

Saharsh Sandeep Barve

✉ saharshbarve@gmail.com  saharshbarve  saharshbarve  saharshbarve.github.io

Work Experience

Software Engineer — ML Platform

Dragonfruit AI

Remote, USA / Menlo Park, CA

Jun 2025 – Present

- Built a config-defined **zero-shot CLIP classifier** framework across 5 services with prompt- and crop-embedding caches, letting customer teams stand up new attribute detectors (posture, phone use) as a JSON config rather than a model release, turning pilots and PoCs from weeks into hours while cutting inference 12.6ms → 8.5ms.
- Re-architected safety alerting from **cloud cron polling** to **event-driven edge push** over the base station's persistent WebSocket, with an at-least-once queue and cloud-side deduplication, cutting alert latency from 25s–5min to ~1–2s ($>10\times$).
- Rebuilt on-device **multi-object tracking** for edge inference with vectorized association, dwell-based static classification and appearance re-ID, cutting per-frame cost 2.91ms → 0.77ms and scaling a single base station to 20 concurrent camera streams with zero dropped frames, while removing a recurring source of customer false alarms.
- Owned the **model development lifecycle** for 5 production detection models across annotation, training, evaluation and release, standardizing it into a gated pipeline that made shipping a model self-serve for the team and caught accuracy regressions on real deployment footage before they reached customers.
- Built a **truck turnaround monitoring** pipeline pairing YOLO detection with license-plate recognition and cloud thumbnail storage, achieving ~300ms on-prem inference per event and replacing manual gate logging for logistics customers.

Machine Learning Scientist

Onward Assist

Bengaluru, India

Jul 2022 – Jul 2023

- Led development of the Nottingham scoring models for breast-cancer histopathology grading, boosting diagnosis accuracy by 30% over the previous baseline.
- Deployed ML models to the web platform on AWS and Kubernetes with MLflow tracking and Apache Parquet data workflows, enabling reproducible and scalable serving.
- Led the team to a top-8% finish in the HuBMAP + HPA Kaggle competition on large-scale biomedical image analysis.

Education

University of Illinois Urbana-Champaign

M.S. Computer Science, GPA: 3.96/4.0

Urbana-Champaign, Illinois

Aug 2023 – May 2025

Manipal Institute of Technology

B.Tech Computer Science and Engineering, GPA: 9.25/10

Manipal, India

Jul 2018 – May 2022

Technologies

Languages: Python, C/C++, Go, TypeScript, SQL

ML & Vision: PyTorch, CUDA, YOLO, CLIP/MobileCLIP, OpenCV, multi-object tracking, CoreML & on-device inference

LLMs & Agents: RAG, prompt optimization, agentic workflows & tool use, MCP, VLM evaluation, evals

Systems & Infra: Docker, Kubernetes, AWS, gRPC/protobuf, WebSockets, Celery/RabbitMQ, real-time & edge deployment, CI/CD

Data & MLOps: Weights & Biases, MLflow, Qdrant, FAISS, Elasticsearch, PostgreSQL

Publications

- *Social stereotypes in AI text-to-image generation* — AI and Ethics (Springer), 2026, 6(3). [\[Link\]](#)
- *Reef-Insight: A Framework for Reef Habitat Mapping with Clustering Methods Using Remote Sensing* — Information, 2023, 14, 373. [\[Link\]](#)
- *Switched Auxiliary Loss for Robust Training of Transformer Models for Histopathological Image Segmentation* — arXiv, 2023. [\[Link\]](#)

Leadership & Additional

- **Teaching Assistant, UIUC** — CS444 Deep Learning for CV, CS441 Applied ML, CS498 Cloud Computing.
- **Open Source** — [BespokeLabsAI/Curator](#): synthetic data curation for post-training.
- **Awards** — J. N. Tata Endowment Scholarship; Head of Finance, IAESTE India LC Manipal.