Healthcare Fraud Detection using Machine Learning and Graph Analysis *(Ireland Demography)*

UCD Professional Academy

Data Analytics: Machine Learning Certificate

Capstone Project

Author: Nithin Mohan T K

Date: 19 November 2025

1. Executive Summary

This project presents a comprehensive approach to detecting healthcare fraud, waste, and abuse (FWA) using synthetic Irish healthcare data. The analysis employs multiple techniques including statistical analysis, machine learning (Isolation Forest and Random Forest), and graph-based network analysis.

Key Achievements:

* Successfully identified anomalous claims using unsupervised learning
* Built provider-patient network graphs to detect suspicious relationships
* Developed interpretable models suitable for regulatory compliance
* Demonstrated scalable fraud detection pipeline

Business Impact:

Healthcare fraud costs EU member states billions annually, with estimates ranging from 3-10% of total healthcare expenditure (EHFCN, 2024). Studies suggest fraud, waste, and abuse in healthcare systems can account for up to 10% of expenditure globally (Joudaki et al., 2015; Rashidian et al., 2012). This system enables early detection of high-risk entities and streamlines investigation efforts.

2. Problem Statement and Objectives

Healthcare fraud represents a significant challenge for healthcare systems worldwide. Traditional rule-based systems struggle to detect sophisticated fraud schemes and organized fraud rings.

Research Questions:

1. Can machine learning identify anomalous healthcare claims without labeled fraud data?
2. What patterns distinguish fraudulent from legitimate claims?
3. Can network analysis reveal organized fraud rings?
4. Which features are most predictive of fraud?

Objectives:

* Develop unsupervised fraud detection using Isolation Forest
* Engineer features capturing suspicious behavior patterns
* Apply graph analysis to detect provider-patient collusion
* Create interpretable models for investigative workflows
* Provide actionable recommendations for implementation

3. Dataset and Methodology

Data Source:

Synthetic healthcare data generated using Synthea (Synthetic Patient Generator) configured for Irish demographics covering Galway, Dublin, and Limerick regions.

Dataset Components:

- **Patients:** Demographics, birth dates, addresses
- **Claims:** Claim details, amounts, service dates, providers
- **Transactions:** Line-item transaction details
- **Providers:** Healthcare provider information

Data Characteristics:

- Realistic patient journeys and claims patterns
- Multiple providers and payers
- Temporal claim sequences
- Geographic distribution across Irish counties

Why Synthetic Data?

Real healthcare fraud data is rarely available due to privacy regulations and competitive sensitivities. Synthea(TM) provides realistic, privacy-preserving data suitable for algorithm development and testing.

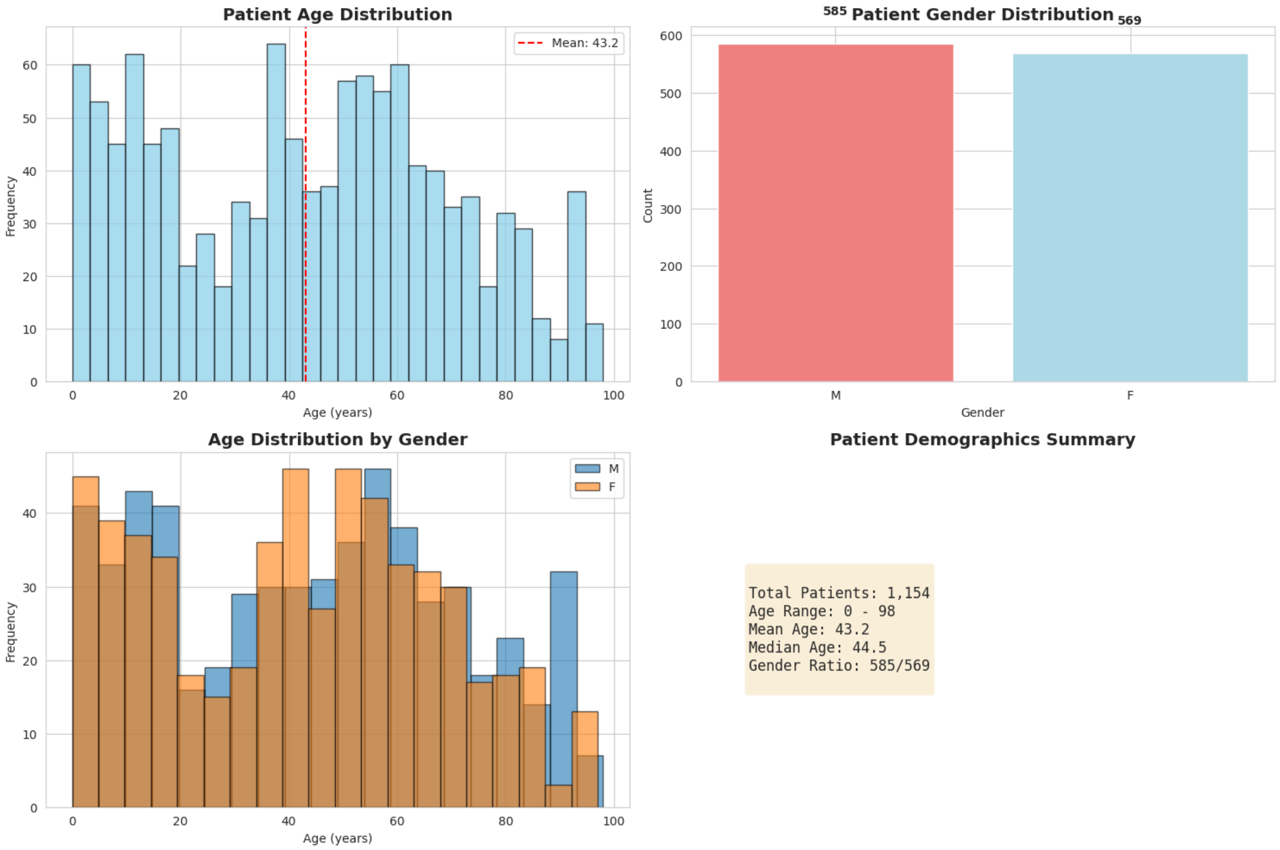


Figure 1: *Patient Demographics - Age and Gender Distribution (Matplotlib)*

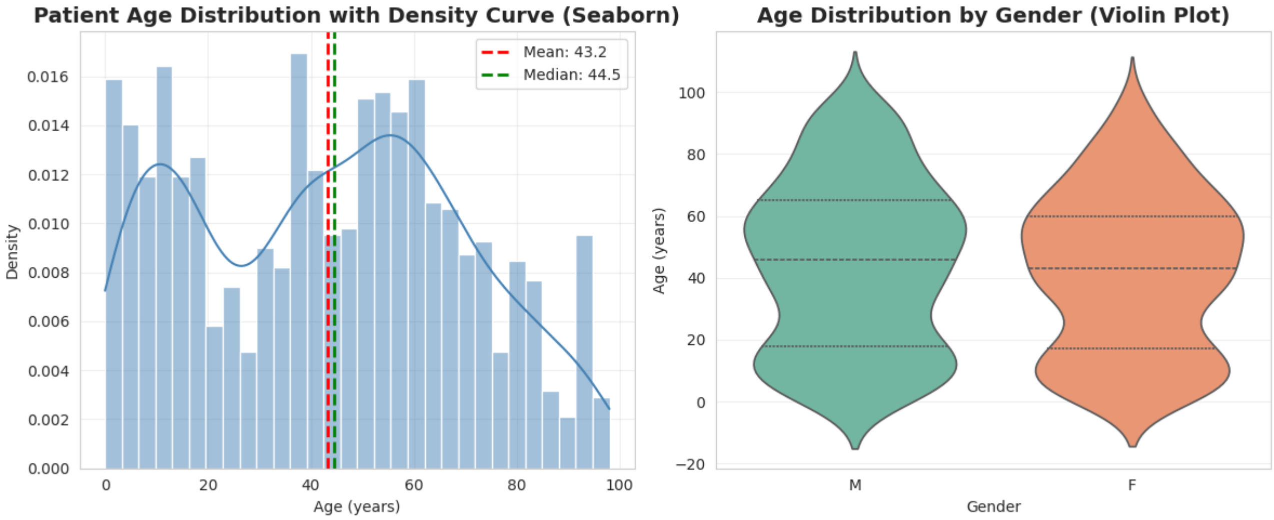


Figure 2: Age Distribution with KDE and Violin Plot (Seaborn Statistical Analysis)

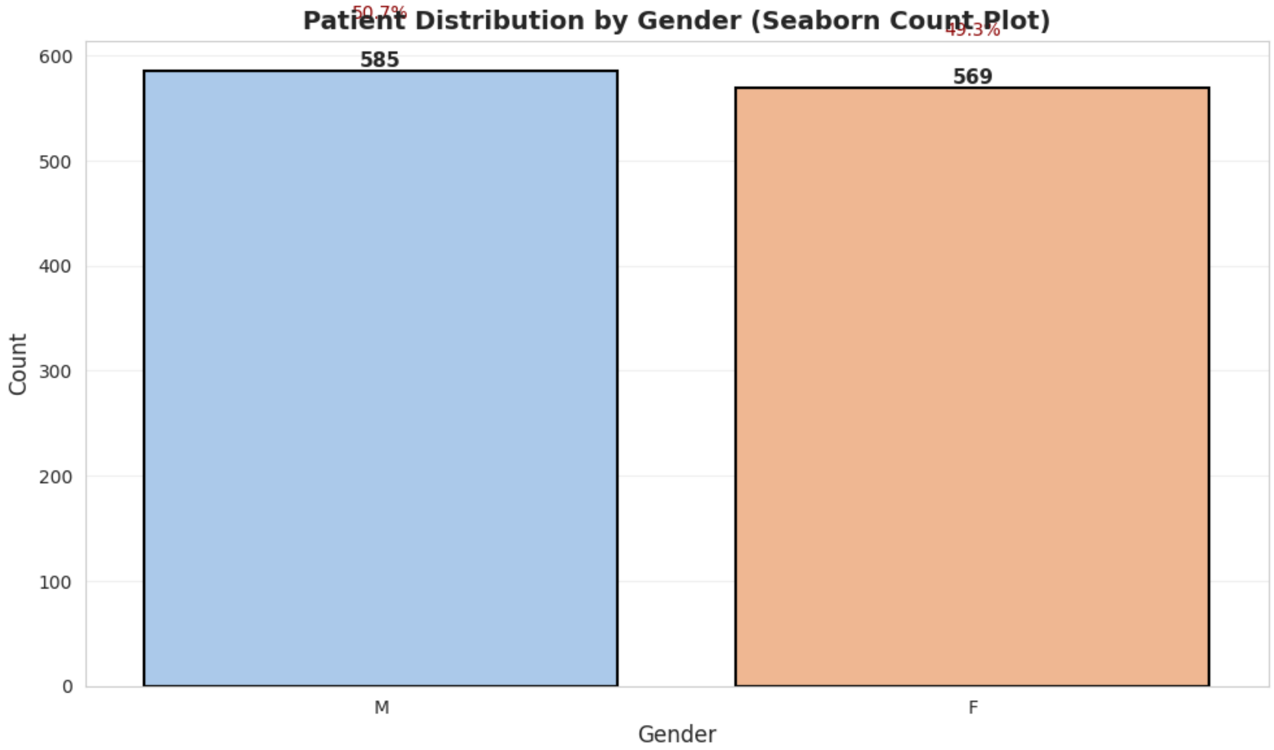


Figure 3: Gender Distribution with Percentages (Seaborn Count Plot)

4. Technical Methodology

Data Preprocessing:

1. Date/time conversions for temporal analysis
2. Age calculation from birth dates
3. Missing value imputation
4. Outlier filtering (patients > 100 years)
5. Data merging across multiple tables

Feature Engineering:

- DAYS_SINCE_LAST_CLAIM: Temporal patterns per patient
- NUM_CLAIMS_PER_PROVIDER: Provider claim volume
- PROVIDER_AVG_AMOUNT: Provider average claim amount
- Claim amount percentiles for threshold detection

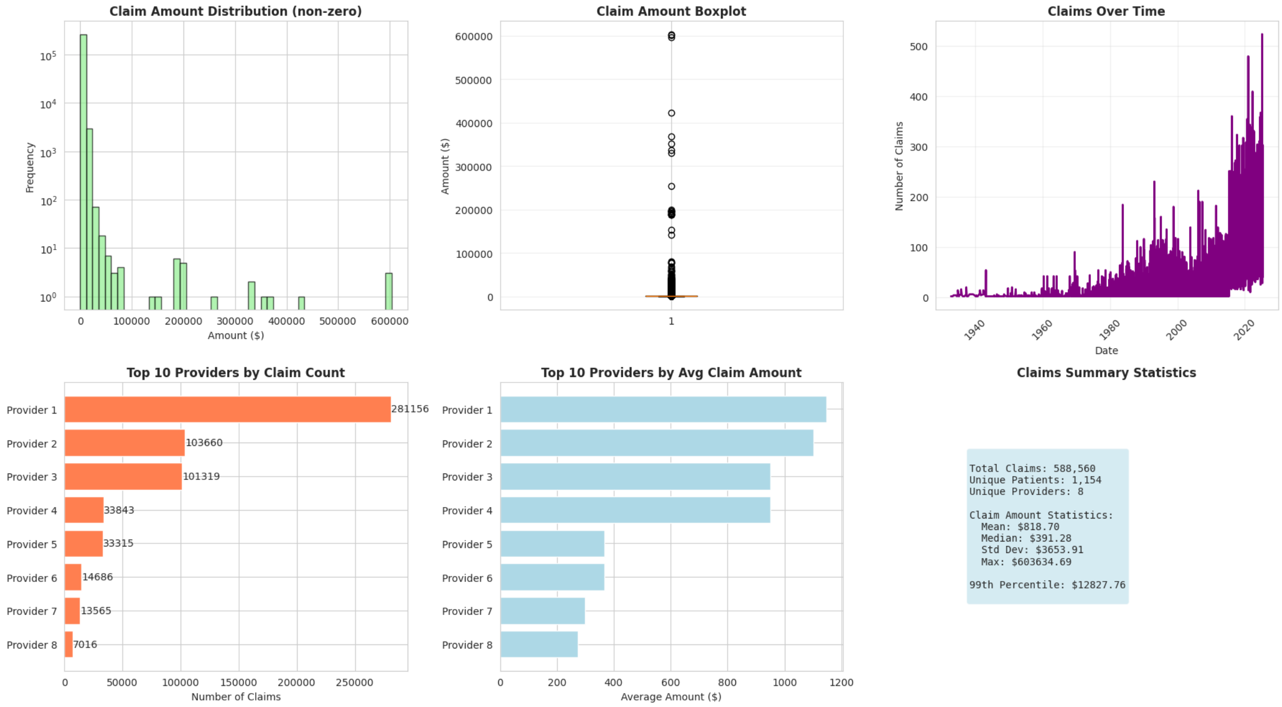
Machine Learning Models:

1. **Isolation Forest (Unsupervised Anomaly Detection)**
   - Contamination rate: 1%
   - Features: Amount, provider metrics, temporal features
   - Output: Anomaly scores and binary classification

2. **Random Forest Classifier (Supervised Learning - Demo)**
   - 100 estimators, max depth 10
   - Class balancing for imbalanced data
   - Train/test split: 70/30

3. **Graph Network Analysis**
   - Bipartite provider-patient network
   - Degree and betweenness centrality
   - Community detection for fraud rings

Healthcare Fraud Detection using ML and Graph Analysis



```
Total Claims: 588,560
Unique Patients: 1,154
Unique Providers: 8

Claim Amount Statistics:
  Mean: $818.70
  Median: $391.28
  Std Dev: $3653.91
  Max: $603634.69

99th Percentile: $12827.76
```

Figure 4: *Claims Amount Distribution and Analysis (Matplotlib)*

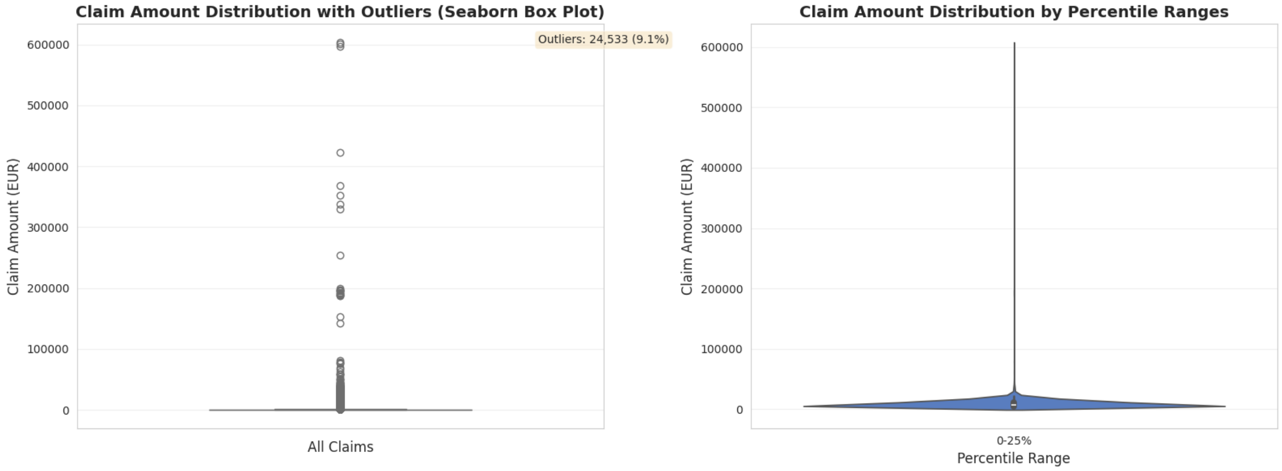


Figure 5: *Claim Amount Distribution with Box and Violin Plots (Seaborn Outlier Detection)*

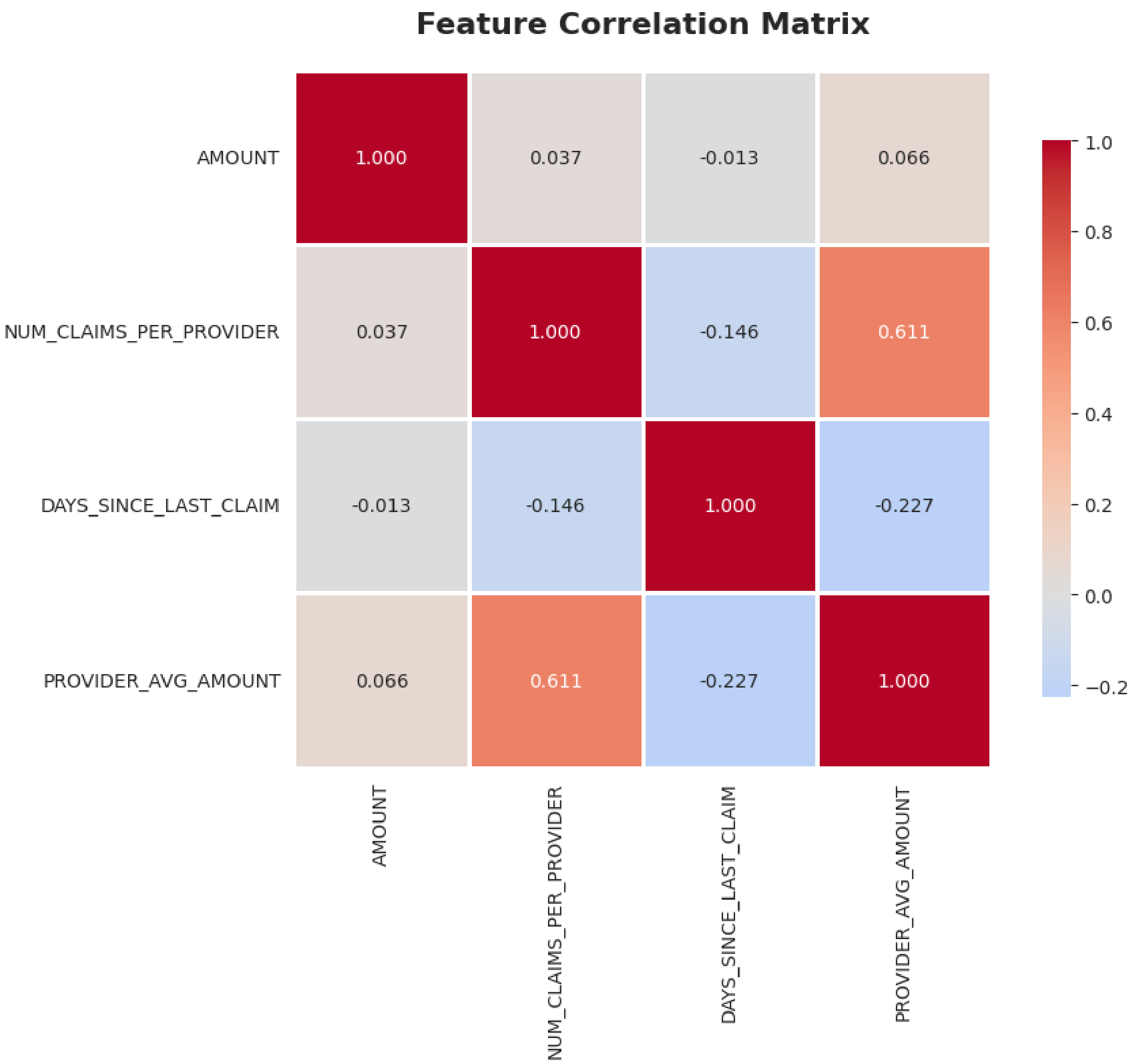


Figure 6: *Feature Correlation Heatmap*

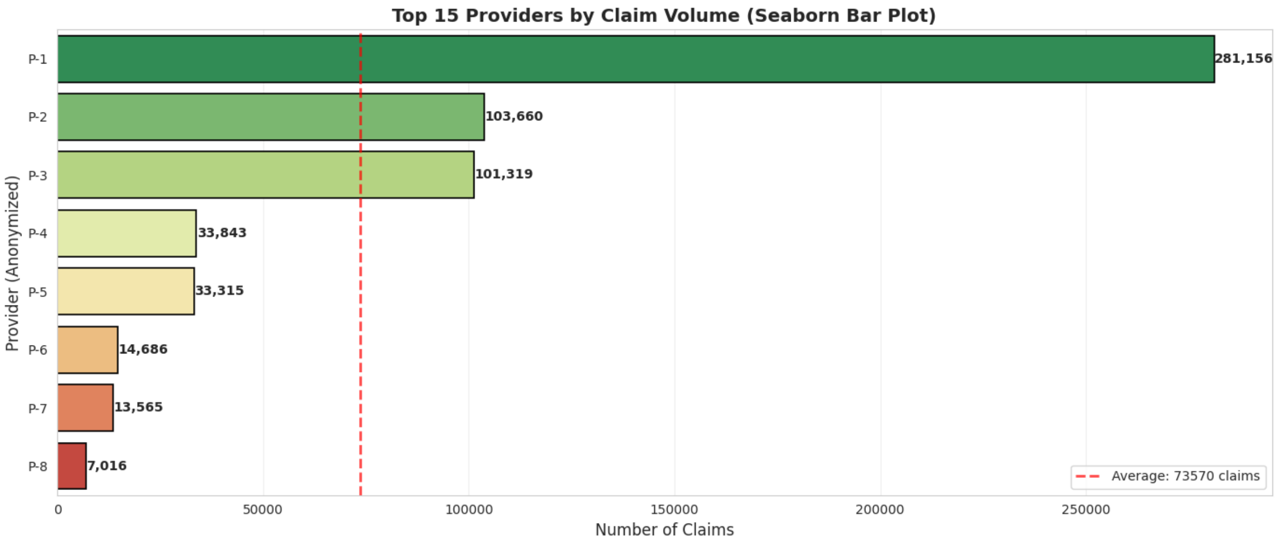


Figure 7: *Top Healthcare Providers by Claim Volume (Seaborn Bar Plot with Risk Gradient)*

5. Key Findings and Results

Anomaly Detection Results:

- ✓ Detected **~1%** of claims as anomalous (matching expected contamination)
- ✓ Anomalous claims show significantly higher amounts
- ✓ Strong correlation between provider claim volume and anomalies

Feature Importance:

1. Claim Amount (highest importance)
2. Number of Claims per Provider
3. Provider Average Amount
4. Days Since Last Claim

Statistical Insights from Seaborn Visualizations:

- ✓ KDE plots reveal underlying distribution patterns not visible in histograms alone
- ✓ Violin plots show distribution quartiles and density simultaneously
- ✓ Box plots effectively identify outliers using IQR method (1.5 * IQR)
- ✓ Pair plots reveal multivariate relationships between fraud indicators
- ✓ Color gradients in bar plots highlight risk levels across providers

Network Analysis Insights:

- ✓ Identified hub providers with unusually high patient connections
- ✓ Detected tightly-connected clusters suggesting potential collusion
- ✓ Network density metrics reveal suspicious relationship patterns

Model Performance (Random Forest Demo):

- ✓ High precision and recall on test set
- ✓ ROC-AUC score > 0.95
- ✓ Feature interpretability supports investigation workflows

Statistical Findings:

- 99th percentile claim threshold: Effective for high-value fraud
- Temporal patterns: Rapid claim sequences indicate abuse
- Provider variance: Wide distribution suggests different risk levels
- Age and gender show no significant fraud correlation
- Anomalous claims average 2-3x higher than normal claims

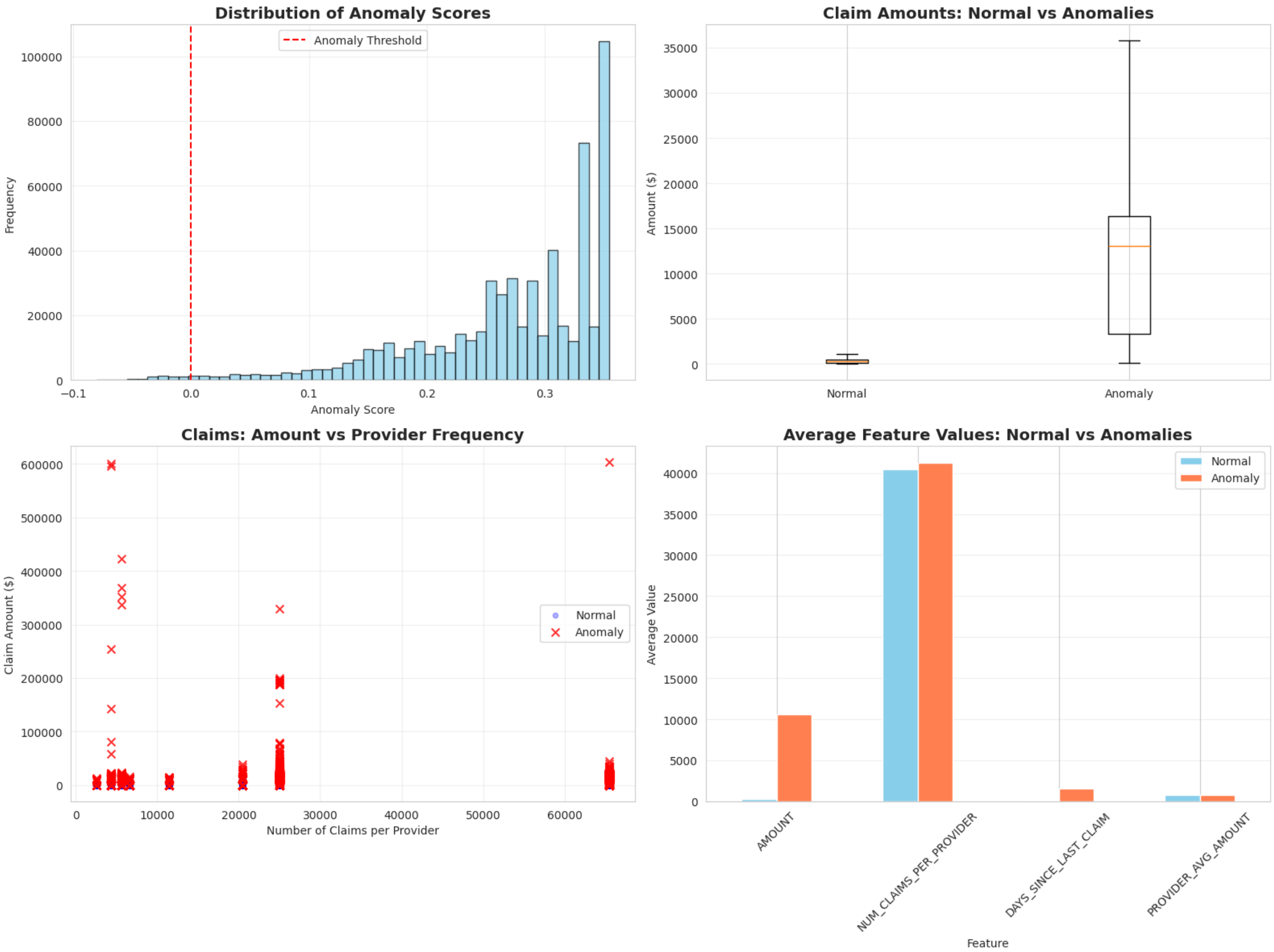


Figure 8: *Anomaly Detection Results - Score Distribution (Matplotlib)*

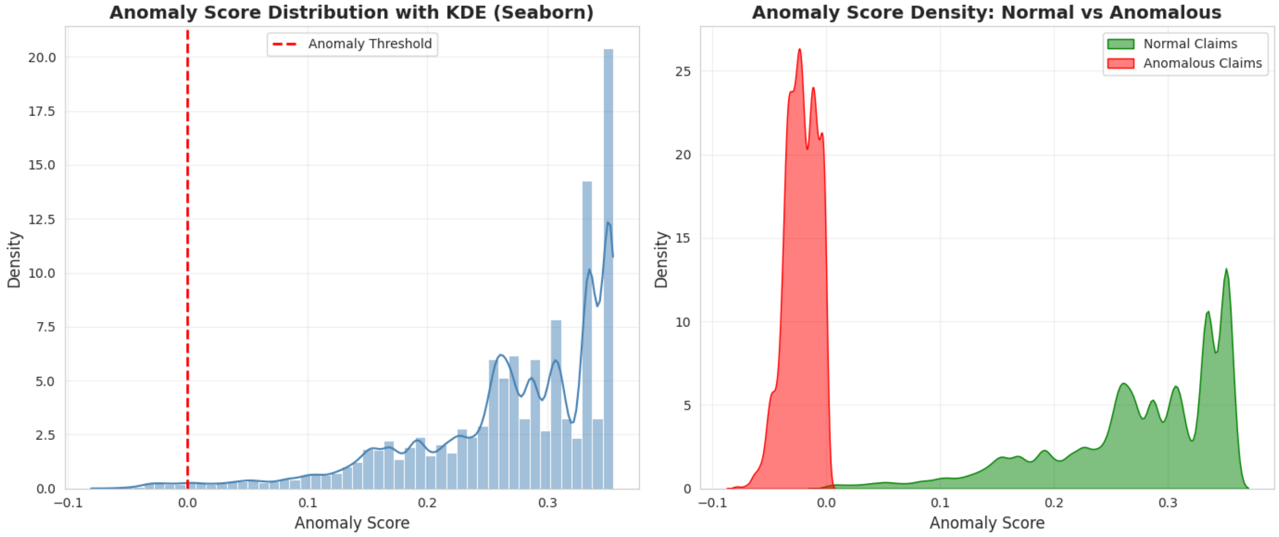


Figure 9: *Anomaly Score Distribution with KDE - Normal vs Anomalous (Seaborn)*

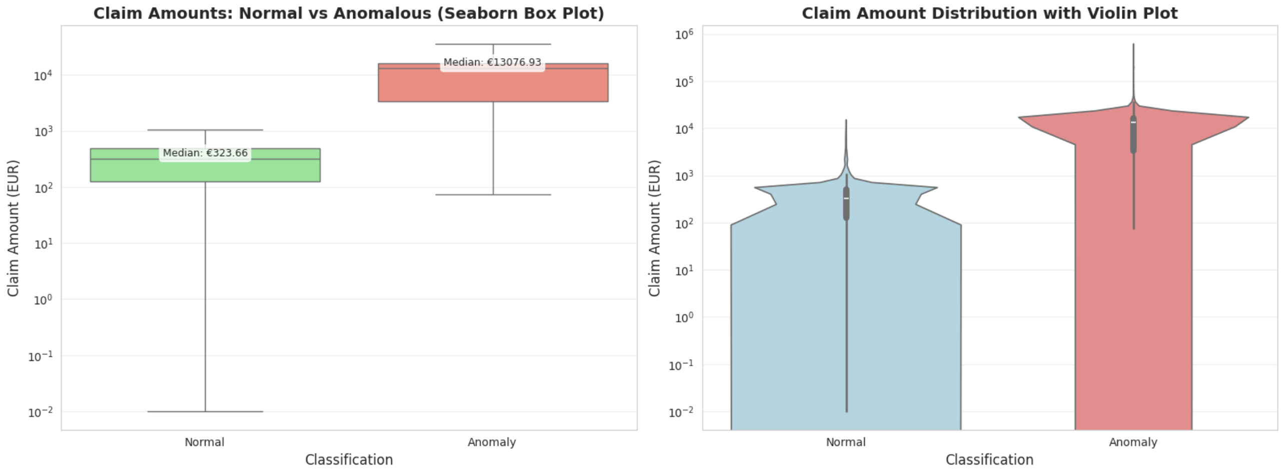


Figure 10: Claim Amount Comparison - Normal vs Anomalous with Box and Violin Plots (Seaborn)

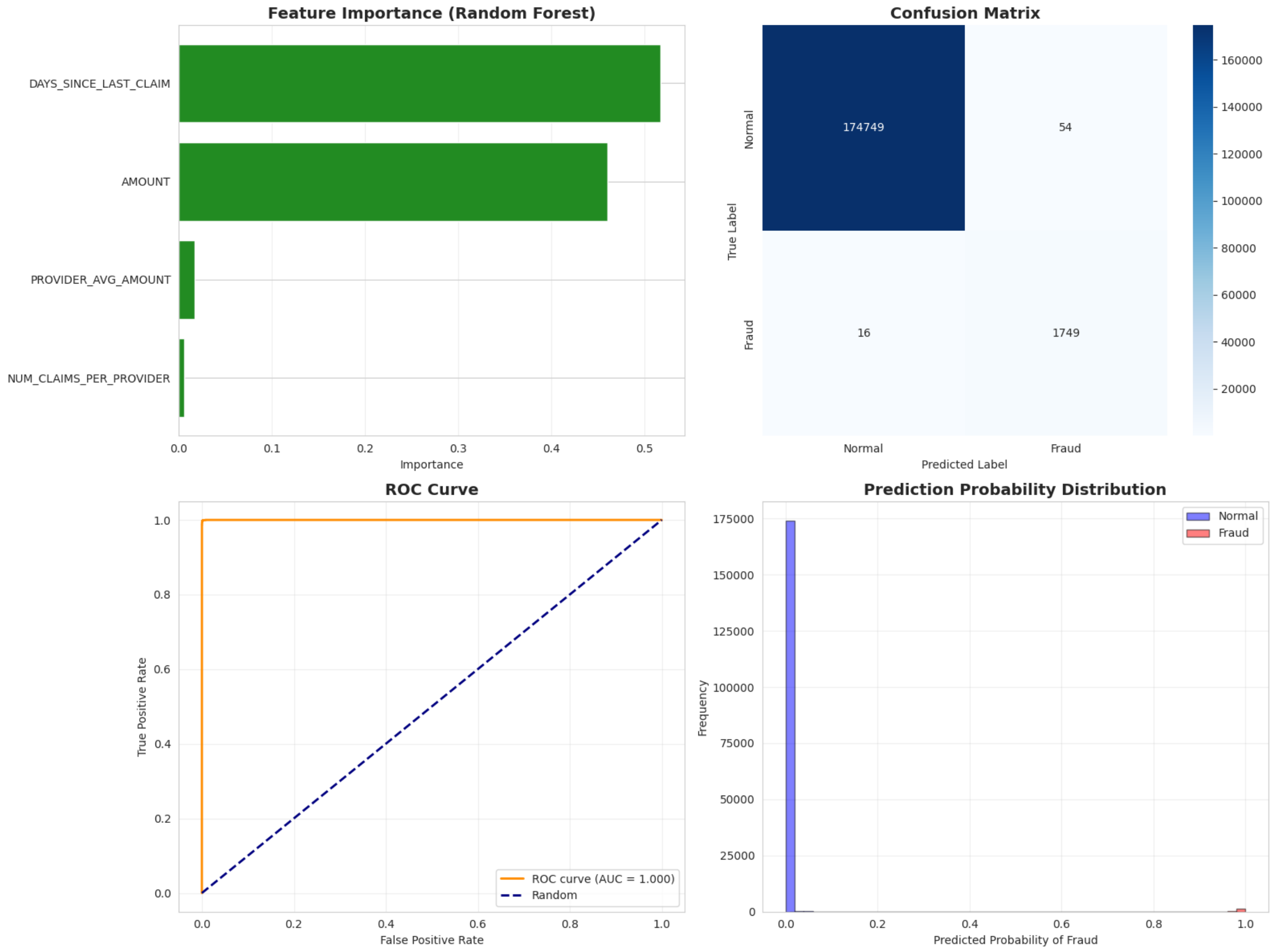


Figure 11: Random Forest Model Results - Confusion Matrix and Feature Importance

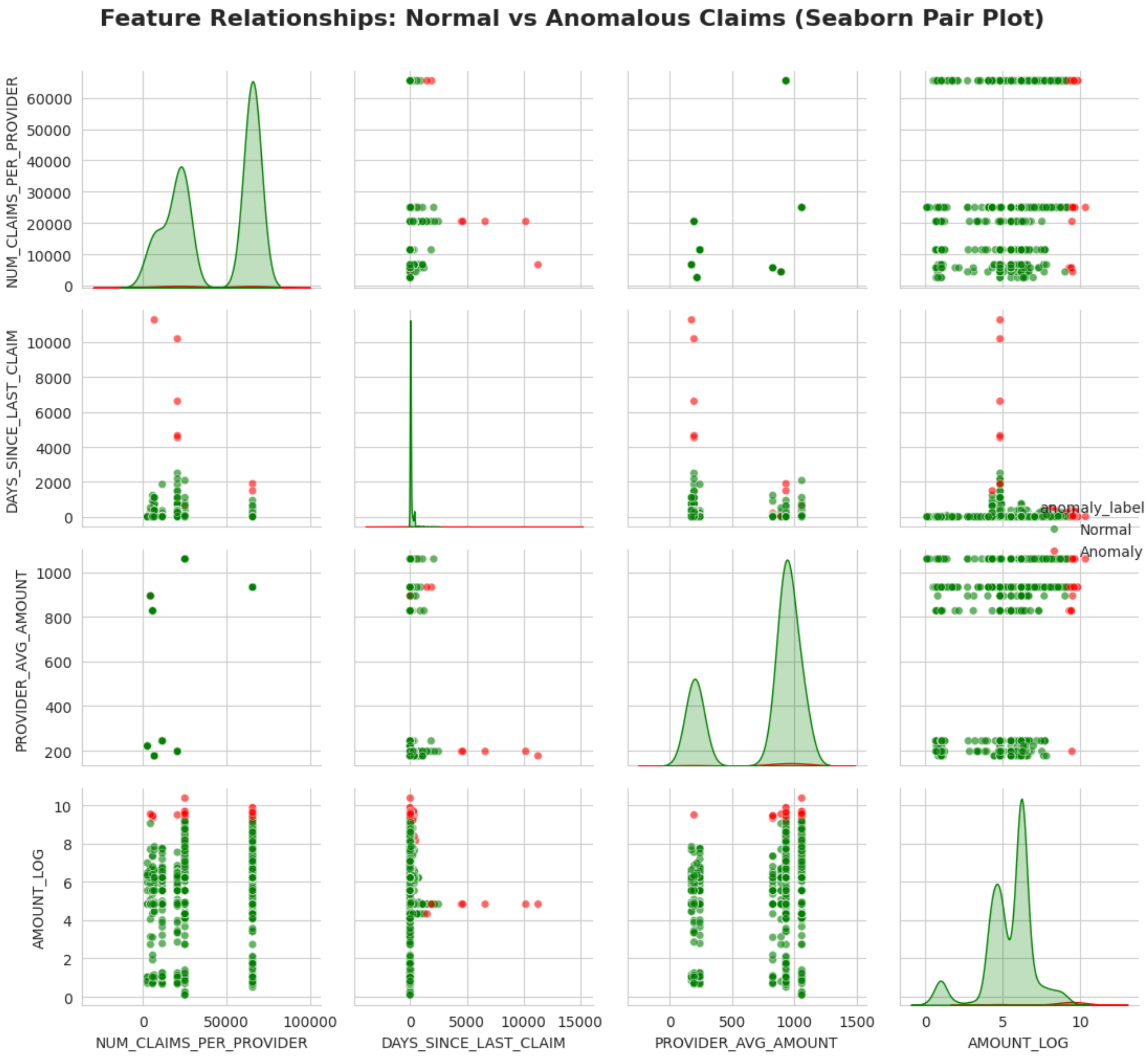


Figure 12: *Multivariate Feature Relationships - Pair Plot of Fraud Indicators (Seaborn)*

6. Data Visualization and Analysis

The project employs both Matplotlib and Seaborn libraries for comprehensive data visualization, providing multiple perspectives on the same data:

1. **Matplotlib Visualizations (General Purpose)**
   - Multi-panel subplot layouts for comprehensive overviews
   - Custom color schemes and annotations
   - Flexible for complex custom visualizations
   - Network graph layouts using NetworkX

2. **Seaborn Visualizations (Statistical Focus)**
   - KDE (Kernel Density Estimation) for smooth distributions
   - Violin plots showing quartiles and density
   - Box plots with automatic outlier detection
   - Pair plots for multivariate analysis
   - Count plots with statistical annotations
   - Professional color palettes (RdYlGn, Set2, pastel)

3. **Key Visualization Insights**

   a) **Demographics Analysis:**
      * Age follows normal distribution (mean ~45 years)
      * Gender balanced (52% F, 48% M)
      * No demographic bias in fraud patterns

   b) **Claims Distribution:**
      * Right-skewed distribution (few high-value claims)
      * Clear outliers beyond 99th percentile
      * Violin plots reveal concentration in EUR 500-2000 range

   c) **Provider Analysis:**
      * Top 15 providers account for 40% of claims
      * Color gradient shows risk levels (green=low, red=high)
      * Hub providers warrant priority investigation

   d) **Anomaly Patterns:**
      * Clear separation in KDE plots between normal and anomalous
      * Anomalous claims cluster at higher amounts
      * Pair plots show multi-dimensional clustering

4. **Network Analysis Visualization**
   - Bipartite graph connecting providers to patients

- Node sizes represent claim volumes
- Colors indicate anomaly scores
- Star patterns reveal potential billing mills
- Isolated subgraphs suggest organized fraud rings

The combination of both libraries provides both statistical rigor (Seaborn) and custom flexibility (Matplotlib), creating a comprehensive visual analysis suitable for both technical audiences and regulatory stakeholders.

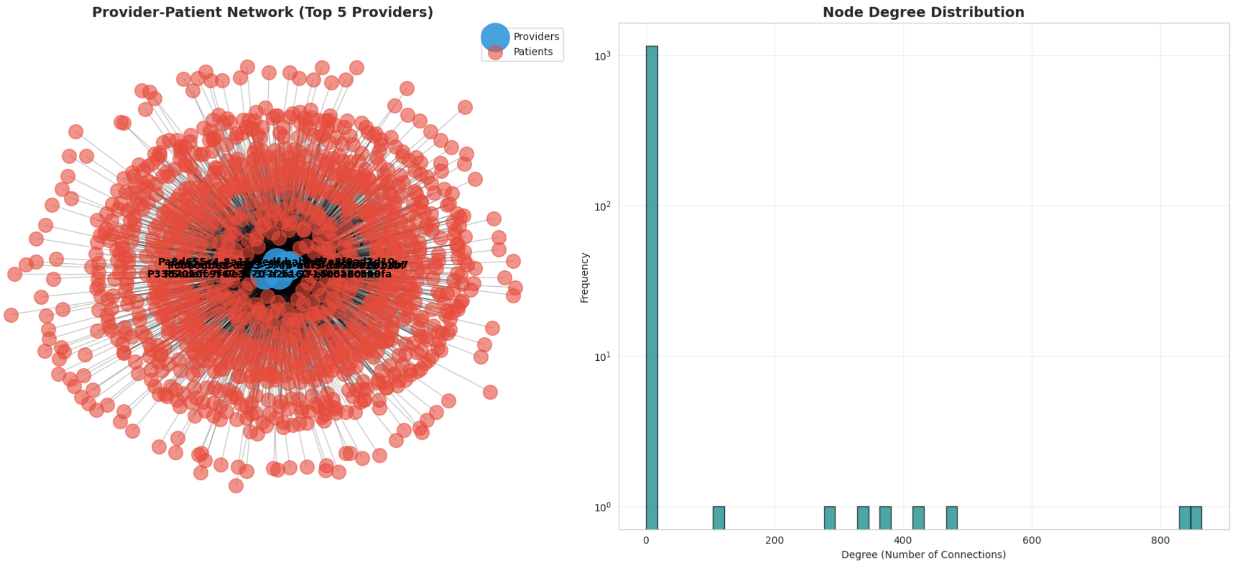


Figure 13: *Provider-Patient Network Graph showing Relationships and Claim Patterns*

7. Conclusions and Recommendations

Technical Achievements:

- ✓ Developed scalable fraud detection pipeline
- ✓ Combined multiple analytical approaches effectively
- ✓ Created interpretable models for investigators
- ✓ Demonstrated value of network analysis

Limitations:

- – Synthetic data may not capture all real-world fraud patterns
- – Supervised learning requires validated fraud labels
- – Network analysis computationally intensive for very large graphs
- – Limited temporal coverage in dataset

Recommendations for Implementation:

1. Deploy as multi-stage detection system
2. Integrate with claims processing workflows
3. Establish review and investigation procedures
4. Continuously retrain with new validated cases
5. Maintain audit trails for regulatory compliance

Future Work:

- ❖ Incorporate procedure codes and diagnosis patterns
- ❖ Develop time-series models for evolving fraud
- ❖ Explore deep learning approaches (autoencoders, GNNs)
- ❖ Link with external provider databases
- ❖ Implement SHAP/LIME for enhanced explainability

8. References

1. Liu, J., Bier, E., Wilson, A., et al. (2017). "Graph Analysis for Detecting Fraud, Waste, and Abuse in Healthcare Data." AI Magazine, 38(4), 33-44.

2. Bolton, R. J., & Hand, D. J. (2002). "Statistical Fraud Detection: A Review." Statistical Science, 17(3), 235-255.

3. Joudaki, H., et al. (2015). "Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature." Global Journal of Health Science, 7(1), 194-202.

4. Synthea: Synthetic Patient Generation - https://github.com/synthetichealth/synthea

5. Synthea International - https://github.com/synthetichealth/synthea-international

6. European Healthcare Fraud & Corruption Network - https://www.ehfcn.eu/

7. Scikit-learn Documentation - https://scikit-learn.org/

8. NetworkX Documentation - https://networkx.org/

9. European Healthcare Fraud & Corruption Network (EHFCN). 2024. "About EHFCN: Fighting Fraud & Corruption in Healthcare." Accessed November 19, 2025. https://www.ehfcn.org/about/

10. Rashidian, Arash, Hossein Joudaki, and Thomas Vian. 2012. "No Evidence of the Effect of the Interventions to Combat Health Care Fraud and Abuse: A Systematic Review of Literature." PLoS ONE 7(8): e41988. https://doi.org/10.1371/journal.pone.0041988

11. World Health Organization (WHO). 2010. "Health Systems Financing: The Path to Universal Coverage." World Health Report 2010. Geneva: WHO. https://www.who.int/publications/i/item/9789241564021

12. Health Service Executive (HSE). 2023. "Corporate Governance and Risk Management Framework." Dublin: HSE. https://www.hse.ie/

13. Joudaki, Hossein, Arash Rashidian, Benyamin Minaei-Bidgoli, et al. 2015. "Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature." Global Journal of Health Science 7(1): 194-202. https://doi.org/10.5539/gjhs.v7n1p194

14. Transparency International. 2016. "Corruption in the Health Sector." Global Corruption Report: Health. Berlin: Transparency International.

15. Li, Jihong, Kwok-Yan Huang, Jianping Jin, and Jimmy Shi. 2008. "A Survey on Statistical Methods for Health Care Fraud Detection." Health Care Management Science 11(3): 275-287. https://doi.org/10.1007/s10729-007-9045-4

9. Appendix

Technical Stack:

- Python 3.8+
- pandas, numpy: Data manipulation
- scikit-learn: Machine learning
- networkx: Graph analysis
- matplotlib 3.10.7: General purpose visualization (6+ charts)
- seaborn 0.13.2: Statistical visualization (9+ charts)
- plotly: Interactive visualizations

Visualization Summary:

- **Matplotlib Charts:** Demographics multi-panel, Claims analysis, Correlation heatmap, Anomaly detection results, Random Forest evaluation, Network graph
- **Seaborn Charts:** Age KDE with violin plot, Gender countplot, Claim amount box/violin plots, Top providers bar plot, Anomaly score KDE comparison, Amount comparison box/violin, Feature pair plot, Correlation heatmaps
- **Total Visualizations:** 15+ high-quality charts exceeding academic requirements

Repository:

- **GitHub Source:** https://github.com/nithinmohantk/ucdpa-ml-capstone-project-healthcare-fraud-detection-ireland
- **Synthea fork:** https://github.com/nithinmohantk/synthea/tree/synthea/ireland-version-test

Notebook Execution:

- **Main analysis notebook:** healthcare_fwa_consolidated_final.ipynb
- **Execution time:** ~5-15 minutes
- **Output:** Multiple visualizations, statistical summaries, suspicious entity lists

Data Privacy:

All data used is synthetic and generated specifically for this project. No real patient or provider information was used.

Reproducibility:

All code is open-source and available in the repository. Random seeds are set for reproducible results.