

Kiran Shrestha

504-377-9716 | shresthakiran2069@gmail.com | <https://linkedin.com/in/kiranrajstha> | <https://github.com/stharajkiran>

EDUCATION

Tulane University

MS in Computer Science - GPA: 3.65

May 2025

Southeastern Louisiana University

BS in Mathematics - GPA: 3.9

2020

SKILLS

Programming Languages: Python, SQL, Java, MATLAB, C, JavaScript

Frameworks And Libraries: PyTorch, TensorFlow, Keras, LangChain, LangGraph, HuggingFace, Sklearn

Visualization And Data Analysis: Matplotlib, Seaborn, Pandas, Tableau

MLOps And Cloud: CI/CD, Model deployment, MongoDB, API, DVC, MLflow, W&B, AWS, Kafka, Docker, Qdrant, GitHub Actions, Git

Machine Learning And AI: Classification, Clustering, Anomaly detection, Computer Vision, Deep Learning, Transformer models, Large Vision Models, Large Language Model (LLM), RAG, Regression, Finetuning, Model Training, Few-shot learning, Unsupervised learning, Segment Anything Model (SAM), Cross-validation, Hyperparameter optimization

Software Engineering: HTML, CSS, React, FastAPI, Node.js, Streamlit, REST APIs

WORK EXPERIENCE

Machine Learning Researcher (Graduate Research Assistant)

2022 - 2025

Tulane University

- Built an unsupervised anomaly detection framework utilizing self-supervised DINO features and multivariate Gaussian distribution modeling; achieved State-of-the-Art (SOTA) benchmarks on MVTec AD (96.3% pAUROC) and ViSA (97.1% pAUROC) in 1-shot scenarios.
- Developed and fine-tuned large vision models (LVMs), implementing custom part-attention modules in PyTorch to discover human-interpretable visual prototypes from image-level labels without part-level annotations.
- Designed an automated defect localization pipeline for high-noise X-ray tomography by combining Fast Fourier Transforms (FFT) for prompt generation with the Segment Anything Model (SAM) for unsupervised segmentation.
- Managed the entire research infrastructure lifecycle, from writing highly optimized CUDA-accelerated training scripts to tracking experiments via Weights & Biases (W&B) and managing codebase version control.

RESEARCH EXPERIENCE

Prototype-Part Interpretable Image Recognition

2024-09

Tulane University

- Fine-tuned DINO with a custom part-attention module to discover meaningful object parts using only image-level labels, requiring no part annotations
- Developed a prototype-based classification framework where prototypes are actual image regions rather than abstract latent vectors, making model decisions directly human-interpretable
- Analyzed part-attention embeddings via clustering to validate semantic consistency of discovered parts and explored clustering-based explanations as a lightweight alternative to prototype matching

Concept Distribution Modeling for Few-Shot Anomaly Detection

2023-09

Tulane University

- Engineered a novel concept-based distribution framework that semantically divides industrial images into unique visual prototypes using k-means clustering and a saliency-assisted voting mechanism
- Developed a robust reference bank of normal prototypes by modeling multivariate Gaussian distributions of deep features, significantly reducing the dependency on large volumes of training data
- Achieved State-of-the-Art (SOTA) performance on the MVTec AD and ViSA benchmarks, securing 96.3% and 97.1% pAUROC respectively in 1-shot scenarios while outperforming existing methods in F1-max and PRO scores
- Optimized anomaly localization precision by implementing Mahalanobis distance scoring for test patches, enabling the model to effectively discern defects in small and complex objects without global feature reliance

Unsupervised Crack Detection in X-ray Tomography using SAM

2023-01

Tulane University

- Designed an unsupervised crack detection pipeline using FFT-based frequency-domain filtering to automatically identify crack regions and generate bounding box prompts for SAM, enabling annotation-free defect localization

- Refined crack boundary segmentation using SAM conditioned on FFT-derived prompts, improving spatial precision of detected defects in high-noise X-ray tomography images
- Characterized the sensitivity of segmentation quality to FFT frequency thresholds and SAM prompt density, establishing reliable operating ranges for robust defect detection

PROJECTS

VisualVault - Semantic Image Search & Digital Asset Management Platform

2026

- Designed a multi-service ML platform from scratch: FastAPI + Celery workers for async inference, Qdrant for vector search, and a Streamlit UI - deployed via Docker Compose with GPU-aware worker containers
- Optimized CLIP inference end-to-end: exported to ONNX, compiled a TensorRT engine, and built a model registry to swap providers at runtime without restarting the API - 3.8x latency reduction in practice
- Closed the retraining loop automatically: YOLO detections get flagged, pushed to Label Studio for human review, written to Postgres, and Celery Beat triggers fine-tuning once corrections hit a threshold - no manual intervention needed
- Added embedding drift detection to catch distribution shift on incoming images before it silently degrades search quality; the OOD test set scored 2.3x above the alert threshold, confirming the signal was real
- Extended the platform from images to video - FFmpeg extraction feeding into the same async Celery pipeline, with a WebSocket live feed for real-time previews

ResearchMind - Personal Research Intelligence System

2026

- Engineered a hybrid retrieval pipeline over 1,317 ML papers combining BM25 and dense retrieval with RRF fusion; benchmarked embedding models and FAISS index types, achieving Recall@10 of 0.97 with MPNet and 91% p95 latency reduction with HNSW32
- Built a 7-tool LangGraph agent with intent-based routing using a configurable LLM backend supporting both local and API-hosted models at temperature 0, achieving 100% classification accuracy on 110 labeled queries, supporting gap detection, methodology comparison, citation graph traversal, and multi-turn memory
- Implemented a 4-stage RAG validation pipeline covering citation grounding, hallucination scoring, and gap schema validation with a rejection threshold of 0.70, averaging a hallucination score of 0.80 on eval; clustered low-rated queries using k-means to surface systematic retrieval failures
- Deployed on HuggingFace Spaces with a FastAPI backend, Redis caching achieving 93% latency reduction, and sustained 15 RPS at 80 concurrent users under Locust load testing with a RAGAS faithfulness baseline of 0.775

ClearSignal - Evaluation-First Fraud Explanation System

2026

- Built the full pipeline that includes domain pretraining, LoRA fine-tuning (SFT and DPO), and vLLM serving with two hot-swapped adapters, reaching sub-second p95 latency (roughly 4 - 10x faster than a GPT-4o call) and cutting GPU memory ~47% by sharing one base model.
- Wrote an evaluation harness with automated hallucination checks, adversarial test cases, and an LLM judge that decided whether the model was good enough to ship, using pass/fail criteria fixed in advance so results couldn't be gamed.
- Added a router that keeps easy cases on the local model and escalates hard ones, which cut frontier-model cost about 63%, plus three safety guards that catch missing or malformed inputs before the model can respond.

Vehicle Insurance Response Prediction - Production ML Pipeline

2025-01

- Designed and implemented a FastAPI application to predict customer response for vehicle insurance using an industry-standard ML pipeline
- Integrated AWS S3 for model deployment, retraining, and versioning
- Containerized the application using Docker and deployed on AWS EC2 instances
- Automated model training, evaluation, and deployment pipeline using GitHub Actions CI/CD

RefHub - Referee & League Management Platform

2025

- Built a full-stack PWA with a React + TypeScript frontend and Python FastAPI backend, featuring a custom HTML scraper that ingests live fixture data from league websites using httpx and BeautifulSoup across 6+ divisions
- Implemented a smart shift builder with multi-filter fixture search across league, date, time range, and field - powered by parallel async API calls and client-side memoized filtering, installable on Android as a native-like app
- Automated post-shift email reporting via SendGrid with incident tracking, suspension logging, and shift history - deployed on Vercel and Render with CI/CD via GitHub push triggers