

SHUBHAM KULKARNI

AI Engineer | Computer Vision • Generative AI & RAG • Edge AI • MLOps

Pune, Maharashtra, India | +91 8308003684 | 21shubhamkulkarni@gmail.com
[linkedin.com/in/shubhkulk21](https://www.linkedin.com/in/shubhkulk21) | github.com/kulkarnishub377 | kulkarnishub377.github.io

PROFESSIONAL SUMMARY

AI Engineer with 1+ year of production experience building Computer Vision, Generative AI, and Edge AI systems for national Intelligent Transportation Systems infrastructure — processing 50,000+ real-time events daily at 95% accuracy and sub-40ms latency on NVIDIA Jetson. Built a FAISS-based retrieval engine delivering 30x faster search across 150,000+ daily forensic queries. Strong in RAG, multi-agent orchestration, and hybrid search, translated into resilient production microservices. National Champion (AIR 1, Smart India Hackathon 2023).

CORE COMPETENCIES

AI / Computer Vision: Deep Learning, LLMs, YOLOv5/v8, RT-DETR, ByteTrack, OSNet-AIN, ResNet, PaddleOCR (PP-OCRv4), NLP, PyTorch, TensorFlow, OpenCV, Hugging Face

Generative AI & Agents: Retrieval-Augmented Generation (RAG), Hybrid Search (FAISS + BM25 + RRF), LangChain, Prompt Engineering, Multi-Agent Orchestration, Model Context Protocol (MCP), Knowledge Graphs, Vector Databases (FAISS, ChromaDB)

Edge AI & Hardware: NVIDIA Jetson (Orin/NX/Xavier/Nano), TensorRT (FP16/INT8), CUDA, ONNX Runtime, GStreamer, Hardware-Accelerated Video Decoding, Raspberry Pi, Rockchip, Arduino

Backend & MLOps: Python, FastAPI, Django, Flask, AsyncIO, REST APIs, Microservices, Docker, Kubernetes, CI/CD, Git, Multi-threading, WebSockets (SSE)

Databases & Cloud: PostgreSQL, Redis, MongoDB, MySQL, AWS, Microsoft Azure, Google Cloud Platform (Vertex AI), SQL

PROFESSIONAL EXPERIENCE

Software Engineer – AI/ML — Arya Omnitalk Wireless Solutions Pvt. Ltd., Pune

Aug 2025 – Present

- Architected edge-native ITS pipelines:** led design of concurrent multi-model inference architecture for enterprise Intelligent Transportation Systems (VIDES, ATMS), scaling to 50,000+ real-time events daily at 95% enforcement accuracy.
- Eliminated CPU bottlenecks:** engineered zero-latency video ingestion via GStreamer and hardware-accelerated decoding (nvv4l2decoder), achieving a sustained sub-40ms processing lock across 6 concurrent HD RTSP streams on NVIDIA Jetson.
- Accelerated inference throughput:** compiled ONNX models into TensorRT FP16 engines to maximize Ampere GPU utilization; orchestrated YOLOv8, RT-DETR, and ByteTrack for multi-lane vehicle taxonomy and spatial trajectory hazard detection.
- Built resilient real-time data relays:** architected asynchronous, decoupled inference-to-UI pipelines using PostgreSQL thread pools and sub-150ms WebSocket (SSE) broadcasts for instant operator alerting.
- Automated highway toll auditing:** engineered advanced OCR pipelines using PaddleOCR consensus logic, eliminating manual reporting workflows and achieving 100% data integrity.
- Delivered 99.9% edge-node uptime:** implemented CI/CD pipelines, automated heartbeat monitors, and exponential-backoff retry protocols for seamless offline network recovery.
- Owned vehicle re-identification platform:** designed a stateless, database-free microservice for real-time forensic video analysis, generating 512-dimensional OSNet-AIN embeddings and achieving sub-100ms FAISS IVF vector search across 150,000+ daily surveillance queries (30x retrieval speedup).

HIGH-IMPACT ENGINEERING PROJECTS

CivicMind — AI-Powered Smart City Decision Intelligence Platform | Gemini 2.5, Vertex AI, FastAPI, Next.js 14, Docker | 2026 |

github.com/kulkarnishub377/CivicMind

- Architected an enterprise-grade multi-agent orchestration platform (4 coordinated AI agents) with a digital-twin simulator and Explainable AI (XAI) layer for transparent, risk-based policy recommendations.

OmniSight — Autonomous Multi-Modal Incident Intelligence Fabric | FastAPI, Agentic AI, RAG, Vector DB, Docker | 2026 |

github.com/kulkarnishub377/OmniSight

- Built an agentic workflow engine that ingests and correlates high-velocity IoT, camera, and audio signals in real time to validate anomalies and auto-generate safety guardrails.

Document AI + RAG Pipeline (DocuAI Studio v3.1) | *Python, FastAPI, FAISS, Mistral, Ollama, LangChain, AsyncIO* | *Mar 2026 – Present* | github.com/kulkarnishub377/Document-AI---RAG-Pipeline

- Architected a secure, locally-hosted RAG platform (40+ API endpoints) processing PDFs, images, and web URLs with sub-2-second latency and zero external data exposure.
- Engineered a hybrid retrieval engine fusing FAISS dense vector search with BM25 sparse keyword search via Reciprocal Rank Fusion (RRF), significantly reducing LLM hallucinations; added interactive knowledge-graph entity extraction and complexity-based routing between local (Mistral 7B) and cloud LLMs.

Vehicle Forensic Matching System (RAC) | *PyTorch, YOLOv5, FAISS, FastAPI, OSNet-AIN, Docker* | *Dec 2025 – Mar 2026* | github.com/kulkarnishub377/Vechile_forensic_match_RAC

- Designed a stateless, database-free vehicle re-identification microservice for real-time forensic video analysis across non-overlapping camera networks.
- Generated 512-dimensional deep feature embeddings (OSNet-AIN) with YOLO-based localization, sustaining high-fidelity matches under extreme occlusion and variable lighting; optimized FAISS IVF indexing to sub-100ms search across 150,000+ records.

JalTantra — Smart Irrigation & AgTech Platform (SPPU Startup Olympiad, Rank 2) | *Python, IoT, Computer Vision, ML, GenAI* | *Dec 2024 – May 2025* | kulkarnishub377.github.io/JalTantra

- Spearheaded an AI-powered precision-agriculture platform combining IoT irrigation logic, predictive yield modeling, and real-time CV-based crop-disease detection, improving water efficiency by 30%.
- Deployed a multilingual (Hindi, Marathi, English) AI advisory chatbot for instant farmer support; pitched live at SPPU Startup Olympiad 2025, securing University Rank 2.

Alumni Management Portal — Enterprise Full-Stack Application (164+ commits) | *Django DRF, PostgreSQL, Flask, WebSockets, React/ES6* | github.com/kulkarnishub377/Alumni_Management_portal

- Architected a hybrid-monolith platform (19 DB models, 50+ REST APIs, 4 role-based dashboards) replacing legacy university systems for mentorship pairing and event management.
- Slashed database query latency from 3,200ms to 12ms via composite indexing and query profiling; engineered a custom WebSocket subscriber queue resolving JWT refresh race conditions.

EDUCATION

Bachelor of Engineering (B.E.), Electronics & Telecommunication — First Class Dec 2021 – Mar 2025
Savitribai Phule Pune University (SPPU) — Dr. VVP College of Engineering

- Engineered the officially adopted Alumni Management Portal; developed AI E-Waste Management System (Winner, AIR 1, Smart India Hackathon 2023).

ACHIEVEMENTS & CERTIFICATIONS

- **All India Rank 1 (AIR 1) — Smart India Hackathon 2023:** National Championship (Software Edition), Team Lead — engineered an intelligent e-waste routing and classification system.
- **Runner-Up (Global) — TIAA Global AI Hackathon:** built an AI investment-intelligence platform with risk-based scoring and portfolio recommendations.
- **University Rank 2 — SPPU Startup Olympiad 2025:** recognized for JalTantra, demonstrating strong commercial viability in AgTech.
- **Deploy Multi-Agent Architectures** — Google (2026) | **Career Essentials in Generative AI** — Microsoft & LinkedIn Learning | **Engineer AI Agents (ADK)** — Google Cloud (2025) | **Introduction to Machine Learning Concepts** — Microsoft (2026) | **Deep Learning Onramp & Statistics Onramp** — MathWorks