Biky Dev

Nepal | +977 9824718889 | deobiky@gmail.com | LinkedIn | GitHub

Summary

AI/ML Engineer with hands on experience building Computer Vision systems, RAG pipelines, and LLM-powered applications. Built and deployed 3 personal AI projects currently live: a multi-agent orchestrator with sub-200ms streaming, an enterprise RAG chatbot reducing irrelevant responses by 60 percent, and a crop disease CNN achieving 90 percent accuracy with 40 percent faster inference after quantization. Strong foundations in full-stack AI architecture, production-ready Python, and scalable MLOps deployment.

Education

Kurukshetra University Kurukshetra, Haryana, India
Bachelor of Technology, Computer Science and Engineering *Graduated May 2026*

Technical Skills

AI and Deep Learning: Computer Vision (OpenCV, MTCNN), CNNs, LLMs (Llama 3.2), RAG (LangChain, LangGraph), NLP (BERT, Transformers), PyTorch, TensorFlow, Keras, Hugging Face

Data Science and MLOps: Pandas, NumPy, ONNX Model Quantization, Vector Databases (FAISS, Pinecone), Scikit-Learn, XGBoost, Random Forest, SMOTE

Engineering and Tools: Python, SQL (PostgreSQL), MongoDB, Redis, FastAPI, React, Next.js, Docker, Git, AWS, Node.js

Professional Experience

LearnXChain (LXC) Haryana, India
AI and Computer Vision Engineer *Feb 2025 – May 2026*

- Engineered an end-to-end Automated Face Attendance System using OpenCV and MTCNN, boosting biometric verification accuracy by **25 percent** across live classroom deployments.
- Built and deployed a concurrent REST API backend to ingest real-time biometric streams and persist structured session logs into a relational database, reducing data processing overhead by **30 percent** at production scale.
- Optimized the multi-threaded video inference pipeline to sustain a stable **30 FPS** throughput under continuous production load, supporting over 100 concurrent users.

Key Projects

Multi-Agent LLM Orchestrator | *LangGraph, Redis, PostgreSQL, Python, SSE* Live Demo

- Architected a multi-agent orchestration system using LangGraph with 4 specialized agents (researcher, planner, executor, validator), enabling autonomous completion of complex multi-step tasks.
- Implemented dual-layer agent memory: Redis for short-term session context and PostgreSQL for long-term task history, enabling reliable state recovery across interrupted sessions.
- Exposed the pipeline via a streaming REST API using Server-Sent Events, delivering real-time agent output with first-token latency under 200 milliseconds.

RAGChat: Intelligent Document Chatbot | *LangChain, Pinecone, Llama 3.2, Python* Live Demo

- Built a production-grade RAG pipeline with hybrid dense and sparse retrieval over enterprise document corpora, reducing irrelevant responses by **60 percent**.
- Engineered custom chunking strategies and iterative prompt templates that improved contextual retrieval precision, cutting average query resolution time by **35 percent**.

CropGuard: Plant Disease Diagnosis | *Custom CNN, TensorFlow, Keras, ONNX, XGBoost* Live Demo

- Designed and trained a custom CNN architecture from scratch on a 15-class crop disease dataset covering Pepper, Potato, and Tomato leaf conditions, achieving **90 percent accuracy**.
- Exported the model to ONNX and applied post-training quantization, cutting inference latency by **40 percent** and enabling deployment on low-resource edge hardware.

Research

Independent ML Research | *Google Colab*

- **Financial Modeling:** Built a Credit Card Fraud Detection classifier using XGBoost and Random Forest; applied SMOTE oversampling and threshold calibration, improving Recall by **18 percent** on imbalanced data.
- **Sequence Modeling:** Implemented LSTM and fine-tuned BERT for Fake News Detection and Sentiment Analysis, with BERT outperforming the LSTM baseline by **12 percent** on F1 score.