Shubham Kulkarni

AI Engineer | Computer Vision | Generative AI | MLOps | Edge AI

Pune, Maharashtra | +91 8308003684 | kulkarnishub377@gmail.com

linkedin.com/in/shubhkulk21 | github.com/kulkarnishub377 | kulkarnishub377.github.io/sk

PROFESSIONAL SUMMARY

Results-driven AI Engineer with proven production experience designing, building, and deploying scalable Computer Vision and Generative AI systems that process 50,000–150,000+ events/day. Specialized in real-time low-latency inference, multi-modal retrieval, Retrieval-Augmented Generation (RAG) pipelines, MLOps, and edge AI deployment on NVIDIA Jetson and Raspberry Pi. Delivered measurable impact across multiple live systems — 30% latency reduction, 95% model accuracy, and 30 times faster vector retrieval. Experienced in end-to-end system design from prototype to production, with strong backend integration, model optimization (TensorRT, ONNX), and cloud deployment skills. National hackathon champion (AIR 1, Smart India Hackathon 2023) with a track record of building AI solutions that create real-world value.

TECHNICAL SKILLS

Languages: Python, SQL, JavaScript

AI / ML: Computer Vision, Deep Learning, NLP, Large Language Models (LLMs), Hugging Face Transformers

Frameworks: YOLOv5/v8, OpenCV, PyTorch, TensorFlow, PaddleOCR, ByteTrack, OSNet, ResNet

Generative AI: Retrieval-Augmented Generation (RAG), Prompt Engineering, LangChain, FAISS, Vector Databases, Embeddings

MLOps & Backend: MLOps, FastAPI, Flask, Django, REST APIs, Async Systems, Microservices, Docker, CI/CD, Git

Model Deployment: TensorRT, ONNX, Model Optimization, Inference Pipeline Tuning, Batch Processing, Multi-threading

Edge AI & Hardware: NVIDIA Jetson Nano, NVIDIA Jetson Xavier, Raspberry Pi, Rockchip

Cloud Platforms: AWS, Microsoft Azure, Google Cloud Platform (GCP)

Databases & Caching: PostgreSQL, Redis, MSSQL

PROFESSIONAL EXPERIENCE

Software Engineer – AI/ML

Aug 2025 – Present

Arya Omnitalk Wireless Solutions Pvt. Ltd. | Pune, Maharashtra | Full-time, On-site

- Built and scaled two production-grade AI and Computer Vision pipelines (VIDES and ATMS), achieving 95% model accuracy and processing 50,000+ events/day in live environments.
- Developed real-time object detection and multi-object tracking models using YOLOv8, OpenCV, and ByteTrack for vehicle detection, classification, and traffic violation analysis with under 100ms inference latency.
- Engineered full end-to-end systems integrating NVIDIA Jetson Nano/Xavier edge devices with backend microservices via async REST APIs for real-time field deployment.
- Optimized inference pipelines using multi-threading and batch processing, reducing end-to-end system latency by 30% in production — directly improving real-time responsiveness for field operators.
- Built OCR-based structured data extraction pipelines using PaddleOCR, reducing manual data entry effort by eliminating manual reporting workflows for the operations team.
- Collaborated with hardware engineers, backend developers, and operations stakeholders across 3 teams to integrate AI capabilities into existing production infrastructure.
- Developed and maintained FastAPI and Flask backend services including real-time dashboards for system monitoring and operator interaction.
- Deployed and maintained AI systems across cloud (AWS/Azure) and edge environments, ensuring 99% uptime and continuous delivery via CI/CD pipelines.

Artificial Intelligence Intern

Jun 2023 – Aug 2023

IBM India Pvt. Ltd. | Remote | Internship

- Developed an NLP-based mental health classification system achieving 90% accuracy using transformer-based models and classical ML techniques.

- Executed full ML pipeline: data collection, preprocessing, feature engineering, model training, hyperparameter tuning, and evaluation — improving baseline model accuracy by 15%.
- Collaborated with IBM mentors and delivered results aligned with IBM enterprise AI standards, presenting findings to senior researchers.

PROJECTS

Vehicle Forensic Matching System

PyTorch | OpenCV | FAISS | FastAPI | Redis | PaddleOCR | ResNet | OSNet

- Problem: Manual vehicle identification from surveillance footage was slow and error-prone across large databases of 150,000+ entries.
- Built a multi-modal vehicle retrieval system combining deep learning embeddings (ResNet + OSNet), OCR-extracted license plate data, and vector search — enabling automated forensic identification at scale.
- Optimized retrieval speed by 30 times using FAISS IVF indexing, reducing average query response time to under 1 second and cutting investigator lookup time significantly.
- Scaled to handle 150,000+ vehicle queries/day with 99% uptime via async processing, Redis caching, and load-balanced FastAPI services.

Document AI + Retrieval-Augmented Generation (RAG) Pipeline

PyTorch | LangChain | FAISS | PaddleOCR | LLM | RAG

- Problem: Teams spent hours manually reviewing large volumes of PDFs and scanned documents to extract key information.
- Built an end-to-end Document AI pipeline using OCR and layout analysis to extract structured data from PDFs and scanned images, eliminating manual document review.
- Implemented a full RAG architecture with semantic chunking, vector indexing, and LLM-based reasoning — enabling natural language Q&A over documents with sub-2 second response time.
- Designed reusable prompt pipelines for summarization, key-field extraction, and table parsing — reducing document processing time by enabling instant querying over large document corpora.

ACHIEVEMENTS & AWARDS

- **AIR 1 — Smart India Hackathon 2023 (Team Lead)** — Developed an AI-based e-waste management system for automated waste classification. Selected as best solution from thousands of teams across India.
- **University Rank 2 — SPPU Startup Olympiad 2025** — Built an AI smart irrigation system using predictive modelling to minimize water usage and maximize crop yield.
- **Runner-Up — AI Hackathon, TIAA Global Pvt. Ltd.** — Developed an AI-powered investment intelligence platform with risk-based scoring and data-driven financial recommendations.

EDUCATION

Bachelor of Engineering (B.E.) — Electronics & Telecommunication 2021 – 2025

Dr. Vithalrao Vikhe Patil College of Engineering | Ahilyanagar (Ahmednagar), Maharashtra | Savitribai Phule Pune University (SPPU)

- Final Year Project: Alumni Management Portal — Designed and developed a full-stack alumni management system that was officially adopted and deployed by the college. The portal is actively used by the institution for alumni tracking, networking, and engagement.
- College Core Committee: Served as Secretary, Refreshment & Food Department — Managed logistics, coordination, and execution of food and refreshment operations for all major college events and functions.

CERTIFICATIONS

- **Deep Learning Onramp** — MathWorks (2024)
- **Python Programming** — GUVI (2023)
- **Geodata Processing using Python** — ISRO (2023)
- **Global Navigation Satellite System (GNSS)** — ISRO (2023)