LinkedIn
jainpalak3286@gmail.com

Palak Jain

Website
Salt Lake City, UT

Data Scientist with 5+ years of experience building scalable AI solutions in computer vision and NLP. Proven track record of reducing digital media fraud, saving costs, and improving decision-making with large vision models deployed in production.

Experience

Data Scientist - Claims Fraud Detection **Feb 2023 - Current**
Verisk Analytics **Lehi, UT**

- **Deepfake Detection-** Developed and deployed a CNN-based deepfake detection system using datasets curated from DALL·E, Midjourney, and Stable Diffusion; achieved 80% precision with <2% false positives, reducing false claim approvals and enabling scalable fraud screening via a Dockerized Gradio demo app.
- **Pixel Manipulation Detection-** Designed and productionized a deep learning image forensics pipeline that detected pixel-level splicing with 94% precision (<1% FPR), preventing millions in fraudulent insurance payouts. Integrated explainable AI with heatmap-based localization, increasing adjuster trust and adoption.
- **Internet Duplication Detection-** Built a hybrid computer vision pipeline combining ORB features, ResNet embeddings, and SSIM scoring to detect duplicate claim images from the internet, flagging fraudulent submissions with >90% precision and reducing manual review effort.
- **Fraud Claim Pre-Filtering System-** Led the end-to-end development and deployment of a 32-class image classification system using CLIP embeddings, enabling automated pre-filtering of irrelevant claims images and improving downstream fraud detection efficiency by 30%.
- Leveraged **FiftyOne** to streamline dataset curation and error analysis, accelerating labeling workflows and boosting model performance across multiple fraud detection pipelines.

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

- Implemented multiple Computer vision models to detect text and face in scene images, safeguarding personally identifiable information (PII) in insurance claims data.
- Designed a computationally efficient CV model using OpenVINO Toolkit, achieving 85% recall and processing speed of ~20 images/second, enabling faster and cost-effective large-scale screening.
- Reduced the need for resurveying 20K+ commercial underwriting cases by building an automated data integration pipeline that extracted and merged information from Verisk and its acquired companies, resulting in significant cost savings.

Research Assistant || National Science Foundation(NSF) **Jan 2021 - Dec 2022**
Indiana University **Indianapolis, IN**

- Extracted causal relationships from 1M+ biomedical and scientific sentences using a hybrid pipeline of semantic role labeling (SRL), dependency parsing, and statistical weighting to quantify causal strength.
- Developed a BiLSTM RNN in PyTorch with attention mechanism to capture bidirectional context, achieving ROC-AUC 0.98 on benchmark datasets, outperforming traditional sequence models.
- Adapted BERT, RoBERTa, and SciBERT on CauseNet corpus for cause-effect extraction; improving F-score by 8% over baseline.
- Optimized BERT attention and embeddings to reduce semantic drift, improving accuracy in detecting outdated or time-sensitive facts.

Data Engineer **Jan 2019 - Nov 2020**
Infosys Limited **Hyderabad, India**

- Catalogued financing data for e-contract utilization from OLTP servers and flat files using Informatica.
- Designed, developed, and tested 350+ ETL mappings and workflows in Informatica PowerCenter, managing data across 150+ tables.
- Optimized SQL queries for unit testing and improved ETL pipeline performance by 200% through partitioning and parallel processing techniques.

Technical Skills

Programming	Python, SQL, HTML
Database and Cloud	RDBMS (MySQL, SQL Server), ETL (Informatica PowerCenter), Cloud (AWS S3, AWS EC2)
Analytics Tools	Power BI, Microsoft Excel (Advanced), Matplotlib, Seaborn
Statistical Skills	Statistical Modeling, Hypothesis Testing, Predictive Modeling, Exploratory Data Analysis, Data Mining, Parameter Optimization
Machine Learning and Deep Learning	Supervised & Unsupervised Learning, Neural Networks, CNNs, RNNs, BiLSTM, Word Embeddings (Word2Vec, GloVe), Dimension Reduction
Natural Language Processing (NLP)	BERT, SpaCy, NLTK, Cause-Effect Extraction, Causal Inference
Computer Vision	OpenCV, OpenVINO Toolkit, Image Classification, Object Detection, Heatmap Localization, FiftyOne
Model Deployment	Flask, Docker, Gradio, Cloud Deployment (AWS EC2)

Publications

1. VanSchaik, J. *et al.* Using transfer learning-based causality extraction to mine latent factors for Sjögren's syndrome from biomedical literature. *Heliyon* (2023).

Education

Master of Science, Applied Data Science, Indiana University Indianapolis — **Jan 2021 - Dec 2022**
Bachelor of Engineering, Electronics and Telecommunication, Devi Ahilya University, India — **July 2014 - May 2018**

Personal Projects

Pre-Owned Car Market | Data Visualization | Prediction Model | PowerBI | Flask | Heroku
- Created interactive PowerBI dashboard to visualize and investigate car price variation with 10+ features of the car.
- Deployed prediction model on Heroku cloud platform for online estimation of pre-owned car prices.

Diabetes Onset Prediction | NLP | Pytorch | Artificial Neural Networks(ANN) | GPU
- Determine the patient's diabetic condition based on rapidly diagnosable measures including Blood Pressure, Glucose level, and BMI.
- Developed an ANN classification model using PIMA Indian Diabetes Dataset and PyTorch framework resulting in accuracy of 80.5%.

All Other Projects