AI-Driven Compliance Risk Intelligence for Life Sciences

Duke MIDS
Master in Interdisciplinary Data Science

Arko Bhattacharya, Jennifer Li, Si Min Loo, Vishesh Gupta, Yirang Liu

Background

What is Pharmaceutical Fraud?

- Illegal practices that distort clinical decision-making.[1,3]
- Includes kickbacks, off-label promotion, fraudulent billing.[2]
- Results in significant legal penalties and reputational damage.[3]

Why is it Difficult to Predict?

- Misconduct is hidden within long, unstructured legal complaints
- Each case is narrative-heavy with inconsistent language
- No standardized dataset across enforcement actions

What Have We Done to Solve This?

Data Sources

Department of Justice (DOJ) Complaints + SEC Filings (8-K, 10-K)

Document Parsing and Text Processing

OCR for DOJ Complaints + Wharton Research Data Services (WRDS) platform for SEC Filings

LLM-Based Extraction of Key Patterns and Compliance Risk Signals

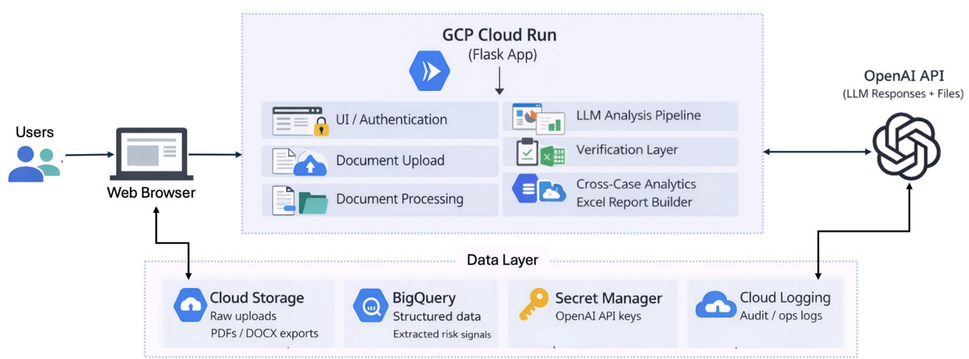
Scalable Platform Deployment on Google Cloud Run

Feature Engineering and Machine Learning Modeling, for Fraud and Non-fraud Cohorts

Deliverables: Compliance Monitoring Platform with Predictive Model

Architecture, Modelling, and Results

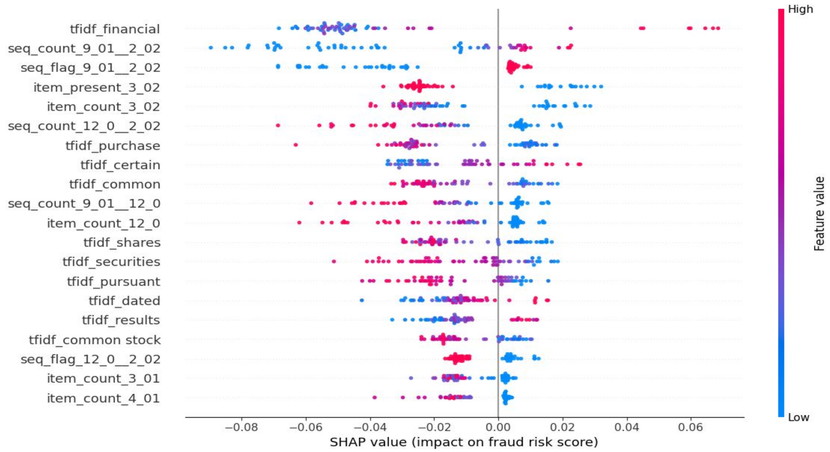
Legal Intelligence Platform Architecture



LLM-powered pipeline converts unstructured legal complaints into structured, company-level risk signals for large-scale compliance analysis.

- **Cases Processed:** 160+ analyzed end-to-end with structured outputs.
- **Verification:** 100% pipeline validation with stakeholder review.
- **Analytics Output:** Cross-case pattern detection with exportable, investigation-ready reports

Predicting Litigation Risk from SEC filings



Analytics outputs from our aforementioned platform are adapted into SEC based behavioral features and scored using a ML pipeline to **estimate litigation risk**.

- **Assumption:** No DOJ complaint ⇒ non-fraud label
- **Best model:** Random Forest | PR-AUC: 0.84 | F1: 0.80
- **Key Signals:** TF-IDF terms, item transitions, item counts

Patterns in SEC filings, including frequency, structure, and language, are strong early indicators of misconduct, with the model accurately identifying most fraud cases on held out data.

Future Work

1. **Expand Dataset & Coverage:** Scale beyond 168 cases by ingesting additional DOJ and SEC data to improve signal diversity and robustness.
2. **Real-Time Modelling:** Ingesting live filings/news streams to move from static analysis to a continuous risk modelling pipeline.
3. **Platform Expansion:** Publishing the predictive risk model on GCP, resulting in an end to end platform for company level compliance analytics.

Acknowledgements

- We gratefully acknowledge GlaxoSmithKline (GSK) for sponsoring this capstone project and for serving as our industry mentor team, providing invaluable domain expertise in life sciences legal and compliance, as well as guidance on real-world workflows and insightful feedback throughout the project.
- We also extend our appreciation to the Duke University MIDS Program, mentor Dr. Jana Schaich Borg, capstone coordinator Dr. Zhongyuan Yu, and all faculty members for their invaluable mentorship and support.

References

1.SciVision Publishing. (2025). A critical review of pharmaceutical industry fraudulent practices.
2.Transparency International. (2016). Corruption in the pharmaceutical sector.
3.U.S. Department of Justice. (2020). Resource guide to the U.S. Foreign Corrupt Practices Act (FCPA).