FIROJ RAUT

AI Developer

Phone: +977-9805352767 Email: firojraut094@gmail.com

GitHub: github.com/feerose111 LinkedIn: linkedin.com/in/firojraut1

Summary

AI-focused developer leveraging the power of **Python** to build intelligent, data-driven, and production-ready systems. Experienced in developing **machine learning**, **deep learning**, and **AI-powered applications** using frameworks like FastAPI, Django, TensorFlow, and PyTorch. I specialize in integrating ML models into scalable backends, designing data pipelines, and optimizing inference for real-time use cases. Passionate about turning research into deployable AI products through efficient code, experimentation, and system-level thinking.

Education

Ambition College, Mid-Baneswor, Kathmandu 03/2021 – 09/2025
BSc. CSIT (Bachelor in Computer Science and Information Technology)

Arniko Awasiya Secondary School, Biratnagar, Koshi 08/2018 – 03/2021
+2 Computer Science

Work Experience

AI/ML Intern – eSewa Pvt. Ltd., Pulchowk, Lalitpur 04/2025 – 07/2025

- Built and deployed machine learning models into production systems using **Python**, **FastAPI**, and deep learning frameworks.
- Developed scalable pipelines for data preprocessing, model evaluation, and API integration.
- Enhanced model optimization and real-time inference performance in production-grade systems.

Projects

ASAPP – AI-Powered Project Planning & Assistant (Personal Project)

- Developed an intelligent project planning system integrating **FastAPI** (backend) and **Streamlit** (frontend) for seamless user interaction.
- Implemented **ChromaDB** as a persistent vector database with **Hugging Face embeddings** to enable contextual and semantic understanding.
- Built a **context-aware chatbot** that retrieves project insight from embeddings leveraging the power of **LangChain**, delevering personalized responses.
- Integrated a **custom logging system** using the **Observer pattern** to track system events, improving transparency, and streamline debugging.
- Containerized the system using **Docker** and deployed using **Render's** services, ensuring persistence, modularity, and scalability across environments.

Semantic AI Search System using LangChain, Milvus & Elasticsearch (Internship Project)

- Engineered a hybrid search system using **LangChain**, **Milvus (vector DB)**, and **Elasticsearch (keyword DB)**.

- Built a **semantic search algorithm** that leverages similarity from past query embeddings and a dissimilarity score to rank results, improving personalization and diversity in search outcomes.
- Ingested structured JSON documents using **Kafka** and implemented upsert and hybrid retrieval.
- Deployed backend using **FastAPI** with monitoring via **Streamlit**, scaling with **Redis** and **Docker**.

Facial Emotion Recognition Using CNN for Song Mapping (Academic Project)

- Implemented CNN-based emotion detection using Python and **Django** for real-time emotion-to-song mapping.
- Optimized the CNN architecture to reduce the parameters using **Depthwise Separable Convolution** with accuracy upto **79.41 %**.
- Developed backend logic integrating webcam input, CNN inference, and a music recommendation pipeline.

Data Analysis and Machine Learning Projects (Personal Projects)

- Created ML pipelines using **Scikit-Learn** and applied algorithms such as Decision Trees, KNN, SVM, and Regression.
- Completed projects like Heart Disease Prediction, Parkinson's Detection, Spam Classification, and Wine Quality Analysis.

Technical Skills

Languages: Python, HTML, CSS, JavaScript
Web Frameworks: Django, FastAPI, Streamlit, RESTful APIs
Databases: MySQL, PostgreSQL
Machine Learning: Scikit-Learn, TensorFlow, PyTorch, LangChain, Sentence-Transformers
Tools & Platforms: Git, GitHub/GitLab, Docker

Soft Skills

Adaptability and continuous learning
Analytical and problem-solving skills
Collaboration and communication
Ownership and accountability

Passions

Artificial Intelligence, Generative AI, Music, Art