

---

Data Scientist with 5+ years of experience driving business impact through advanced analytics and neural networks. Expertise in NLP, computer vision, and scalable model deployment, delivering actionable insights that power data-driven decisions.

## EXPERIENCE

---

### Data Scientist - Claims Fraud Detection Verisk Analytics

Feb 2023 - Current  
Lehi, UT

- **Deepfake and Image Forensics**
  - Designed and deployed a CNN-based deepfake detection system trained on synthetic datasets (DALL-E, Midjourney, Stable Diffusion), achieving 80% precision with <2% false positives, reducing false claim approvals at scale, supported by a Dockerized Gradio inference app.
  - Built a pixel-manipulation detection pipeline for splicing and tampering using deep CNNs and heatmap-based localization to visually highlight manipulated regions, achieving 94% precision (<1% False positive rate) and improving adjuster trust.
- **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 full lifecycle 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**
  - Streamlined dataset curation and error analysis using FiftyOne, improving labeling speed and model robustness across fraud pipelines.
  - Built Dockerized inference apps using Flask and Gradio, enabling scalable internal demos and real-time fraud-screening solutions.

### Data Science Intern Verisk Analytics

June 2022 - August 2022  
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 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 - National Science Foundation(NSF) Indiana University

Jan 2021 - Dec 2022  
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 Infosys Limited

Jan 2019 - Nov 2020  
Hyderabad, India

- Designed, implemented, and optimized 350+ ETL workflows across 150+ tables using Informatica PowerCenter.
- Catalogued financing data for e-contract utilization from OLTP datasets and unstructured data sources to support analytics and ML initiatives.
- Improved ETL pipeline throughput by 200% using SQL optimization, partitioning, and parallelization strategies.

## 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**

## TECHNICAL SKILLS

---

<b>Languages</b>	Python, SQL, HTML
<b>ML Frameworks</b>	Pytorch, Scikit-learn
<b>Computer Vision</b>	OpenCV, OpenVINO Toolkit, Image Classification, Object Detection, Heatmap Localization, FiftyOne
<b>NLP</b>	BERT, RoBERTa, SciBERT, Causal Inference, Dependency Parsing
<b>ML Deployment</b>	Docker, Flask, Gradio, AWS EC2/S3
<b>Database and Cloud</b>	RDBMS (MySQL, SQL Server), ETL (Informatica PowerCenter), Cloud (AWS S3, AWS EC2)
<b>Analytics Tools</b>	Power BI, Microsoft Excel (Advanced), Matplotlib, Seaborn
<b>Statistical Skills</b>	Statistical Modeling, Hypothesis Testing, Predictive Modeling, Exploratory Data Analysis, Data Mining, Parameter Optimization

## PERSONAL PROJECTS

---

**For detailed project descriptions, demos, and source code, please visit my project portfolio:**

[palak-j.github.io/](https://palak-j.github.io/) 

## 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).

