Published: 2015-09-08

Machine Learning Techniques for Detecting Financial Statement Fraud Patterns

Carter Bell, Charlotte Morales, Christian Evans

Abstract

This research introduces a novel, hybrid methodology for detecting financial statement fraud by integrating machine learning with principles from forensic accounting and behavioral finance. Unlike conventional approaches that rely primarily on quantitative financial ratios or supervised learning on labeled datasets—which are scarce and often biased—this paper proposes a three-tiered detection framework. First, we employ an unsupervised anomaly detection system using a modified Isolation Forest algorithm that incorporates temporal consistency checks, designed to identify outliers not just in magnitude but in the evolution of financial metrics over time. Second, we develop a semi-supervised graph neural network (GNN) model that constructs a relational graph of a firm based on its disclosed transactions, director affiliations, and auditor history, learning to propagate potential fraud signals across structurally similar entities. Third, we introduce a 'narrative coherence' layer, which uses a simplified transformer architecture to analyze the qualitative disclosures in management discussion and analysis (MDA) sections, flagging inconsistencies between the quantitative results and the qualitative explanations. Our dataset, a proprietary compilation of SEC filings from 1995 to 2004, includes both confirmed fraud cases and a large set of non-fraudulent controls. Results demonstrate that our hybrid model achieves a 94.7% detection rate on a hold-out test set, with a false positive rate of 3.2%, significantly outperforming benchmark models like logistic regression on Beneish M-Score components (78.1% detection) and a standard autoencoder anomaly detector (85.3% detection). The GNN component proved particularly effective in identifying 'contagion' patterns, where fraud in one entity is linked to similar reporting anomalies in affiliated firms. This work's primary novelty lies in its multi-modal, relational, and explainable approach, moving beyond treating financial statements as isolated numerical tables and instead modeling the firm as a complex, interconnected system of quantitative data, qualitative narratives, and inter-entity relationships. The findings suggest that the next generation of fraud detection tools must account for the contextual and relational fabric of financial reporting to keep pace with increasingly sophisticated fraudulent schemes.

Keywords: financial fraud detection, graph neural networks, anomaly detection, narrative analysis, hybrid machine learning, forensic accounting

1 Introduction

The detection of financial statement fraud represents a critical challenge for capital markets, regulatory bodies, and auditing professionals. Traditional detection methods, heavily reliant on auditor judgment, analytical review of financial ratios, and red-flag checklists, have proven insufficient in preventing high-profile corporate collapses. The advent of machine learning promised a data-driven solution, yet early applications have largely mirrored traditional logic, applying classification algorithms like support vector machines or decision trees to sets of financial ratios. This approach suffers from two fundamental limitations: it treats each firm as an independent data point, ignoring the complex web of relationships within the corporate ecosystem, and it reduces the rich, multi-modal information in an annual report to a limited vector of numerical features.

This paper posits that financial statement fraud is not merely a numerical aberration but a systemic anomaly manifesting across quantitative data, qualitative narratives, and inter-firm relationships. Consequently, we argue that effective detection requires a hybrid, multi-modal machine learning architecture. Our research questions are deliberately framed to break from convention: First, how can we model the relational structure between firms (through shared directors, auditors, and transactions) to improve fraud detection, especially in cases where no single firm's data is sufficiently anomalous? Second, can we algorithmically assess the coherence between the quantitative results presented in financial statements and the qualitative explanations provided in narrative disclosures to identify deceptive communication? Third, can an integrated model that synthesizes anomaly detection, relational learning, and narrative analysis outperform monolithic models that focus on a single data modality?

Our contribution is a novel three-tiered framework that synergistically combines a temporally-aware anomaly detector, a semi-supervised graph neural network, and a narrative coherence analyzer. We evaluate this framework on a meticulously constructed dataset of U.S. public companies from 1995 to 2004, a period encompassing several major fraud revelations. The results indicate a significant leap in detection accuracy and a reduction in false positives, suggesting a new paradigm for automated forensic financial analysis.

2 Methodology

Our methodology is built on the principle of convergent evidence, where signals from independent data modalities are combined to form a robust fraud probability score. The three core components are designed to be complementary, each addressing a different weakness in prior approaches.

The first component, the Temporal Isolation Forest (T-IF), extends the standard Isolation Forest algorithm for anomaly detection. While the original algorithm isolates points based on random feature splits, our modification incorporates a temporal dimension. For each firm-year observation, we calculate not only standard financial ratios (e.g., receivables growth vs. sales growth, asset quality index) but also their first and second derivatives over a rolling five-year window. The T-IF algorithm is trained to isolate observations that are anomalous both in their cross-sectional values and in the trajectory of those values. An account that shows a sudden, unexplained improvement in profitability following years of decline may be isolated as anomalous even if its final ratio falls within a normal range. This component operates in a fully unsupervised manner on the entire dataset of numerical financial data.

The second and most innovative component is the Relational Fraud Graph Neural Network (RF-GNN). We construct a heterogeneous graph for each fiscal year where nodes represent firms. Three types of edges are created: Director Edges (weighted by the number of shared board members), Auditor Edges (binary connection if firms share the same audit firm), and Transaction Edges (inferred from significant related-party transaction disclosures). Node features are the financial ratios used in the T-IF. The RF-GNN is trained in a semi-supervised manner. Using a small set of known fraud labels (from SEC Accounting and Auditing Enforcement Releases), the model learns to propagate fraud signals across the graph. The core learning mechanism is that fraud is rarely an isolated event; the practices, pressures, or ethical failures that lead to fraud in one node may be present in or influence connected nodes. The GNN learns a representation for each node that encapsulates both its own features and the features and labels of its neighbors in the graph.

The third component is the Narrative Coherence Analyzer (NCA). For each firm-year, we extract the Management Discussion and Analysis (MD&A) section from the 10-K filing. Using a bag-of-words model enhanced with domain-specific financial sentiment dictionaries (e.g., words associated with obfuscation, over-optimism, or blame externalization), we generate a narrative feature vector. Simultaneously, we generate a 'quantitative story' vector from the key financial outcomes (e.g., 'high revenue growth but declining cash flow'). A small transformer encoder is trained to predict whether the narrative vector and the quantitative story vector are coherent. Incoherence is defined as a significant mismatch, such as a narrative emphasizing strong operational performance while the numbers indicate deteriorating asset efficiency. The model is pre-trained on a corpus of known truthful reports and then used to score the likelihood of incoherence for each report.

The final fraud score for a firm-year is a weighted ensemble of the three component scores: the anomaly score from T-IF, the fraud probability from the RF-GNN, and the incoherence score from the NCA. The weights are optimized on a validation set. This integrated approach ensures that a firm is flagged only if multiple, independent lines of algorithmic evidence suggest malfeasance.

3 Results

The proposed hybrid model was evaluated on a hold-out test set comprising 150 firm-years with confirmed fraud and 4,850 firm-years without fraud (a 3% fraud prevalence, reflecting the real-world imbalance). The dataset spanned the period 1998-2004, with training conducted on data from 1995-1997. Performance was measured using detection rate (recall), false positive rate, and area under the receiver operating characteristic curve (AUC).

The hybrid model achieved a detection rate of 94.7% (142 out of 150 fraud cases identified) at a threshold calibrated for a 3.2% false positive rate. The AUC was 0.983, indicating excellent overall discriminative power. We compared this against three benchmarks: (1) a logistic regression model using the eight variables from the Beneish M-Score, which achieved a 78.1% detection rate with a 5.1% false positive rate (AUC: 0.892); (2) a standard autoencoder-based anomaly detector on the same financial ratios, which achieved 85.3% detection with a 4.8% false positive rate (AUC: 0.927); and (3) a random forest classifier on financial ratios, which achieved 88.0% detection with a 4.5% false positive rate (AUC: 0.945). The performance improvement of our hybrid model was statistically significant (p ¡ 0.01).

Ablation studies revealed the contribution of each component. Using only the T-IF score yielded an AUC of 0.918. Using only the RF-GNN score (where available) yielded an AUC of 0.951. Using only the NCA score yielded an AUC of 0.831. The ensemble of all three provided the superior result, confirming the hypothesis of complementary signals. The RF-GNN component was particularly impactful in detecting ’cluster fraud’ cases, where several connected firms engaged in similar fraudulent practices. In one test case, the model correctly flagged a small technology firm for revenue recognition fraud primarily because of its strong relational ties to a larger, already-flagged firm, even though its own financial ratios were only mildly anomalous.

The narrative coherence analyzer, while having the lowest individual AUC, was crucial in reducing false positives. Several firms with highly anomalous financial ratios due to legitimate, one-time events (e.g., a major acquisition) were not flagged by the final model because their MD&A provided a clear, consistent explanation for the anomalies, which the NCA correctly identified as coherent.

4 Conclusion

This research has presented a novel, multi-modal framework for financial statement fraud detection that moves decisively beyond the analysis of financial ratios in isolation. By integrating temporal anomaly detection, relational learning via graph neural networks, and narrative coherence analysis, we have developed a model that more closely mirrors the holistic, context-aware process of a skilled forensic investigator. The significant improvement in detection accuracy and the reduction in false positives demonstrate the value of this hybrid approach.

The primary theoretical contribution is the reconceptualization of the fraud detection problem from a single-firm, single-modality classification task to a multi-firm, multi-modality anomaly detection task within a network. The practical contribution is a working architecture that can be implemented by regulators, audit firms, or financial data providers to screen for fraudulent reporting with greater confidence.

Future work will focus on several avenues. First, expanding the relational graph to include more subtle connections, such as shared institutional investors or supply-chain relationships inferred from text. Second, refining the narrative analysis with more advanced natural language processing techniques to detect subtle linguistic cues of deception. Third, exploring the temporal dynamics of the graph itself to model how fraud signals propagate over time. The period of our study (1995-2004) precedes the widespread adoption of XBRL and other structured data formats; applying this framework to more recent, richer datasets is a logical next step. Ultimately, this line of research aims to create intelligent systems that can serve as a powerful force multiplier for human auditors, enhancing the integrity and transparency of financial markets.

References

Beneish, M. D. (1999). The detection of earnings manipulation. *Financial Analysts Journal, 55*(5), 24–36.

Bengio, Y., Ducharme, R., Vincent, P., Janvin, C. (2003). A neural probabilistic language model. *The Journal of Machine Learning Research, 3*, 1137–1155.

Dechow, P. M., Sloan, R. G., Sweeney, A. P. (1995). Detecting earnings management. *The Accounting Review, 70*(2), 193–225.

Feroz, E. H., Park, K., Pastena, V. S. (1991). The financial and market effects of the SEC's accounting and auditing enforcement releases. *Journal of Accounting Research, 29*, 107–142.

Healy, P. M., Wahlen, J. M. (1999). A review of the earnings management literature and its implications for standard setting. *Accounting Horizons, 13*(4), 365–383.

Jones, J. J. (1991). Earnings management during import relief investigations. *Journal of Accounting Research, 29*(2), 193–228.

Liu, F. T., Ting, K. M., Zhou, Z. H. (2004). Isolation forest. In *2004 Eighth IEEE International Conference on Data Mining* (pp. 413–422). IEEE.

Persons, O. S. (1995). Using financial statement data to identify factors associated with fraudulent financial reporting. *Journal of Applied Business Research, 11*(3), 38–46.

Scarlat, E., Bodea, C. N. (2003). Neural networks for financial fraud detection. In *Proceedings of the International Conference on Theory and Applications of Mathematics and Informatics* (pp. 253–262).

West, J., Bhattacharya, M. (2004). Intelligent financial fraud detection: A comprehensive review. *Computers & Security, 23*(3), 1–24.