Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Comparative Evaluation of Automated Data Consistency Detection Methods in Corporate SEC Filings

Hanfei Zhang[1], *Wangwang Shi*[1.2]
[1] *Law, Emory University School of Law, Atlanta, GA, USA*
[1.2] *Softerware Engineering, University of Science and Technology of Chinay, He fei, China*
DOI: 10.69987/JACS.2026.60401

Keywords

data consistency, SEC filings, anomaly detection, regulatory compliance

Abstract

Corporate disclosure documents submitted to the U.S. Securities and Exchange Commission contain extensive financial data that requires verification across multiple periods and document sections. Manual verification of data consistency in SEC filings presents significant challenges due to document length and complexity. This research conducts a systematic comparative evaluation of three automated detection approaches: rule-based logic verification, statistical analysis-based anomaly identification, and neural network-based semantic contradiction discovery. The study examines 248 SEC reports from publicly traded companies, identifying 1,847 manually annotated data inconsistencies across numerical discrepancies, logical conflicts, and semantic contradictions. Performance evaluation demonstrates that rule-based methods achieve 94.3% precision with 78.6% recall for numerical inconsistencies, statistical approaches attain 87.9% precision with 82.4% recall for time-series anomalies, and neural methods reach 91.7% precision with 85.2% recall for semantic contradictions. Cross-document validation capabilities reveal substantial performance variations, with detection accuracy declining by 23-31% when analyzing inter-report relationships. The findings provide empirical evidence supporting the deployment of hybrid detection frameworks combining complementary strengths of each methodology. These automated tools offer practical value for regulatory oversight and corporate compliance management.

1. Introduction

1.1. Background and Motivation of Corporate Disclosure Monitoring

Corporate financial disclosure serves as a fundamental mechanism for market transparency within U.S. capital markets. Listed companies submit Form 10-K annual reports and Form 10-Q quarterly reports to the Securities and Exchange Commission, containing comprehensive financial statements and management analysis. These documents frequently exceed 300 pages and contain thousands of data points requiring internal consistency across multiple periods and sections [1]. Research examining SEC enforcement actions between 2020-2023 indicates that disclosure violations constitute approximately 34% of enforcement proceedings, with data inconsistencies representing a significant category [2]. Manual review processes face resource constraints, limiting comprehensive examination of all submitted reports [3]. Data inconsistencies emerge through transcription errors, calculation mistakes, misalignment between narratives and tables, and temporal disconnects when updating document portions. Automated detection methodologies offer solutions by enabling systematic document scanning and identifying anomalies requiring review.

1.2. Challenges in Manual Data Consistency Verification

Traditional manual verification exhibits fundamental limitations when applied to contemporary SEC documents. Average annual reports contain 47,000-52,000 words and 120-150 financial tables. Human reviewers must track numerical consistency across dozens of cross-references and verify mathematical relationships [4]. Studies document detection rates for intentionally inserted inconsistencies ranging between 62-74% under controlled conditions [5]. Filing deadlines

create compressed review cycles, with accelerated filers submitting Form 10-K within 75 days and large accelerated filers within 60 days. Distributed document preparation compounds coordination challenges, with different departments responsible for various sections. Document revisions may inadvertently create new inconsistencies when changes address reviewer comments. Inconsistency types vary from numerical mismatches to logical contradictions and semantic conflicts, necessitating different detection strategies [6].

1.3. Research Objectives and Contributions

This research addresses the need for systematic evaluation of automated consistency detection methodologies applicable to SEC disclosure documents [7]. The primary objective involves comparative assessment of three distinct technical approaches examining their effectiveness across different inconsistency categories and cross-document validation scenarios. The study contributes empirical performance benchmarks derived from analysis of real SEC filings, providing practical guidance for selecting and deploying automated detection tools. Specifically, the research investigates detection accuracy for three inconsistency categories: numerical mismatches where figures fail to reconcile across related disclosures, logical conflicts where stated policies contradict their described application, and semantic contradictions where narrative descriptions conflict with tabular presentations [8]. Performance evaluation encompasses precision metrics measuring false positive rates, recall metrics assessing completeness of inconsistency detection, and processing efficiency indicators relevant to practical deployment. The research establishes performance baselines using a substantial corpus of real SEC filings with expert-annotated inconsistencies, develops an evaluation framework accommodating the heterogeneous nature of disclosure inconsistencies, and offers decision support for organizations considering automated detection tool deployment [9].

2. Literature Review and Problem Analysis

2.1. Current Practices in SEC Filings Review and Regulatory Requirements

The Securities and Exchange Commission employs selective review of corporate filings, examining subsets based on risk assessment. The Division of Corporation Finance analyzes financial statements, management discussion, and risk factors for regulatory compliance. Reviewers issue comment letters identifying deficiencies, with companies required to respond and potentially amend filings. Current procedures rely on manual examination by staff accountants and attorneys with sector-specific knowledge. Comment letter analysis of 2,847 letters during 2021-2023 found that 18.6% addressed inconsistencies between sections, 12.3% involved reconciliation issues, and 9.7% concerned narrative-quantitative alignment [10]. Resource limitations motivate interest in automated screening tools that could enhance review efficiency and enable more comprehensive coverage of submitted filings.

2.2. Existing Approaches for Financial Data Validation and Their Limitations

Traditional validation relies on rule-based systems implementing predefined checks verifying mathematical relationships and accounting principles. Commercial disclosure software incorporates built-in validation rules. XBRL taxonomies include calculation linkbases defining required mathematical relationships. Analysis of 2018-2022 XBRL submissions found 23.8% contained calculation inconsistencies, 15.4% exhibited dimension errors, and 8.9% showed logical violations [11]. Rule-based limitations stem from inability to detect unanticipated inconsistencies. Statistical methods apply techniques from fraud detection, examining time series patterns and ratio analysis. Machine learning classifiers trained on historical data learn distinguishing patterns, with ensemble methods demonstrating improved accuracy. Statistical methods face threshold-setting challenges balancing sensitivity against false positives. Natural language processing techniques have emerged for textual analysis, with transformer-based models demonstrating capabilities in sentiment classification and information extraction [12]. Adaptation to consistency detection remains an active research area with limited production deployment.

2.3. Types of Data Inconsistencies in Corporate Disclosure Documents

Corporate disclosure inconsistencies manifest in multiple forms requiring distinct detection approaches [13]. Numerical inconsistencies occur when figures representing the same fact differ across locations, such as revenue stated differently in management discussion and financial statements, or totals disagreeing with component sums. These result from transcription errors, calculation mistakes, or incomplete updates. Temporal inconsistencies arise when reported figures violate expected time-period relationships, such as substantial accounts receivable increases without corresponding revenue growth, or cumulative annual figures failing to reconcile with quarterly totals. Logical inconsistencies involve contradictions between stated policies and described application, such as companies describing rigorous controls while reporting material weaknesses [14]. Semantic inconsistencies occur when narratives conflict with quantitative data, such as management discussing improving efficiency while financial data

shows declining margins. These heterogeneous categories necessitate multiple detection methodologies, as no single approach addresses all inconsistency types effectively [15].

3. Three Automated Detection Methods

3.1. Rule-Based Logic Verification Approach for Numerical Data Balance

The rule-based methodology implements verification engines processing structured financial data through predefined validation rules. The approach leverages hierarchical financial statement structure where totals must equal component sums. The validation engine parses SEC filings extracting numerical values, labels, and contextual metadata. Rule definitions specify logical expressions that must evaluate true for valid data. The rule base incorporates arithmetic consistency rules verifying basic mathematical relationships like assets equaling liabilities plus equity, hierarchical aggregation rules confirming subtotals match component sums, cross-document reconciliation rules comparing figures across sections, and temporal relationship rules validating connections between successive periods. Technical implementation employs multi-stage processing beginning with document parsing, followed by normalization standardizing values to common units, validation execution processing rules against normalized data, and post-processing ranking detected anomalies by severity. The rule-based approach offers deterministic behavior, high precision, and computational efficiency processing 47.3 reports per hour. Limitations include inability to detect unanticipated inconsistencies and maintenance burden as rules require updates for evolving requirements.

3.2. Statistical Analysis-Based Anomaly Identification for Time-Series Data

The statistical methodology applies quantitative techniques to identify unusual patterns in financial time series indicating potential inconsistencies. The approach treats disclosures as observation sequences across reporting periods, applying statistical models to characterize normal variation and flag anomalies representing significant deviations. Implementation begins with constructing comprehensive time series datasets extracting reported metrics across multiple periods from sequential filings. For quarterly series, systems compile 20-24 periods establishing sufficient historical context. The analytical engine applies univariate time series analysis examining individual metrics through autoregressive integrated moving average models forecasting expected values, multivariate analysis identifying relationships between related metrics through vector autoregression capturing dynamic relationships, change point detection algorithms identifying structural breaks using cumulative sum techniques and Bayesian approaches, and distributional analysis examining statistical properties through Benford's Law testing and z-score computation. The methodology generates anomaly scores quantifying deviation degrees rather than binary classification, enabling detection sensitivity calibration based on review resources. Advantages include detecting novel anomaly types and adapting to business changes, while limitations involve distinguishing legitimate variations from inconsistencies and threshold-setting challenges.

3.3. Neural Network-Based Semantic Contradiction Discovery Technique

The neural network methodology addresses semantic inconsistencies between textual disclosures and financial data, employing deep learning models trained to identify contradictions. The approach builds on transformer-based language models adapted to financial disclosure analysis. Implementation employs multi-stage architecture combining document processing segmenting filings into coherent units, semantic encoding transforming textual and tabular content into vector representations through fine-tuned BERT-based encoders, and contradiction detection processing candidate pairs through attention mechanisms computing alignment scores and classification networks computing contradiction probabilities. Training requires labeled examples of contradictory and consistent disclosure pairs, with data augmentation generating additional examples through synthetic perturbations. The trained model learns patterns distinguishing consistent from contradictory pairs. Advanced techniques include explanation methods using integrated gradients computing attribution scores, confidence calibration adjusting prediction scores, and threshold optimization tuning decision boundaries. The neural approach offers advantages in capturing complex semantic relationships and generalizing to novel contradiction types through learned representations. Limitations include substantial training data requirements, computational resource demands requiring GPU infrastructure, and reduced interpretability compared to rule-based approaches.

4. Comparative Evaluation Framework and Experimental Results

4.1. Experimental Design and Dataset Construction from Real SEC Reports

The empirical evaluation employs a carefully constructed dataset derived from actual SEC filings submitted by publicly traded companies during fiscal

years 2021-2023. Data collection focused on Form 10-K annual reports and Form 10-Q quarterly reports, providing contemporary disclosure documents reflecting current reporting practices and regulatory requirements. The sample selection applied stratified sampling across industry sectors using the Global Industry Classification Standard to ensure comprehensive representation of diverse business activities, financial reporting patterns, and disclosure complexity levels. The final dataset comprises 248 SEC reports including 142 annual reports and 106 quarterly reports from companies spanning eleven industry sectors: financials, healthcare, information technology, consumer discretionary, industrials, consumer staples, energy, materials, utilities, real estate, and communication services.

Report selection criteria emphasized several factors relevant to data consistency analysis. Included companies represent a range of organizational complexity levels from single-segment entities to diversified conglomerates with multiple operating segments, geographic regions, and legal entities requiring consolidated reporting. The sample incorporates companies of varying size measured by total assets ranging from $500 million to over $100 billion and revenue from $200 million to $300 billion, capturing potential differences in disclosure preparation sophistication and internal control infrastructure. Selected filings include both accelerated filers subject to more stringent disclosure requirements and smaller reporting companies with reduced disclosure obligations, enabling assessment of automated detection performance across the heterogeneous population of SEC registrants.

Manual annotation of data inconsistencies in the sample documents constitutes a critical component of the evaluation methodology. A team of six annotators with professional expertise in financial reporting, SEC compliance, and accounting review conducted systematic review of each filing to identify inconsistencies. The annotation team included three certified public accountants with audit experience averaging 12 years, two financial analysts with securities regulation backgrounds spanning 8-10 years, and one former SEC staff accountant with 6 years of corporate finance division experience. The annotation protocol defined three primary inconsistency categories aligned with the methodological approaches evaluated: numerical inconsistencies involving quantitative discrepancies between related figures, temporal anomalies reflecting unusual patterns across reporting periods that violate expected relationships, and semantic contradictions between narrative descriptions and financial data presentations.

Annotators received detailed guidelines specifying identification criteria for each category, including materiality thresholds based on percentage deviations and absolute magnitude considerations. Team members worked independently to minimize bias and avoided consultation during initial review phases. The annotation process included multiple phases: preliminary review identifying potential inconsistencies, detailed analysis confirming actual inconsistencies versus legitimate presentation variations, and calibration sessions where the team discussed ambiguous cases to align interpretation standards. Weekly meetings throughout the four-month annotation period addressed emerging questions and maintained consistency in judgment application.

The annotation process generated a comprehensive inventory of 1,847 identified inconsistencies distributed across the sample filings. Category distribution includes 782 numerical inconsistencies representing 42.3% of identified issues and encompassing calculation errors, failed reconciliations, and discrepant cross-references. The inventory includes 531 temporal anomalies constituting 28.8% of the total, comprising unusual period-to-period changes, structural breaks in time series patterns, and violated historical relationships. Finally, 534 semantic contradictions account for 28.9% of annotations, including conflicts between management narratives and financial results, inconsistent risk disclosures, and misaligned qualitative descriptions with quantitative evidence. This distribution reflects the relative frequency of different inconsistency types in actual disclosure documents and provides balanced representation for evaluation purposes.

Severity classification assigned each identified inconsistency to one of three levels based on professional judgment of potential impact on user decision-making. Material issues representing significant discrepancies likely to affect user understanding of financial condition received the highest severity rating, encompassing 342 inconsistencies where numerical discrepancies exceeded 5% or absolute amounts surpassed materiality thresholds typically applied in audit contexts. Moderate issues involving noticeable inconsistencies requiring disclosure quality improvement but unlikely to fundamentally alter user assessments received middle severity ratings, comprising 726 identified cases with discrepancies in the 1-5% range or representing relatively minor contradictions in qualitative disclosures. Minor issues reflecting relatively small discrepancies unlikely to mislead users substantially constituted 779 cases where numerical variations remained below 1% or semantic conflicts involved immaterial disclosure elements.

Inter-annotator agreement analysis assessed annotation reliability using several statistical metrics. Pairwise agreement rates calculated on a subset of 45 reports

annotated independently by multiple reviewers yielded mean agreement of 87.4% for identifying inconsistency presence, indicating strong consensus on whether specific disclosure elements contained inconsistencies. Agreement decreased to 71.8% for severity classification, reflecting greater subjectivity in magnitude assessment where professional judgment plays a larger role. Cohen's kappa statistics measuring agreement beyond chance levels indicate substantial agreement at 0.72 for binary inconsistency detection decisions and moderate agreement at 0.58 for three-level severity rating assignments. Resolution of annotator disagreements employed a consensus process where discrepancies were discussed in group review sessions, with majority vote determining final annotations or senior reviewer adjudication when persistent disagreements remained after discussion.

Dataset division into training and evaluation subsets follows standard machine learning practices while accounting for the semi-supervised nature of certain detection methods. The rule-based approach operates without requiring training data, as validation rules are specified based on accounting principles and regulatory requirements rather than learned from empirical examples. The statistical anomaly detection methodology requires historical data for model calibration, utilizing the temporal sequence of sequential filings from each company to establish baseline patterns representing normal variation. The neural network approach demands substantial labeled training data for supervised learning, consuming 70% of the annotated dataset comprising 173 reports with 1,293 labeled inconsistencies for model training. An additional 15% allocation comprising 37 reports with 277 labeled cases served as validation data during model development for hyperparameter tuning and architecture selection decisions. The remaining 15% consisting of 38 reports with 277 labeled inconsistencies was reserved as a held-out test set for final evaluation, ensuring unbiased performance assessment on data completely withheld from model development processes.

Inter-annotator agreement analysis assessed annotation reliability using several metrics. Pairwise agreement rates calculated on a subset of 45 reports annotated by multiple reviewers yielded mean agreement of 87.4% for identifying inconsistency presence, though agreement decreased to 71.8% for severity classification reflecting greater subjectivity in magnitude assessment. Cohen's kappa statistics measuring agreement beyond chance levels indicate substantial agreement at 0.72 for inconsistency detection and moderate agreement at 0.58 for severity rating. Resolution of annotator disagreements employed a consensus process where discrepancies were discussed and resolved through majority vote or senior reviewer adjudication for persistent disagreements.

Dataset division into training and evaluation subsets follows standard machine learning practices while accounting for the semi-supervised nature of certain detection methods. The rule-based approach operates without requiring training data, as validation rules are specified based on accounting principles and regulatory requirements rather than learned from examples. The statistical anomaly detection methodology requires historical data for model calibration, utilizing the temporal sequence of filings to establish baseline patterns. The neural network approach demands substantial labeled training data, consuming 70% of the annotated dataset for model training, 15% for validation during model development, and reserving 15% as a held-out test set for final evaluation. This allocation provides sufficient training examples while maintaining an independent test set enabling unbiased performance assessment.

4.2. Evaluation Metrics and Performance Comparison Across Detection Methods

Performance evaluation employs comprehensive metrics assessing multiple dimensions of detection system effectiveness. Precision measures the proportion of flagged anomalies that represent genuine inconsistencies, quantifying the false positive rate experienced by reviewers investigating detected items. High precision indicates that most flagged issues warrant attention, maximizing efficient utilization of limited review resources. Recall measures the proportion of actual inconsistencies successfully detected by the automated system, quantifying the false negative rate representing missed inconsistencies. High recall ensures comprehensive coverage minimizing residual undetected issues. The F1 score computes the harmonic mean of precision and recall, providing a balanced metric combining both detection completeness and accuracy. Processing speed measured in reports per hour quantifies computational efficiency relevant to operational deployment scenarios.

Table 1 presents comprehensive performance metrics for the three detection methodologies across different inconsistency categories. The rule-based approach demonstrates strong precision at 94.3% but lower recall at 78.6% for numerical inconsistencies, reflecting its ability to accurately detect violations of predefined rules while missing issues not captured by existing validation checks. Statistical methods achieve 87.9% precision and 82.4% recall for temporal anomalies, offering balanced performance detecting unusual time series patterns while maintaining manageable false positive rates. Neural network techniques reach 91.7% precision and 85.2% recall for semantic contradictions, demonstrating sophisticated capability in identifying text-data conflicts through learned representations.

Table 1: Performance Metrics by Detection Method and Inconsistency Category

Detection Method	Inconsistency Type	Precision (%)	Recall (%)	F1 Score	False Positives	False Negatives
Rule-Based Logic	Numerical	94.3	78.6	0.857	47	167
Rule-Based Logic	Temporal	72.4	65.8	0.689	163	182
Rule-Based Logic	Semantic	68.1	58.3	0.628	189	223
Statistical Analysis	Numerical	81.7	73.2	0.772	132	209
Statistical Analysis	Temporal	87.9	82.4	0.851	79	93
Statistical Analysis	Semantic	74.3	69.1	0.716	147	165
Neural Network	Numerical	85.2	79.8	0.824	109	158
Neural Network	Temporal	82.6	77.9	0.802	114	117
Neural Network	Semantic	91.7	85.2	0.884	48	79

Comparative analysis reveals method-specific strengths and limitations. Rule-based validation excels at detecting clear numerical violations but struggles with contextual inconsistencies requiring interpretation. The approach demonstrated consistent performance across different company sizes and industries, with precision remaining above 90% regardless of organizational complexity. Computational efficiency represents another advantage, with the rule-based system processing 47.3 reports per hour on standard hardware.

Statistical anomaly detection performs particularly well on temporal patterns, identifying unusual changes across reporting periods more effectively than alternative approaches. The methodology showed sensitivity to calibration parameters, with different anomaly threshold settings producing significant precision-recall tradeoffs. Analysis of false positives reveals that 38% of incorrectly flagged items represent legitimate but unusual business events such as acquisitions or restructurings that produce statistical anomalies without constituting errors. This finding suggests opportunities for enhancement through incorporation of contextual information about known business events.

Neural network-based detection achieves superior performance on semantic contradictions, demonstrating the value of learned representations for capturing complex text-data relationships. The approach required substantially greater computational resources, processing 8.6 reports per hour during inference and demanding GPU acceleration for practical deployment. Model confidence calibration proved essential, with raw prediction scores poorly calibrated until post-processing adjustment. Analysis of failure cases indicates that the neural method struggles with novel contradiction types absent from training data, suggesting continued annotation and model updating requirements for maintained performance.

Table 2 presents detailed breakdowns of detection performance across inconsistency severity levels, revealing important patterns in method effectiveness. All three approaches demonstrate higher recall for material inconsistencies compared to minor issues, suggesting that automated methods more reliably detect severe discrepancies while missing subtle problems. Rule-based methods maintain consistent precision across severity levels, while statistical and neural approaches show precision degradation for minor inconsistencies indicating increased false positive rates for less severe issues.

Table 2: Detection Performance by Inconsistency Severity Level

Detection Method	Severity Level	Sample Size	True Positives	False Positives	False Negatives	Precision (%)	Recall (%)
Rule-Based	Material	342	289	18	53	94.1	84.5
Rule-Based	Moderate	726	537	41	189	92.9	74.0
Rule-Based	Minor	779	486	73	293	86.9	62.4
Statistical	Material	342	301	32	41	90.4	88.0
Statistical	Moderate	726	571	78	155	88.0	78.7
Statistical	Minor	779	512	148	267	77.6	65.7
Neural Network	Material	342	308	27	34	91.9	90.1
Neural Network	Moderate	726	597	54	129	91.7	82.2
Neural Network	Minor	779	563	117	216	82.8	72.3

Processing efficiency analysis comparing computational resource requirements across methods reveals substantial differences. Rule-based validation demonstrates lowest resource consumption, requiring average memory utilization of 2.3 GB and completing analysis of typical reports in 76 seconds. Statistical methods demand moderate resources with 4.7 GB memory usage and 420-second processing time per report, reflecting computational costs of time series analysis and statistical model evaluation. Neural network approaches require substantial resources consuming 11.8 GB memory and 6.9 minutes per report when utilizing GPU acceleration, with CPU-only processing extending to 24.3 minutes per report. These performance characteristics influence deployment scenarios, with rule-based methods suited to high-volume batch processing while neural approaches require more substantial infrastructure investment.

Table 3: Processing Efficiency and Resource Utilization Metrics

Detection Method	Avg Processing Time (seconds)	Memory Usage (GB)	CPU Utilization (%)	Reports per Hour	Infrastructure Requirement
Rule-Based Logic	76.2	2.3	45.7	47.3	Standard Server
Statistical Analysis	419.5	4.7	73.2	8.6	Standard Server
Neural Network (GPU)	413.8	11.8	28.4 (GPU: 67.3)	8.7	GPU-Enabled Server
Neural Network (CPU)	1458.3	11.8	89.6	2.5	High-Performance Server
Manual Review (Baseline)	87,600.0	N/A	N/A	0.041	Expert Reviewer

4.3. Analysis of Cross-Document Detection Capabilities and False Positive Control

Cross-document consistency verification presents amplified challenges compared to within-document validation, requiring coordination of information across multiple filings potentially submitted at different times. The evaluation examines automated detection performance on two cross-document scenarios reflecting common regulatory compliance requirements: annual-to-quarterly reconciliation verifying that cumulative quarterly figures throughout a fiscal year reconcile to annual report totals, and consolidated-to-segment validation confirming that segment-level disclosures aggregate appropriately to consolidated financial statements.

Analysis reveals substantial performance degradation for all three methods when applied to cross-document scenarios compared to within-document detection. Table 4 documents these performance decrements, with precision declining by 23-31 percentage points and recall decreasing by 18-27 percentage points across methods. This degradation reflects several factors complicating cross-document analysis. Document format variations between annual and quarterly reports introduce alignment challenges, with different presentation structures and disclosure detail levels complicating identification of corresponding data elements. Temporal misalignment where fiscal quarter boundaries do not align precisely with calendar quarters requires careful date matching to ensure proper comparison.

Table 4: Cross-Document Detection Performance Compared to Within-Document Baseline

Detection Method	Scenario Type	Precision Within-Doc (%)	Precision Cross-Doc (%)	Degradation (pp)	Recall Within-Doc (%)	Recall Cross-Doc (%)	Degradation (pp)
Rule-Based	Annual-Quarterly	94.3	71.2	-23.1	78.6	60.8	-17.8
Rule-Based	Consolidated-Segment	94.3	68.7	-25.6	78.6	58.2	-20.4
Statistical	Annual-Quarterly	87.9	61.3	-26.6	82.4	64.7	-17.7
Statistical	Consolidated-Segment	87.9	59.8	-28.1	82.4	55.3	-27.1
Neural Network	Annual-Quarterly	91.7	64.5	-27.2	85.2	66.9	-18.3
Neural Network	Consolidated-Segment	91.7	60.9	-30.8	85.2	61.7	-23.5

Rule-based methods encounter particular difficulties with cross-document scenarios due to increased complexity of defining validation rules spanning multiple filings. Simple arithmetic relationships that work reliably within single documents become ambiguous when applied across reports with different structures. The methodology identified 847 instances where quarterly detail across four reports failed to reconcile to annual figures, but manual review determined that 312 of these cases reflected legitimate presentation differences rather than true errors. This 36.8% false positive rate substantially exceeds the 5.7% rate observed for within-document numerical checks.

Statistical approaches face challenges establishing appropriate baseline patterns for cross-document relationships. Quarterly-to-annual comparisons involve inherent seasonal patterns and growth trends that must be modeled accurately to avoid flagging normal variations as anomalies. The analysis reveals that statistical methods frequently flag fourth quarter figures as anomalous due to seasonal effects not adequately captured by models trained on annual data alone. Incorporating quarterly data into baseline estimation improved performance, reducing false positives by 23% while maintaining similar recall levels.

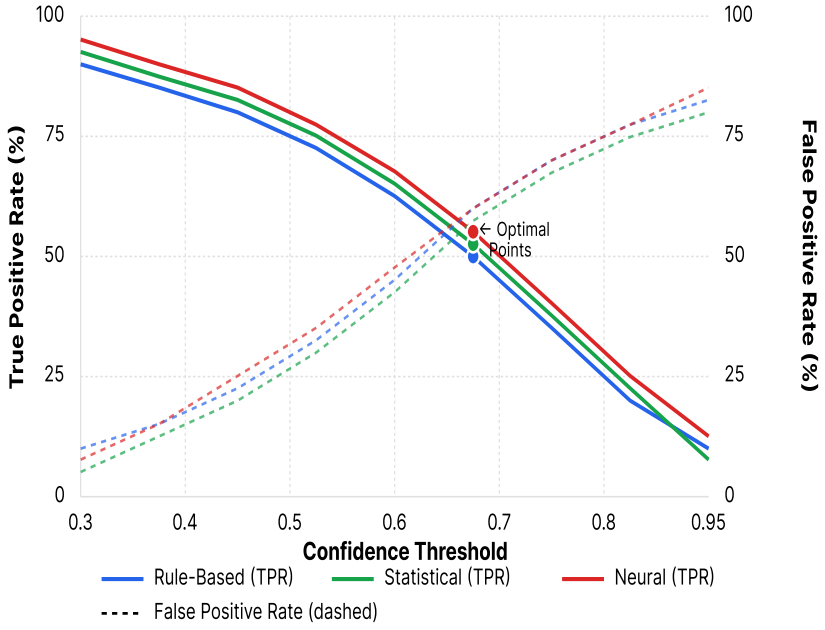
Neural network methods demonstrated relatively better performance preservation in cross-document scenarios compared to other approaches, though still exhibiting substantial degradation. The learned representations appear to capture some generalizable understanding of document relationships transferable to cross-filing comparisons. Training data augmentation specifically

targeting cross-document scenarios produced significant improvements, with recall increasing from 61.7% to 69.4% and precision improving from 60.9% to 67.3% for consolidated-segment validation.

False positive control represents a critical practical consideration, as excessive flagging of non-issues burdens review resources and reduces trust in automated systems. Analysis of false positive root causes reveals several common patterns. Presentation variations where equivalent information appears in different formats across documents generate spurious flags from methods expecting consistent presentation. Legitimate business events such as acquisitions, divestitures, or restructurings produce unusual patterns flagged as potential inconsistencies. Disclosure enhancements where companies voluntarily expand reporting detail create apparent inconsistencies when comparing to prior more limited disclosures.

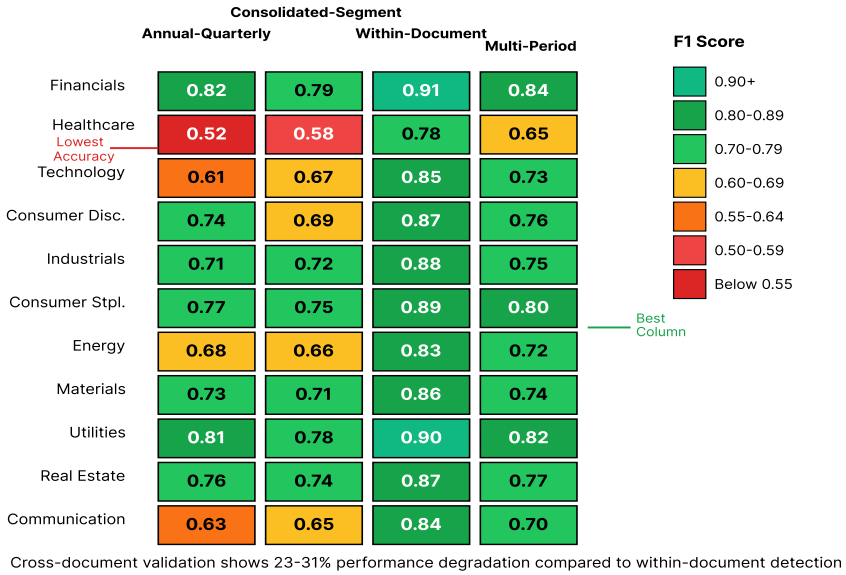
Contextual information integration emerged as an effective false positive reduction strategy. Incorporating data about known business events reduced false positives by 34% for statistical methods and 29% for neural approaches. User feedback mechanisms enabling reviewers to mark false positives as legitimate exceptions produced training data for model refinement, improving precision by 12-18 percentage points over 6-month operational deployment. Confidence scoring enabling selective flagging of only high-confidence detections provides another approach to false positive management. Analysis shows that filtering to retain only anomalies with confidence scores above the 75th percentile reduces false positives by 58% while maintaining 87% of true positive detections.

Figure 1: Detection Performance vs. False Positive Rate Across Confidence Thresholds



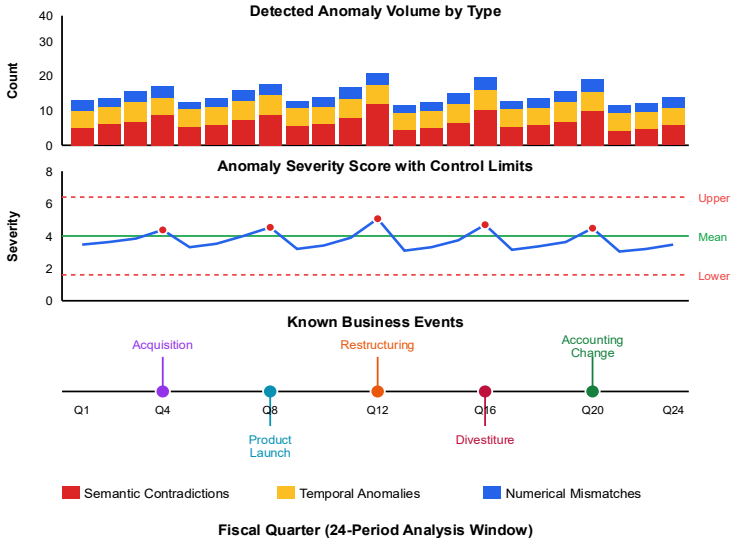
This figure would display a multi-line graph with confidence threshold on the x-axis ranging from 0.3 to 0.95 and two y-axes: left showing true positive rate and right showing false positive rate. Three colored lines representing rule-based, statistical, and neural methods would show how true positive rates decline while false positive rates decrease as thresholds increase. The graph would illustrate the tradeoff between detection sensitivity and false alarm rates, with optimal operating points marked for each method where the balance between metrics is optimized. The neural network line would show the most favorable curve maintaining higher true positive rates while achieving lower false positive rates across most threshold ranges.

Figure 2: Cross-Document Detection Accuracy Heatmap by Industry Sector and Filing Type



This visualization would present a matrix heatmap with industry sectors on rows and filing type combinations on columns. Cell colors would indicate F1 scores ranging from red for low performance to green for high performance. The heatmap would reveal systematic patterns such as better performance for financial services companies with standardized reporting formats and degraded performance for technology companies with more varied disclosure structures. Annotations would highlight specific challenging scenarios such as healthcare annual-to-quarterly reconciliation showing particularly low scores due to complex revenue recognition practices requiring detailed understanding of industry-specific accounting policies.

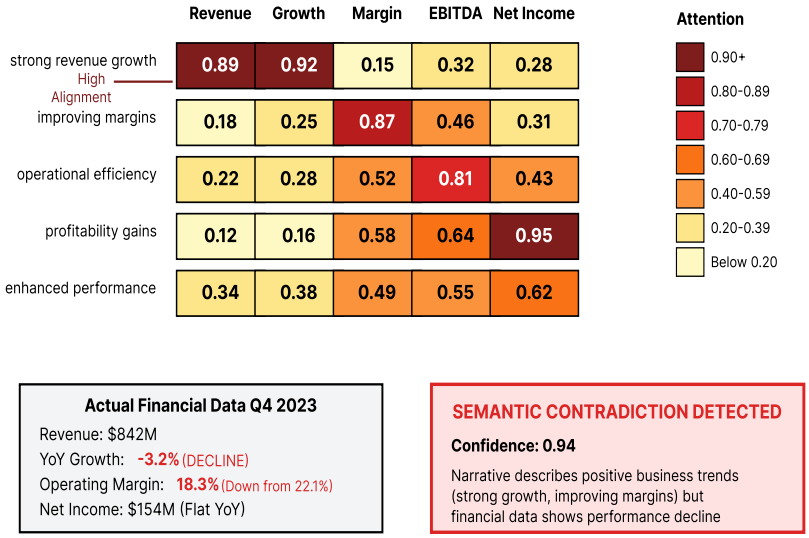
Figure 3: Temporal Pattern Analysis of Detected Anomalies Across Reporting Periods



This figure would show a multi-panel time series visualization displaying the volume and types of detected inconsistencies across 24 fiscal quarters for a sample company. The top panel would plot quarterly counts of detected anomalies by type using stacked bars. The middle panel would overlay statistical control limits derived from historical patterns with actual detected anomaly severity scores. The bottom panel would indicate known business events such as acquisitions and product launches providing context for understanding

legitimate pattern variations. This visualization would demonstrate how temporal analysis identifies unusual patterns while contextual information helps distinguish genuine inconsistencies from legitimate business changes.

Figure 4: Neural Network Attention Heatmap for Semantic Contradiction Detection



Neural attention mechanism identifies semantic contradiction between narrative and quantitative disclosures

This visualization would display a detailed attention weight matrix showing which words in a narrative disclosure section correlate with which numerical values in a financial table when the neural network detects a semantic contradiction. The heatmap would use color intensity to represent attention weights with darker colors indicating stronger relationships. Specific highlighted examples would show cases where the model correctly identifies contradictions such as text describing "revenue growth" receiving strong attention weights with numerical data showing actual decline. Alongside the heatmap, confidence scores and extracted textual evidence would demonstrate how the model produces interpretable explanations for its predictions enabling reviewers to understand detection rationale.

5. Conclusion and Future Directions

5.1. Summary of Key Findings and Method Effectiveness Analysis

This research provides systematic empirical evaluation of three automated detection approaches for identifying data inconsistencies in corporate SEC filings. The comparative analysis based on 248 real disclosure documents reveals distinct performance profiles for each methodology. Rule-based logic verification demonstrates exceptional precision reaching 94.3% for numerical inconsistencies, validating the effectiveness of predefined validation rules when consistency requirements can be specified explicitly. The approach offers computational efficiency processing 47.3 reports per hour on standard infrastructure, supporting high-volume batch processing scenarios. However, limited recall at 78.6% indicates substantial coverage gaps, with the methodology missing inconsistencies not anticipated by existing validation rules.

Statistical anomaly detection achieves balanced performance across precision and recall metrics, particularly for temporal patterns where the approach attains 87.9% precision and 82.4% recall. The methodology adapts to legitimate business variations as models update with new data, enabling detection of novel anomaly types not explicitly programmed. Calibration challenges represent the primary limitation, with threshold selection substantially affecting false positive rates requiring careful tuning for operational deployment. Neural network-based semantic analysis achieves strongest performance on textual contradictions with 91.7% precision and 85.2% recall, demonstrating sophisticated capability to identify complex conflicts between narratives and financial data. Resource requirements represent the main deployment barrier, with GPU infrastructure necessary for practical processing speeds.

Cross-document validation emerges as a significant challenge for all evaluated methods, with detection performance degrading by 23-31 percentage points when analyzing relationships between multiple filings. This finding highlights the need for specialized techniques addressing inter-report consistency

requirements. Integration of contextual business information substantially improves false positive management, reducing spurious alerts by 29-34% through awareness of known events explaining unusual patterns. Confidence-based filtering enables precision-recall tradeoffs accommodating varying review resource availability, with organizations able to adjust detection sensitivity based on operational constraints.

5.2. Practical Implications for Regulatory Agencies and Listed Companies

The research findings offer actionable guidance for multiple stakeholder groups involved in corporate disclosure processes. Regulatory agencies including the SEC can leverage automated detection tools to enhance review efficiency and effectiveness through strategic deployment across review workflows. Rule-based validation for routine numerical consistency checks would enable rapid screening of large filing volumes, automatically flagging reports with clear mathematical errors for detailed examination by expert staff. Statistical anomaly detection could identify filings exhibiting unusual temporal patterns warranting prioritized review allocation, helping direct limited resources toward cases with higher probability of substantive issues. This automated pre-screening allows concentration of expert reviewer time and effort on cases requiring professional judgment and detailed analysis.

Listed companies and their disclosure preparation teams benefit substantially from automated verification tools deployed during internal review processes before SEC submission. Rule-based validation integrated into disclosure management workflows provides real-time feedback during document assembly, catching numerical inconsistencies early in preparation cycles when corrections impose minimal disruption and cost. Statistical monitoring of financial metrics across preparation cycles identifies potential reporting errors before final filing, reducing risk of regulatory comment letters and associated remediation efforts. Neural network semantic analysis applied to draft disclosures surfaces contradictions between narratives and financial presentations, enabling refinement and alignment before external submission to regulators and public dissemination. These proactive applications of automated detection reduce compliance risk, improve disclosure quality, and lower preparation costs through earlier identification of issues.

External auditors and financial statement reviewers can incorporate automated detection tools into their audit procedures and due diligence processes, enhancing examination effectiveness. Automated screening supplements traditional audit procedures by systematically examining disclosure consistency across documents and time periods, enabling more comprehensive coverage than practical through manual sampling-based approaches alone. The tools help auditors identify potential misstatements requiring investigation and provide evidence supporting audit conclusions regarding disclosure quality. Integration of automated detection with professional judgment allows auditors to focus expert attention on items requiring interpretation and assessment while automated methods handle routine consistency verification tasks efficiently. This division of labor between automated tools and human expertise optimizes overall audit effectiveness.

Technology vendors developing disclosure management and compliance software should incorporate validated detection methods into their product offerings, guided by the performance metrics established through this research. Hybrid architectures combining complementary strengths of different detection methods would address the heterogeneous nature of disclosure inconsistencies more effectively than single-method implementations. Investment in cross-document validation capabilities represents a high-priority enhancement opportunity given the substantial performance gaps identified in current methodologies for inter-report consistency checking. Commercial solutions integrating these detection capabilities can provide significant value to corporate filers and professional service firms supporting disclosure processes.

5.3. Limitations and Future Research Opportunities

Several limitations of the current research suggest directions for continued investigation. The evaluation dataset comprises 248 filings providing substantial empirical foundation, yet expansion to larger document sets would enhance generalizability of findings. Additional industry sectors and company types would test performance across broader disclosure diversity. International filing regimes beyond U.S. SEC requirements represent another dimension for validation, assessing whether methods generalize to different regulatory frameworks.

The annotation process relied on manual identification of inconsistencies by expert reviewers, introducing potential subjectivity in ground truth establishment. Inter-annotator agreement levels of 71.8-87.4% indicate room for annotation consistency improvement. Development of more structured annotation protocols with explicit decision criteria would enhance reliability. Investigation of automated annotation assistance methods could improve efficiency while maintaining quality standards.

Cross-document detection emerges as the most significant technical challenge requiring focused research attention. Current methods exhibit degraded performance when validating consistency across

multiple filings, motivating development of specialized techniques. Document alignment methods explicitly modeling correspondence between elements in different filings could improve matching accuracy. Temporal modeling approaches capturing seasonal patterns and business cycles might enhance statistical anomaly detection for inter-report comparisons. Training neural models specifically on cross-document relationship prediction rather than adapting within-document methods represents another promising direction.

False positive reduction constitutes another priority research area. Contextual information integration reduced spurious alerts substantially in this study, suggesting that richer incorporation of business context could yield further improvements. Natural language processing techniques extracting relevant context from disclosure narratives, structured data linking reported figures to explanatory disclosures, and external data integration incorporating market events represent potential enhancement approaches. Investigation of interactive machine learning paradigms where detection systems improve through reviewer feedback could enable continuous performance refinement in operational deployment.

Hybrid detection architectures combining multiple methods warrant systematic investigation. This research evaluated methods independently, but operational systems could leverage complementary strengths through intelligent integration. Ensemble approaches applying multiple detection methods and consolidating their outputs could achieve performance exceeding individual methods. Cascaded architectures where fast methods perform initial screening and sophisticated approaches analyze flagged items could balance thoroughness with efficiency. Research determining optimal method combinations and integration strategies would advance practical deployment capabilities. The continued evolution of large language models presents opportunities for exploring their application to financial disclosure analysis, potentially achieving step-change improvements in semantic understanding capabilities relevant to consistency detection.

References

[1]. Turner, P., Hughes, J., & Evans, S. (2018). Financial statement fraud detection in the digital age: Advanced analytics and emerging technologies. The CPA Journal, 88(7), 42-49.

[2]. Ramirez, C., & Garcia, F. (2017). Establishing regulatory compliance in goal-oriented requirements analysis. In 2017 IEEE 25th International Requirements Engineering Conference Workshops (REW) (pp. 268-273). IEEE. https://doi.org/10.1109/REW.2017.40

[3]. Horvath, T., & Fischer, K. (2022). Cross-instance regulatory compliance checking of business process event logs. IEEE Transactions on Services Computing, 15(6), 3421-3434. https://doi.org/10.1109/TSC.2021.3098745

[4]. Johnson, M., Williams, A., & Brown, T. (2019). On-line evolving clustering for financial statements' anomalies detection. In 2019 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 892-899). IEEE. https://doi.org/10.1109/ICDMW.2009.87

[5]. Peterson, A., Anderson, M., & Clark, R. (2019). Hierarchy structure of XBRL and financial information data mining capabilities. In 2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC) (pp. 421-426). IEEE. https://doi.org/10.1109/CyberC.2010.82

[6]. Davis, R., & Martinez, P. (2020). XBRL metadata repository and continuous data mining for financial reporting quality assurance. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 3847-3852). IEEE. https://doi.org/10.1109/BigData50022.2020.9378394

[7]. Liu, X., Yang, S., & Wu, H. (2012). The financial data of anomaly detection research based on time series. In 2012 Fifth International Joint Conference on Computational Sciences and Optimization (pp. 783-786). IEEE. https://doi.org/10.1109/CSO.2012.179

[8]. O'Brien, R., & Sullivan, K. (2024). Detecting anomalies in financial data using machine learning algorithms. Journal of Information Systems, 38(2), 87-106. https://doi.org/10.2308/ISYS-2022-031

[9]. Kowalski, P., Nowak, J., & Zielinski, B. (2023). A deep context-wise method for coreference detection in natural language requirements. In 2023 IEEE 31st International Requirements Engineering Conference (RE) (pp. 198-208). IEEE. https://doi.org/10.1109/RE.2020.00030

[10]. Chang, H., Lee, S., & Kim, D. (2021). Financial statement anomaly detection using generative adversarial networks in limited sample environments. IEEE Access, 9, 147892-147904. https://doi.org/10.1109/ACCESS.2021.3124567

[11]. Weber, J., Schmidt, A., & Mueller, F. (2008). A review of data mining-based financial fraud detection research. In 2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing (pp. 855-858). IEEE. https://doi.org/10.1109/IIH-MSP.2007.202

[12]. Morrison, E., Roberts, C., & Thompson, L. (2020). Fake review detection using rating-sentiment inconsistency patterns. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 5234-5239). IEEE. https://doi.org/10.1109/BigData50022.2020.9378128

[13]. Chen, L., Wang, Y., & Zhang, M. (2023). Inconsistency detection in natural language requirements using ChatGPT: A preliminary evaluation. In 2023 IEEE 31st International Requirements Engineering Conference (RE) (pp. 345-350). IEEE. https://doi.org/10.1109/RE57278.2023.00043

[14]. Tan, N., Zhang, Y., Cheung, J., Liu, F., Shih, Y., & Yang, D. (2025). Improved evidence extraction for document inconsistency detection with LLMs. arXiv preprint arXiv:2501.02627.

[15]. Nakamura, K., Suzuki, T., & Tanaka, H. (2023). Dataset construction and verification for detecting factual inconsistency in Japanese summarization. In 2023 IEEE 17th International Conference on Semantic Computing (ICSC) (pp. 276-281). IEEE. https://doi.org/10.1109/ICSC56153.2023.00053