

# Kshitij Sharma

## Lead AI/ML Engineer & Data Scientist

Gurugram, India · ksharma3@zohomail.in

LinkedIn · GitHub · kshitijworks.com

14+ years of experience across software engineering, data science, and AI/ML. Hands-on work in computer vision, predictive analytics, recommendations, probabilistic modeling, RAG, and AI agents, with ownership from research and experimentation through production implementation and system architecture.

## Experience

---

### Independent AI products

Jul 2025 – Present

#### Lead AI/ML Engineer & Architect

- Architected and built a multi-camera retail vision platform for live store environments, integrating customer analytics with Petpooja, Zoho, and delivery data through Django/FastAPI microservices and REST APIs.
  - Engineered real-time detection with YOLO12s, Kalman/Hungarian tracking, and cross-camera re-identification and facial recognition with OSNet/ArcFace and FAISS indexing, sustaining 65–85 ms median per-camera latency on 2.7K feeds.
  - Developed homography-based world-space fusion to maintain identity continuity across overlapping camera views and derive dwell time, table occupancy, customer retention, movement flows, and zone activity.
  - Built a multi-tenant restaurant AI platform for web and Telegram using LLMs, LangChain, RAG, and FAISS for menu discovery, cart management, ordering, and order tracking. Integrated persistent context, semantic caching, POS systems, human-agent takeover, and natural-language sales analytics.
  - Deployed Dockerized edge/cloud services using FastAPI/Django, PostgreSQL, GPU infrastructure, CI/CD, Cloudflare security, and production monitoring.
- 

### Mercedes-Benz R&D North America

Apr 2021 – Jun 2025

#### Lead Data Scientist · Consultant, remote

- Developed context-aware vehicle intelligence using probabilistic suggestions and self-learning routines for comfort settings, heating, onboard applications, and location-based suggestions.
- Applied Bayesian inference and variational modeling to user behavior and in-car action prediction, supporting adaptive suggestions and personalized driver experiences.
- Led causal inference and machine learning initiatives to optimize vehicle climate control, EV charging strategies, and other operational prediction problems.
- Owned the ML lifecycle from Python research to embedded C++ implementation, authoring production software deployed onto 2026 Mercedes-Benz production vehicle models.

---

## Wynk Limited

Apr 2020 – Mar 2021

### Lead Data Scientist

- Built hyper-personalized playlist models serving 50M+ users through music embedding spaces, nonparametric clustering, and recommendation techniques.
- Designed and deployed a learning-to-rank model for personalized search, improving NDCG with deep NLP architectures integrated with Elasticsearch.
- Engineered Python, TensorFlow, and Spark pipelines to learn song embeddings from audio features (MFCC) and user consumption patterns, powering similarity and ranking engines.

---

## Mercedes-Benz R&D North America

Aug 2016 – May 2019

### Machine Learning Engineer

- Developed vehicle prediction models using Bayesian inference and Bayesian nonparametric methods, including mixture-model approaches to location clustering and automation.
- Designed, implemented, and tested ML models in C++ for production deployment within vehicle head units.
- Built causal graphs and counterfactual models to distinguish low-value head-unit applications from those underused because of availability or placement, informing application discontinuation decisions.

---

## SAS Institute Inc.

Jun 2015 – Apr 2016

### Software Engineer, Research Intern

- Developed backend features for SAS Graph Template Language in a C++ graph framework, including new SAS procedures and visualization functionality.
- Migrated graphing procedures and functionality from Java to the newer C++ platform.

---

## Publicis Sapient

Nov 2011 – Jun 2014

### Software Engineer

- Developed backend services for a multi-channel Java commerce platform, with primary ownership of order capture and associated commerce functionality.
- Integrated payment processing, shopping baskets, and web services for end-to-end transactions across channels.
- Designed an in-memory cache for high-speed inventory transactions to reduce repeated database and backend lookups.

---

## Independent projects

### Offline facial recognition attendance

Built an Android tablet kiosk using React, TypeScript, Capacitor, SQLite, CameraX, MediaPipe, and LiteRT. Supports guided enrollment, face-based check-in/check-out, scheduling, reports, role-based access, encrypted backup, and English/Hindi interfaces. Recognition runs on-device; optional sync and exports support workplace administration.

## IoT laundromat automation

Built a functional Raspberry Pi Pico and Django prototype connecting machine availability, web payments, relay control, current-based cycle detection, and peristaltic-pump detergent dispensing.

---

## Technical skills

### AI / ML

Computer Vision, Deep Learning, Object Detection, Tracking, Re-ID, RAG, AI Agents, Recommendation Systems, Bayesian Modeling, Time Series, Causal Inference, Edge AI

### Languages

Python, C++, SQL, R, C, Java, JavaScript, TypeScript

### AI / Data Technology

TensorFlow, OpenCV, Scikit-learn, LangChain, FAISS, TensorRT, ONNX Runtime, LiteRT, MediaPipe, Spark

### Backend

Django, Django REST Framework, FastAPI, PostgreSQL, MongoDB, Redis, Celery

### Infrastructure

Docker, Nginx, Gunicorn, AWS, S3 / R2, Cloudflare

### Frontend / App

React, TypeScript, Capacitor, Electron, Android

---

## Education

### North Carolina State University

M.S. in Computer Science · 2016 · GPA: 3.96 / 4.00

### National Institute of Technology, Silchar

B.Tech. in Computer Science & Engineering · 2011

---

## Publication

### A Scalable Approach for Outlier Detection in Edge Streams Using Sketch-based Approximations

S. Ranshous, S. Harenberg, K. Sharma, N. F. Samatova. SIAM International Conference on Data Mining (SDM), 2016.