Kiran Punna

Protfilo.in |Github.com| linkedin.com | kiranpunna58@gmail.com | +91 9381911235

Summary

AI/ML & Generative AI Engineer experienced in building end-to-end ML solutions, including OCR pipelines and real time detection systems.Skilled in Python, deep learning, and deploying production-ready models using FastAPI and Streamlit. Focused on scalable AI systems, RAG-based applications, and personalized recommendation models.

Work Experience

Infosys Springboard — AIML Intern (Nov 2025 – Feb 2026)

- Built an end-to-end Medical Risk Classification and Personalized Diet Recommendation System using OCR-extracted clinical data with robust DVC pipelines for data versioning and reproducible preprocessing workflows.
- Implemented ensemble models (XGBoost, Random Forest) achieving 99% accuracy, along with a pretrained BERT-based medical intent classifier for precise query understanding.
- Designed a hallucination-free RAG framework with vector databases to generate personalized diet plans.
- Built with MLOps readiness using GitHub Actions CI/CD and AWS for scalable deployment, monitoring,and continuous improvement.

Projects

WildVision — Real-Time Animal Detection

A real-time animal detection system identifying 95+ animal species with 97% accuracy using CNN models. Built with PyTorch and Flask, it streams live camera input with a responsive interface for identification and classification.

Tech Stack: Python, PyTorch, Flask, OpenCV, HTML, CSS, Bootstrap, NumPy, Pandas

AI Vision Sentinel — Real-Time Face Recognition

A face recognition system that detects faces in real time using PyTorch and OpenCV. Integrates web scraping with BeautifulSoup to retrieve information about the recognized person from Wikipedia, providing an intelligent data display.

Tech Stack: Python, PyTorch, OpenCV, BeautifulSoup, Flask, HTML, CSS, JavaScript

Multi-Prediction ML System

A machine learning system with prediction models for diabetes, book recommendations, and digit recognition. Built with Python and Streamlit for an interactive, user-friendly interface.

Tech Stack: Python, Pandas, NumPy, Scikit-learn, PyTorch, Streamlit, Matplotlib, Seaborn

SatCap — Satellite Image Captioning

A real-time satellite image captioning system that generates accurate,descriptive captions for images using pretrained CLIP and GPT-2 models. Built with FastAPI and Flask, it allows users to upload images and receive instant captions through a responsive web interface.

Tech Stack: Python, PyTorch, Transformers (CLIP & GPT-2),FastAPI, Flask, OpenCV, HTML, CSS, JavaScript, NumPy, Pandas

DocQGen — Document-Based Exam Q&A Generator

An AI-powered system that converts uploaded documents into exam-ready short-answer questions using RAG, ensuring zero hallucinations and strictly document-grounded outputs. Designed for exam preparation, question paper automation, and AI-assisted learning.

Tech Stack: Python, LangChain, NVIDIA LLaMA 3.1, Pinecone, FastAPI, HTML, CSS, JavaScript

Education

B.Tech (CSE–AI), Annamacharya Institute of Technology & Sciences

2023 – Present | CGPA: 9.3

Skills

Programming Languages: Python, Java, SQL, JavaScript

Databases: MongoDB, MySQL, PostgreSQL

Vector Databases: Pinecone, Chroma

Frameworks & Libraries: PyTorch , Scikit-learn, NumPy, Pandas, FastAPI, Flask, React, Streamlit

Tools: Git, GitHub, Postman

MLOps & Cloud: Docker, GitHub Actions (CI/CD), AWS

ML & Data Skills: ML, DL (CNNs, RNNs, LSTMs, Transformers), NLP, Data Preprocessing, Visualization

Generative AI Applications: LLMs, LangChain, Hugging Face, Prompt Engineering, RAG

Achievements

- **ML Mania — National Level Tech Fest:** Secured **2nd place** at Project Explore for the **NoteCraft** project.
- **Touchstone Elocution Competition:** Winner of college-level elocution and essay writing competitions.
- **Young Trucks Competition — Naukri:** Secured **98.5 percentile** at the national level.
- **ALGO Scholarship Program — Nationwide:** Achieved **Top 7.8%** in coding rounds and applications.