Dockerized FastAPI-based service with Supabase persistence and Slack-based routing/ alerts.
- GitHub (active development): https://github.com/Vinay-R-S/neutrinos-mcp-hackathon

XpressiveAI - Sign Language & Emotion Recognition

Tech: ResNet-50, AGFN, MediaPipe, TensorFlow, PyTorch

- Developed parallel pipelines for emotion recognition and sign-language classification using facial images and hand landmarks.
- Enhanced ResNet-50 with Adaptive Global Feature Normalization (AGFN), trained on a 52K-image dataset and improved emotion recognition accuracy from **80.29% → 86.03%** with class-wise F1 up to **0.965**.
- Implemented a MediaPipe-based hand-landmark extractor and an MLP classifier achieving real-time webcam inference latency of **\~45 ms**.
- GitHub: https://github.com/Vinay-R-S/xpressiveai

Achievements & Recognition

- **2nd Place** - ReClaim-AI, National-Level GDG Tech Sprint Hackathon (RV University)
- **2nd Place** - Capture-The-Flag (CTF) Cybersecurity, Tarang Fest (RV University)
- **Merit Scholarship** - Awarded for Top 5% rank in 1st & 2nd year at RV University

Technical Interview Simulation

Interviewer: Hi Vinay, let's start with your technical round. Can you briefly introduce yourself from a technical perspective?

You: Hi, I'm Vinay, a Computer Science undergraduate specializing in AI and ML. Most of my work revolves around building end-to-end ML systems - from data pipelines and model training to deployment using FastAPI, Docker, and cloud services. I enjoy working on computer vision and LLM-integrated automation problems.

Interviewer: I see you worked on ReClaim-AI. Can you explain the system architecture at a high level?

You: Sure. ReClaim-AI follows a three-tier architecture. The first layer handles real-time video ingestion and object detection using YOLOv8. The second layer performs semantic matching using NLP embeddings combined with visual similarity scores. The final layer manages ownership transfer and auditability using Ethereum smart contracts deployed on Sepolia.

Interviewer: Why did you choose YOLOv8 over other object detection models?

You: YOLOv8 offered a good balance between inference speed and accuracy. Since the system processes CCTV streams in near real time, latency was critical. YOLOv8 also simplified deployment because of its modular architecture and strong performance on COCO-style datasets.

Interviewer: How did you evaluate and improve model accuracy?

You: I focused on dataset curation first, removing noisy samples and balancing classes. I used standard metrics like mAP and class-wise precision-recall. For improvements, I applied targeted augmentation and threshold tuning based on real-world false positives rather than just validation loss.

Interviewer: Can you explain one challenge you faced during deployment?

You: Handling real-time streams with limited compute was challenging. I optimized inference by batching frames, using lower-resolution inputs when confidence was high, and containerizing services with Docker to isolate workloads and manage resource usage.

Interviewer: How comfortable are you with LLM integration?

You: I've integrated LLMs using LangChain and LangGraph for routing, classification, and semantic matching tasks. I'm comfortable with prompt design, tool calling, and building deterministic pipelines around probabilistic outputs.

Interviewer: That's good for now. Let's move to the next round.

Aptitude Interview Simulation

Interviewer: This round focuses on problem-solving and logical thinking. Ready?

You: Yes, ready.

Interviewer: If a system processes 5 tasks in 10 seconds, how long will it take to process 20 tasks assuming linear scalability?

You: If performance scales linearly, 20 tasks are 4 times the workload, so it would take 40 seconds.

Interviewer: You have two APIs. One has 99 percent uptime but higher latency, the other has lower latency but 95 percent uptime. Which would you choose for a real-time ML inference system?

You: I would choose based on SLA requirements. For real-time inference, low latency is usually more critical, but I'd mitigate uptime risk using retries, fallbacks, or a hybrid approach where the low-latency API is primary and the high-uptime API is a backup.

Interviewer: How would you estimate the number of CCTV cameras in a city?

You: I'd use a top-down estimation. Start with population or number of commercial and residential buildings, estimate average cameras per building type, then refine using known data like traffic junctions, public infrastructure, and private installations.

Interviewer: A model gives 90 percent accuracy but performs poorly in production. What could be the reason?

You: Possible reasons include data leakage, train-test distribution mismatch, poor handling of edge cases, or an unrepresentative validation dataset. Monitoring real-world inputs and retraining with production data usually helps.

Interviewer: Good. Let's proceed to the HR round.

HR Interview Simulation

HR: Hi Vinay, how was the interview so far?

You: It's been good. I enjoyed discussing my projects and problem-solving approaches.

HR: Tell me about yourself beyond your resume.

You: Beyond academics, I enjoy building systems that solve practical problems. Hackathons and projects motivate me because they simulate real-world constraints. I also like breaking down complex problems into simple, deployable solutions.

HR: Why should we hire you?

You: I bring a strong mix of hands-on engineering and problem-solving mindset. I don't just build models; I focus on how they behave in production. I'm adaptable, quick to learn, and comfortable taking ownership of end-to-end systems.

HR: Tell me about a failure and what you learned.

You: In one project, I over-focused on model accuracy and ignored deployment constraints. The system struggled in real-time usage. That taught me to think about scalability, latency, and monitoring early in the design phase.

HR: Where do you see yourself in 3 to 5 years?

You: I see myself as an AI engineer working on production-scale systems, contributing to architecture decisions, and mentoring juniors while continuously improving my technical depth.

HR: Do you have any questions for us?

You: Yes. How does your team move models from experimentation to production, and how much ownership do engineers have over deployed systems?

HR: That concludes the interview. Thank you for your time.

You: Thank you for the opportunity.