

Safal Kumar Shrestha

Chitwan, Nepal

+977-9811867187 | safalkumarshrestha08@gmail.com | github.com/safalkumarshrestha | linkedin.com/in/safalkumarshrestha | safalkumarshrestha.com.np

PROFESSIONAL SUMMARY

AI/ML Engineer with 3+ years of experience building ML, computer vision, and generative AI systems. Built and deployed 3 applications used by 150+ users. Experienced in Python, PyTorch, FastAPI, LangGraph, RAG, LLMs, Docker, Kubernetes, and MLflow. Skilled in model integration, API development, feature engineering, testing, debugging, and deployment.

EDUCATION

B.Tech in Artificial Intelligence, Kathmandu University 2022 – Present
CGPA: 3.81

Relevant Coursework: Machine Learning, Deep Learning, Computer Vision, NLP, LLMs, Probability & Statistics, Data Structures & Algorithms

Diploma in Computer Engineering, Nepal Polytechnic Institute Completed 2022
Percentage: 81.62%

EXPERIENCE

Independent AI/ML and Application Developer bala.ai 2023 – Present
Part-time

- Built Python backend services for **3 shipped applications**, supporting production AI features.
- Integrated ML models and REST APIs serving **150+ users** across deployed applications.
- Developed computer vision and LLM features across **3 production applications** using PyTorch and modern AI workflows.
- Reused model inference and feature-engineering components across **3 applications** to support consistent AI integration.
- Debugged ML, API, and backend issues across **3 applications** using logs, model outputs, and API responses.
- Validated deployed models across **3 applications** using logs, outputs, API responses, and user feedback.
- Deployed and maintained **5 self-hosted workloads** using Linux and Docker.
- Configured production security across **5 workloads** using SSL/TLS, reverse proxies, and firewall rules.

ML, Computer Vision and IoT Researcher Hydroponics AI, Kathmandu University 2026
Part-time

- Developed a hydroponics automation prototype supported by **1 STEAM research grant**.
- Secured **1 DFRobot hardware sponsorship** for AI and IoT development.
- Built **6 sensing and control workflows** across **2 ESP32 platforms**.
- Integrated **7 environmental signals** into the IoT monitoring pipeline.
- Monitored **7 variables**: pH, EC, CO2, temperature, humidity, light, and water level.
- Built a monitoring application integrating **4 technologies**: FastAPI, MQTT, WebSockets, and Flutter.
- Developed PPO environments for **3 optimization objectives**: plant performance, environmental quality, and energy efficiency.

High-Performance Computing Core Engineer AI HPC Initiative, Kathmandu University 2025
Part-time

- Supported distributed AI workloads on **1 Kubernetes** infrastructure environment.
- Configured HAProxy for distributed AI workload routing across the HPC environment.
- Analyzed **RTX 3080 Ti 12GB** GPU topology for AI resource planning.
- Evaluated GPU virtualization constraints for distributed AI workloads.

PROJECTS

TrekSathi / TrekMate – Travel Platform and AI Agent LangGraph, RAG, Qdrant, Django, MongoDB, Flutter Jul 2025 – Present

- Built a travel platform integrating **5 core technologies**: LangGraph, RAG, Qdrant, Django, and Flutter.
- Generated personalized itineraries using **2 primary inputs**: user preferences and trail information.
- Integrated **2 external services** for geospatial and weather-aware itinerary planning.
- Built retrieval pipelines covering **650+ destinations** and **3,000+ accommodations**.
- Built a LangGraph agent with **2 external tool categories** for location-aware recommendations.

KU Krishi – Agricultural AI Assistant

PyTorch, Computer Vision, LangGraph, FastAPI, Django, Flutter

- Built **3 computer vision features**: Leaf Color Chart analysis, late-blight prediction, and blight segmentation.
- Developed late-blight prediction using PyTorch and integrated the model into a production application used by **150+ users** across deployed projects.
- Built a LangGraph agent integrating **2 tool categories** for crop-disease and market information.
- Published the application with **180+ Google Play downloads**.
- Reached the **2024 National ICT Award semi-finalist** stage.

Satellite-Image Land Segmentation

Python, Computer Vision, GIS, Image Processing

Jun 2026 – Jul 2026

- Built satellite-image preprocessing pipelines across **2 core stages**: image preprocessing and feature preparation.
- Applied feature engineering and compared **2 segmentation approaches** for satellite imagery.
- Evaluated the workflow across **4 stages**: preprocessing, feature extraction, training, and segmentation.

Wilms Tumor Segmentation

PyTorch, Tiny U-Net, Medical Image Segmentation

Jun 2025 – Jul 2025

- Developed **1 Tiny U-Net** architecture for medical image segmentation.
- Built PyTorch workflows covering **4 stages**: preprocessing, training, inference, and segmentation evaluation.

RESEARCH

Preference-Conditioned Multi-Objective Reinforcement Learning for Hydroponics

Recurrent PPO, GCN, LSTM

2026

- Developed a preference-conditioned RL framework using **1 Recurrent PPO** policy with GCN and LSTM integration.
- Optimized **3 objectives**: plant performance, environmental quality, and energy efficiency.
- Compared **2 architectures**, GCN and MLP, across **6+ environmental variables**.
- Evaluated **6+ variables**: pH, EC, temperature, humidity, CO2, and lighting.
- Compared models using **4 measures**: reward, convergence, entropy, and evaluation performance.

Autonomous Driving with CARLA

Reinforcement Learning, Simulation

Aug 2026 – Present

- Researching reinforcement learning for autonomous vehicle control in **1 CARLA simulation environment**.
- Studying **2 control requirements**: safe and traffic-aware vehicle behavior.

RL for Software-Defined Networking

Reinforcement Learning, Network Optimization

Aug 2026 – Present

- Researching reinforcement learning for **2 network-control areas**: adaptive resource allocation and network optimization.
- Studying RL-based approaches for adaptive resource allocation and network control.

TECHNICAL SKILLS

Programming: Python, SQL, C/C++, Java, JavaScript

Machine Learning: PyTorch, TensorFlow, scikit-learn, Neural Networks, Deep Learning, Feature Engineering, NLP, GNNs, Reinforcement Learning

Computer Vision: OpenCV, YOLOv8, Tiny U-Net, Image Processing, Object Detection, Segmentation

Generative AI: LLMs, LangGraph, LangChain, RAG, Qdrant, Embeddings, Semantic Search, Prompt Engineering, Tool Calling, AI Agents, Chatbots

APIs & Model Integration: OpenAI API, Anthropic API, Gemini API, Structured Outputs, REST APIs, Model Inference

Backend & Software Engineering: FastAPI, Django, Django REST Framework, API Design, Backend Development, Testing, Debugging

Databases & Messaging: MySQL, MongoDB, Redis, RabbitMQ

MLOps & Infrastructure: MLflow, Docker, Kubernetes, Linux, Git, CI/CD, Monitoring, SSL/TLS, Reverse Proxies

Cloud: AWS, Cloud AI Services

IoT: ESP32, Arduino, PlatformIO, MQTT, WebSockets, Sensor Data Pipelines

LEADERSHIP & TEACHING

- Held **3 leadership and teaching roles**: Operations Lead, Mentor, and Instructor at Kathmandu University.
- Operations Lead, AI Convergence – Kathmandu University 2025 – Present
- Mentor, DataCamp / AI CANVAS – KUAIC Dec 2025
- Instructor, Python X Git Bootcamp – KUAIC Jul 2025

CERTIFICATIONS, AWARDS & FUNDING

- Earned **1 LangGraph Certificate** from LangChain Academy.
- Received **1 STEAM Research Grant** and **1 DFRobot Hardware Sponsorship** for Hydroponics AI. 2026
- Achieved **3rd place** at AI Crusade. 2024
- Achieved the **highest score** in the KUAIC Python Bootcamp Coding Test. 2023
- Completed a PCB Design and Embedded Systems Workshop at Kathmandu University. Apr 2026