Prashanna Raj Pandit

ppandit@siue.edu | linkedin.com/in/ra-prashanna | prashanna.vercel.app | github.com/Prashanna-Raj-Pandit

Summary

Data Scientist and ML Engineer with 2.5+ years of experience owning the full ML lifecycle - from feature engineering and model development through production deployment, monitoring, and GenAI integration. Proficient in Python, SQL, and R across supervised/unsupervised learning, deep learning, NLP, computer vision, and agentic AI. Experienced translating complex models into business insights for technical and non-technical stakeholders. AWS Certified ML Engineer (Associate).

Technical Skills

Languages: Python, R, SQL, Bash
ML/DL: scikit-learn, TensorFlow, PyTorch, XGBoost, LightGBM, GLM, Ridge/Lasso, BiLSTM, CNNs, YOLO, BERT, Transformers, Bayesian optimization (Optuna), statistical modeling, hypothesis testing
GenAI/LLMs: OpenAI API, LangChain, LangGraph, RAG pipelines, ChromaDB, FAISS, prompt engineering, agentic workflows, Hugging Face Transformers, fine-tuning
Data/Viz: Pandas, NumPy, EDA, feature engineering, signal processing, Tableau, Power BI, matplotlib, seaborn
MLOps/Cloud: AWS (SageMaker, EC2, S3, ECR, CloudWatch), Docker, GitHub Actions, CI/CD, FastAPI, Flask, MLflow, PostgreSQL, ETL pipelines

Experience

Research Assistant - Machine Learning Jan 2025 - Present
Southern Illinois University Edwardsville Edwardsville, IL

- Designed end-to-end clinical AI pipeline for pediatric autism detection; engineered 15+ biomechanical features from 500,000+ raw gait video frames (OpenPose, MediaPipe, Python, NumPy); applied Savitzky-Golay filtering and spline interpolation to reduce time-series noise by 40%.
- Developed and validated a BiLSTM model via nested cross-validation and Bayesian optimization (Optuna), achieving 88% accuracy; ensemble deployment cut false negatives by 25% - a clinically significant outcome.
- Deployed scalable real-time inference on AWS SageMaker (Docker/ECR, Flask, EC2, GitHub Actions) with ∼200ms latency; implemented CloudWatch monitoring, alerting, and drift detection for long-term model health.

Graduate Teaching Assistant - CS 325 & CS 584 Jan 2026 - Present
Southern Illinois University Edwardsville Edwardsville, IL

- Guided 38 students in ML and GenAI projects including building RAG chatbots with LLMs and vector databases; mentored in Python, statistical modeling, and data visualization.

Machine Learning Engineer *(Promoted from Software Engineer)* Jan 2023 - Jan 2025
Vrit Technologies Kathmandu, Nepal

- Built and shipped end-to-end supervised ML pipelines - data ingestion, feature engineering, model training, REST API deployment on AWS - reducing release time by 40% via CI/CD (GitHub Actions, Docker).
- Engineered automated ETL pipelines from heterogeneous structured and unstructured sources using SQL and Python, reducing manual processing time by 60% and enabling real-time analytics dashboards for 3+ business teams.
- Conducted EDA and statistical modeling on large datasets; translated findings into actionable visualizations for stakeholders; implemented production monitoring, alerting, and retraining triggers to ensure model reliability.
- Implemented production monitoring, automated retraining triggers, and alerting for deployed ML services, reducing model degradation incidents by 30% and improving system uptime.

Projects

EvidenceCV - Agentic RAG Application | *Python, FastAPI, OpenAI API, ChromaDB, LangChain* | GitHub 2026

- Architected a production agentic LLM system with OpenAI function calling, multi-source retrieval (GitHub APIs, PDFs), HNSW-indexed ChromaDB embeddings, and an evaluation framework benchmarking retrieval accuracy and generation quality; shipped full-stack product (FastAPI, React) in 36 hours – 3rd place, eHacks 2026.

MedNLPify - Medical Abstract Classifier | *Python, TensorFlow, BERT, CNN, Transformer, spaCy* | GitHub 2025

- Designed a Tribrid deep learning architecture combining transfer learning (Universal Sentence Encoder, TensorFlow Hub), character-level CNNs, and positional token embeddings to classify 200,000+ PubMed RCT sentences into 5 clinical categories.
- Applied dropout, L2, batch normalization, and LR scheduling – improving accuracy from 72.5% to 85.6% (18% gain, F1 0.84); built full pipeline (spaCy, Flask REST API, Docker), shipped as Chrome extension; GitHub Actions CI/CD cut build time by 45%.

Production ML Pipeline on AWS | *Python, AWS SageMaker, Fargate, CloudWatch, Docker, Flask, GitHub Actions* 2025

- Architected and deployed a production-grade ML inference pipeline on AWS SageMaker – containerized model serving via Fargate (serverless) and Flask REST API, automated retraining triggered by data drift signals, and full CI/CD via GitHub Actions; implemented CloudWatch dashboards with latency, throughput, and drift monitoring to maintain long-term model health in a real clinical environment.

Education & Awards

MS, Computer Science Aug 2024 - May 2026
Southern Illinois University Edwardsville GPA: 3.9/4.0

BE, Electronics, Communication and Information Engineering Nov 2018 - Mar 2023
Tribhuvan University, IOE Pulchowk Campus GPA: 3.75/4.0

AWS Certified ML Engineer - Associate | Credly Badge | Best Paper ICT-CEEL 2023 (YOLO 95% mAP) | Nepal Hackathon Winner 2023 | eHacks 2026 3rd Place