LinkedIn
jainpalak3286@gmail.com

Palak Jain

Website ↗

Salt Lake City, UT

Senior Data Scientist with 6+ years of experience building production grade computer vision and NLP systems for fraud detection, document intelligence, and multimodal understanding. Experienced in translating research prototypes into scalable ML solutions and collaborating closely with cross-functional teams and domain experts.

Experience

Data Scientist - Claims Fraud Detection **February 2023 - Present**
Verisk Analytics **Lehi, UT**

- **Deepfake and Image Forensics**
  - Designed and productionized a CNN-based deepfake detection system trained on synthetic data (DALL·E, Midjourney, Stable Diffusion), achieving 80% precision with <2% false positives, reducing false claim approvals and enabling scalable deployment via a Dockerized Gradio inference app.
  - Built a pixel-level manipulation detection pipeline using deep CNNs and heatmap-based localization, achieving 94% accuracy (<1% False positive rate) while providing visual evidence that increased adjuster trust in fraud review workflows.
- **Internet Duplication Detection**
  - Developed a hybrid internet image duplication detection pipeline combining ORB features, ResNet embeddings, and SSIM scoring to identify externally sourced claim images, achieving >90% precision and significantly reducing adjuster review workload.
- **Automated Claims Pre-Filtering System**
  - Led end-to-end development of a 32-class claims image pre-screening classifier using CLIP embeddings and CNN models, enabling automated filtering of irrelevant claim images and improving downstream fraud-detection throughput by 30%.
- **ML Ops & Data Quality**
  - Owned dataset curation, error analysis, and iterative model improvement using FiftyOne, resulting in faster labeling cycles and more robust fraud detection models across multiple pipelines.
  - Built Dockerized inference apps using Flask and Gradio, enabling scalable internal demos and real-time fraud-screening solutions.

Data Science Intern **June 2022 - August 2022**
Verisk Analytics **Jersey City, NJ**

- Built lightweight face and text detection pipelines for PII (Personally Identifiable Information) redaction using computer vision models optimized with OpenVINO, achieving 85% recall at the rate of 20 images/sec.
- Implemented efficient real-time inference pipelines further reducing compute overhead for large-scale image processing.
- Designed automated data-integration workflows that reduced the need for resurveying 20K+ underwriting cases, resulting in major cost savings.

Research Assistant **January 2021 - December 2022**
Indiana University- Data Science **Indianapolis, IN**

- Extracted causal relationships from 1M+ biomedical sentences using SRL (Semantic role labeling), dependency parsing, and statistical weighting.
- Developed a BiLSTM-Attention model in PyTorch to capture bidirectional context, achieving ROC-AUC 0.98 on benchmark datasets, outperforming previous causal extraction baselines.
- Fine-tuned BERT, RoBERTa, SciBERT models on the CauseNet corpus, achieving +8% F-score improvement over baseline.
- Optimized transformer attention behaviors to mitigate semantic drift, improving accuracy on time-sensitive biomedical facts.

Data Engineer **January 2019 - November 2020**
Infosys Limited **Hyderabad, India**

- Designed and optimized 350+ ETL workflows across 150+ tables, improving pipeline throughput by 200% through SQL optimization, partitioning, and parallelization laying the data foundation for analytics and ML initiatives.
- Catalogued financing data for e-contract utilization from OLTP datasets and unstructured data sources.

Technical Skills

Programming & Data	Python, Bash, SQL, HTML Pandas, NumPy, Matplotlib, Seaborn
ML & DL	Supervised, Unsupervised learning, Model Evaluation, Hyperparameter Tuning, Error Analysis CNNs, RNNs, BiLSTM, Attention Mechanisms
Computer Vision	Image Classification, Object Detection, Heatmap Localization, Image Fraud Detection, Feature Matching OpenCV, OpenVINO Toolkit, FiftyOne
NLP & LLMs	Transformers (BERT, GPT etc.), Causal Inference, Semantic Search, Dependency Parsing, Embeddings, Fine-tuning
Multimodal & Applied AI	Vision-language models (CLIP), Multimodal Retrieval (text, image, video), Speech-to-Text Pipelines for Downstream ML
MLOPs & Deployment	Dataset Curation, Model Validation, Real-time and Batch Inference pipelines Docker, Flask, Gradio
Cloud, Data & Analytics	AWS (EC2, S3), MySQL, SQL Server, ETL
Analytics & Statistics	Statistical Modeling, Hypothesis Testing, Predictive Modeling, Exploratory Data Analysis, PowerBI

Education

Master of Science, Applied Data Science, Indiana University Indianapolis **January 2021 - December 2022**

Bachelor of Engineering, Electronics and Telecommunication, Devi Ahilya University, India **July 2014 - May 2018**

Personal Projects

Multimodal Video Understanding System

- Built an end-to-end multimodal video understanding system converting long-form videos (5–90 mins) into searchable semantic embeddings and structured summaries using speech-to-text, embedding models, retrieval, and summarization pipelines.
- Designed the system with modular inference stages(video ingestion, speech-to-text conversion, embedding generation, retrieval and summarization) to enable scalability and experimentation with different embedding.

RAG based Document Intelligence System

- Designed and built an end-to-end Document Intelligence Retrieval-Augmented Generation (RAG) system to extract, embed, and semantically retrieve information from complex PDF documents for grounded question answering.
- Implemented a semantic retrieval and ranking pipeline that converts unstructured text, tables, and images into a vector space, significantly improving answer relevance and reducing hallucinations compared to keyword search.

Other projects ↗

Publications

[1] Sydney Anuyah, Jack Vanschaik, Palak Jain, Sawyer Lehman, and Sunandan Chakraborty. ***An Empirical Study of Causal Relation Extraction Transfer: Design and Data***. 2025. URL: https://arxiv.org/abs/2503.06076.

[2] Jack VanSchaik, Palak Jain, Anushri Rajapuri, Biju Cheriyan, Thankam Thyvalikakath, et al. ***Using transfer learning-based causality extraction to mine latent factors for Sjögren's syndrome from biomedical literature***. 2023. URL: https://www.cell.com/heliyon/fulltext/S2405-8440(23)06473-3.