

SURAJ PRASAI

Winston-Salem, NC | suraj.prasai97@gmail.com | LinkedIn | GitHub | Portfolio | (743) 240-7785

EXPERIENCE

Wake Forest University

Winston-Salem, NC

Graduate Research Assistant

Aug. 2025 – Present

- Designed and engineered **GeoTALOS**, an AOI-driven geospatial AI platform (evolved from an earlier prototype, **AwakeForest**, **accepted as an IJCAI-ECAI 2026 demo paper**, Bremen, Germany) that replaces ad hoc, duplicated-TIFF workflows across lab drives with one centralized system; owned the backend infrastructure and spatial schema, cutting peak memory usage up to **8x** (66.5GB → 8.4GB, SAM3 segmentation) and data-load latency to near-zero via tiled raster streaming – **submitted, ACM SigSpatial 2026**.
- Built a no-code visual workflow builder so non-programmers can run and automate ML inference pipelines (detection, segmentation, temporal NDVI monitoring) from the map UI; independently built the full **React** frontend using AI-assisted coding tools.
- Built **KnowThyself (AAAI 2026 Demo Track)**, an agentic LLM interpretability assistant; designed an orchestrator-router pipeline (**LangGraph + LLAMA 70B**) that routes natural-language queries to specialized analysis modules, consolidating 5+ fragmented interpretability tools into one chat interface – presented in Singapore, Jan. 2026.

Wiseyak Solutions

Kathmandu, Nepal

AI Engineer

Aug. 2020 – May 2025

- Architected and deployed to production a **RAG chatbot** (hybrid dense + sparse retrieval, LangChain) for **Nepal Telecom** serving real-time customer queries at national scale; co-built a real-time telephony AI pipeline (ASR → LLM → TTS) on **Kafka** producer-consumer streaming and authored the technical proposal that secured **£150K in GSMA** international program funding.
- Curated and augmented audio datasets for automatic speech recognition (**300 hrs supervised, ~2,000 hrs unsupervised**); designed and ran **LoRA** fine-tuning pipelines for **Whisper** ASR models, improving **WER from 22 to 12.8** on Nepali transcription.
- Built a Nepali instruction-tuning corpus using translation models; fine-tuned **LoRA**-based LLAMA3 8B on the resulting dataset and analyzed cross-lingual representations in low-resource settings – published at **COLING 2025 (CHiPSAL Workshop)**.
- Interviewed and grew the ML engineering team from scratch to **10 engineers**; stood up on-premise **GPU server infrastructure (3x RTX 4090)** and LLM serving pipelines (Docker, model quantization) for enterprise clients; **led and mentored** the team throughout tenure.

Integrated ICT Pvt. Ltd.

Kathmandu, Nepal

ML Research Intern

Jul. 2019 – Mar. 2020

- Built an end-to-end **Nepali ASR pipeline**: cleaned a Devanagari speech corpus (font normalization, non-UTF filtering, spelling correction), fine-tuned Wav2Vec2, and fused an n-gram + Transformer LM, improving **WER from 54 to 18.3**; corpus work published in *Nepalese Linguistics* (2022).

PUBLICATIONS

Prasai, S., et al. (2026). *GeoTALOS: A Geoscience Platform for Automated Landscape Observation and Spatial Analysis*. **Submitted, ACM SigSpatial 2026**.

Prasai, S., et al. (2026). *AwakeForest: An Interactive Geospatial Platform for Large-Scale Forest Imagery*. **Accepted, IJCAI-ECAI 2026 Demo Track**, Bremen, Germany.

Prasai, S., Du, M., Zhang, Y., & Yang, F. (2025). *KnowThyself: An Agentic Assistant for LLM Interpretability*. arXiv:2511.03878. **Accepted, AAAI 2026 Demo Track**.

Duwal, S., Prasai, S., & Manandhar, S. (2025). *Domain-adaptative Continual Learning for Low-resource Tasks: Evaluation on Nepali*. **CHiPSAL Workshop, COLING 2025**, pp. 144–153.

EDUCATION

Wake Forest University

Winston-Salem, NC

M.S. Computer Science

Aug. 2025 – May 2027

Tribhuvan University

Kathmandu, Nepal

B.S. Computer Science and Information Technology | 77.1%

2015 – 2019

SKILLS & LEADERSHIP

Agentic & LLM: LangGraph, LangChain, RAG (hybrid dense+sparse), LLAMA, Tool-use Agents, Multi-agent Orchestration

ML & DL: PyTorch, HuggingFace, LoRA fine-tuning, Whisper, Wav2Vec2, Scikit-learn

Geospatial & Data: FastAPI, PostgreSQL/PostGIS, pgSTAC, TiTiler, Data-centric AI, Pandas, NumPy

Systems & Cloud: Docker, Kubernetes, AWS, GCP, Azure, Kafka, Celery/Redis, LLM serving and quantization, REST APIs, OpenTelemetry/distributed tracing

Languages & Fundamentals: Python (primary), JavaScript/React, Data Structures & Algorithms

Leadership: Co-founder, *Return to Heritage* NGO (digitizing Nepal's cultural heritage & manuscripts); taught C and Python to undergraduates; expert talks at Kathmandu University on bridging academia and industry