

Birat Gautam

[LinkedIn](#) · [+977 9861061260](#) · biratgautam09@gmail.com · [GitHub](#) · [Portfolio](#)

PROFILE

Computer Science student specializing in Artificial Intelligence, with hands-on experience in Machine Learning, Full Stack, Computer Vision, and Data Engineering. Started with Python and progressed into building real-world, end-to-end systems through projects and internships. Passionate about understanding how systems work under the hood and aiming to grow as an AI/ML Engineer.

SKILLS

Languages: Python, TypeScript, SQL

Frameworks & Libraries: FastAPI, Django, Langraph, PyTorch, NumPy, Airflow, MLFlow, OpenCV, FastMCP

ML & Data: RAG pipelines, LoRA fine-tuning, CNN, YOLO, U-Net, Whisper ASR, ResNet50

Databases: PostgreSQL, ChromaDB (vector), Redis, MariaDB

MLOps & Infra: Docker, Git, llama.cpp, ollama

Cloud & Tools: AWS/Azure/GCP (familiarity), WebSockets, WebRTC, DRF

EXPERIENCE

Jr. AI/ML Engineer

Nov 2025 – Present

Synapse Technologies – On-site

- Engineered a multilingual NLP pipeline for English, Nepali, and Romanized Nepali queries using transformer-based intent classification and script-aware NER; reduced inference latency from ~10s to ~30ms via ONNX optimization and INT8 quantization.
- Developed and productionized the KanoonBox RAG pipeline using ChromaDB vector search and NepSearch-based legal grounding; optimized intent-routing architecture to reduce reasoning latency from ~5s to ~25ms while maintaining response quality.
- Deployed Nepali document OCR using PaddleOCR-VL-0.9B via llama.cpp within 2 GB VRAM (7 s/page); integrated Qwen2.5-VL-3B-Instruct-GGUF into Synapse Intelligence for enterprise document understanding and knowledge delivery.
- Built Nepal's first structured Nepali ASR corpus from YouTube audio using VAD, NFC normalization, and Gemini validation; LoRA fine-tuned Whisper-large-v3-turbo on resource-limited hardware (WER: 33, WandB monitored).
- Benchmarked lightweight Gemma 4 variants on CPU infrastructure using Ollama, llama.cpp, and GGUF quantization formats for on-device LLM deployment feasibility.

Python, PyTorch, Hugging Face, ONNX, LoRA, RAG, ChromaDB, Whisper, llama.cpp, PaddleOCR-VL, Qwen2.5-VL, GGUF, ASR, OCR

PROJECTS

Atri Code: Local-First Agentic Coding Harness

Python, FastAPI, llama.cpp, Gemma 4 26B MoE (GGUF), FastMCP, SQLite, Tavily

- Built a local-first agentic coding harness running quantized LLMs (Gemma 4 26B MoE via llama.cpp) with a deterministic multi-turn ReAct loop, custom MCP tool server (32 tools), and multi-provider web search; engineered hybrid GPU/CPU MoE inference on a 6 GB VRAM GPU by routing attention layers to CUDA and all 118 expert layers to CPU, achieving stable 5–8 tok/s without OOM.
- Designed the full tool-call pipeline - ToolRegistry for MCP tool normalization and risk-tiering, an LLM adapter parsing OpenAI-format and Gemma-native tokens, sandboxed filesystem access with atomic writes and diff-based edit review, and a hook system for lifecycle events (pre/post tool-call, compaction, budget warnings).
- Implemented a premium terminal UI with slash commands, streaming tok/s counters, session tree branching, context compaction, and PLAN mode; added auto-memory and context distillation to persist agent knowledge across sessions, and a prompt-policy engine with per-request profile overrides and configurable thinking mode.

Attendance Management System Using Computer Vision

React, Django, WebSockets, OpenCV, TensorFlow, NumPy

- Led a team of five developers to build a web application that automated attendance tracking using face recognition.
- Implemented real-time face detection and recognition pipeline with OpenCV and TensorFlow, integrated via WebSockets for live updates.

MatchVision: End-to-End EPL Prediction Platform

Airflow, Redis, PostgreSQL, Docker, DRF, MLFlow, Node.js, React.js

- Containerized ML pipeline using Airflow, MariaDB, Redis, and DVC to automate data ingestion, cleaning, validation, and model training for EPL match prediction.
- Tracked experiments with MLflow and served real-time predictions via DRF, integrated into a Django web app displaying live stats, fixtures, and betting probabilities.

Baali Bigyan: AI-Powered Agriculture App

PyTorch, OpenCV, YOLO, Django, PostgreSQL

- Developed a mobile application leveraging CNN (ResNet32) to detect diseases in paddy plants from images, providing treatment recommendations, video tutorials, and in-app product purchases.
- Integrated an expert consultation feature connecting farmers with agro-vet professionals, enhancing accuracy and trust in disease diagnosis and crop management.

Automated Nuclei Segmentation in Microscopy Images

CNN, TensorFlow, U-Net Architecture, OpenCV

- Built a U-Net-based deep learning model for nuclei detection in microscopy images using the 2018 Data Science Bowl dataset, achieving an IoU score of 0.88.
- Automated segmentation reduces manual annotation, handles complex cellular structures, and enhances diagnostic efficiency in resource-constrained medical environments.

EDUCATION

BSc (Hons) Computer Science with Artificial Intelligence

2023 – 2027

Birmingham City University – Birmingham, United Kingdom

High School | Science | GPA: 3.78

2020 – 2022

Uniglobe SS/College – Kamaladi, Kathmandu

PUBLICATIONS

Automated Nuclei Detection in Microscopy Images

May 2025

International Journal on Engineering Technology (InJET)

- Published in InJET. Developed a U-Net-based deep learning model for nuclei detection in microscopy images (2018 Data Science Bowl dataset), achieving IoU 0.88. Automates segmentation, improves accuracy over manual annotation, and enhances diagnostic capabilities in resource-constrained medical environments.
- DOI: <https://doi.org/10.3126/injet.v2i2.78595>

AWARDS & HONORS

ICT Award 2024 – Top 5 Finalist, Rising Star Innovation

2024

Project Baali Bigyan selected as finalist in ICT Award Rising Star Innovation 2024

Sunway College Innovation Fest 2024 – Gold Medal

July 2024

Sunway College Kathmandu – Maitidevi, Kathmandu

- Awarded Gold Medal among 10 teams for Attendance Management System Using Face Recognition. Led a team of 5 and contributed as Backend Developer.

UGSS Achiever Award – Best Student of the Year

April 2022

Uniglobe SS/College – Kamaladi, Kathmandu (GPA: 3.78)

- Awarded Best Student of the Year for the 2020–2022 batch for excellent performance in both academic and non-academic activities.

CERTIFICATIONS

- **Programming with Python Professional Certificate** – OpenEDG Python Institute
- **JavaScript Foundations Professional Certificate** – Mozilla
- **Docker Foundations Professional Certificate** – Docker, Inc
- **Diploma in Python Programming** – Deerwalk Training Center, Kathmandu (July 2023)