SURAJ MURARI

(Machine Learning Engineer | Deep Learning & Hardware-Accelerated Solutions)

Noida, INDIA • surajmurari02@gmail.com • +91-9410727311 • LinkedIn • GitHub

SUMMARY

- Experienced Machine Learning Engineer with a proven track record of building scalable, production-ready AI solutions using Deep Learning, Computer Vision, LLMs, and GenAI in containerised (Docker-based) environments.
- Skilled in Python, PyTorch, OpenCV, CUDA, DeepStream, TensorRT, and OpenVINO, with hands-on expertise in YOLO, Siamese Networks, Triplet Loss, MobileNet, and embedding-based retrieval systems.
- Developed optimised neural networks and high-throughput GPU-accelerated pipelines for both edge and cloud deployments.
- Proficient in LLM integration, RAG pipelines, and workflow automation tools such as n8n and LangChain.
- Strong foundation in CI/CD, cloud deployment, inference optimisation (quantisation, sparsity), and modular AI system architecture.
- Passionate about shipping real-world AI/ML solutions with modular, maintainable design patterns.

TECHNICAL SKILLS

- **Languages:** Python, MySQL
- **Frameworks & Libraries:** PyTorch, TensorFlow, Keras, Scikit-learn, XGBoost, OpenCV, PaddleOCR, Ultralytics, MobileNet, AlexNet, DINOv2, Hugging Face, LangChain, Streamlit, Seaborn, OpenAI, Ollama
- **Tools & Technologies:** Docker, Git, VS Code, MongoDB, Postman, CVAT, Stable Diffusion, FASIS, Cursor, n8n, Flask, FastAPI, CUDA, DeepStream, TensorRT, OpenVINO
- **Concepts:** CNNs, Triplet Loss, Siamese Networks, Transfer Learning, LLMs, Retrieval-Augmented Generation (RAG), Agentic AI, Federated Learning, Vector Databases, RESTful APIs, Model Deployment, Quantisation, Sparsity, CI/CD, Reinforcement Learning

WORK EXPERIENCE

Samajh.ai, Noida — **Aug 2024 - Present**

Machine Learning Engineer Intern

- Built scalable computer vision pipelines using PyTorch, Siamese Networks, and Embedding models with batch inference and real-time feedback.
- Developed lightweight embedding-based object matching models using auxiliary inputs like camera ID and time delay.
- Integrated YOLO with OpenCV, DeepStream, TensorRT, and CUDA to optimise real-time object detection inference, achieving a 360% increase in speed (from 500 to 2300 FPS).
- Designed a queue-based video inference system with batch processing, status tracking, and robust logging.
- Deployed AI modules such as face matching, ANPR, ATCC, VIDS, and MLFF across locations like Zojila, DME, Pune-Satara, and for clients including PMO, LG, CMS, and Mahindra University.
- Reduced latency and improved model efficiency through quantisation, sparsity, and end-to-end training optimisations.
- Worked on LLM integration, RAG pipelines, batch workflows, and Dockerized edge deployments for cost-effective AI delivery.
- Worked on Stable Diffusion deployment via Flask API on GPU-based Ubuntu server.

PROJECTS

RetinoDeepAI — **Feb 2024 – May 2024**

- Built an AI-based diabetic retinopathy detection system using MobileNetV3 and CNNs.
- Applied PyTorch, Keras, and XGBoost for classification and preprocessing (image normalisation, augmentation).
- Achieved improved accuracy through model fine-tuning and cross-validation on medical image datasets.

Epiderma Lens — **Aug 2023 – Dec 2023**

- Developed a skin disease classification system using CNNs and PyTorch.
- Enhanced model performance via data augmentation, transfer learning, and subtle pattern recognition.

EDUCATION

- **Master of Computer Applications** — **Aug 2023 - Jul 2025**
  Bennett University, Greater Noida
- **Bachelor of Computer Applications** — **Jul 2020 - Jul 2023**
  D.S.B. Campus, Nainital

CERTIFICATIONS

- **Deep Learning** Certificate — **Sep 2023 - Oct 2023**
  IBM, Coursera
- **ML-Powered Image & Video Processing** Certificate — **Mar 2024 - Apr 2024**
  Duke University, Coursera
- **Introduction to Agile Development and Scrum** Certificate — **Mar 2024 - Apr 2024**
  IBM, Coursera

INVOLVEMENT & LEADERSHIP

- **Senior Division NCC 'B' Certificate**
- **Played Nationals in Karate & Kickboxing**