Scientific Research Publishing

Journal of Computer and Communications, 2025, 13(12), 80-99
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

SHAP-Driven Interpretability in Financial Fraud Detection: A Multimodal Data Approach

Nie Hui

School of Information Management, Sun Yat-Sen University, Guangzhou, China
Email: issnh@mail.sysu.edu.cn

How to cite this paper: Hui, N. (2025) SHAP-Driven Interpretability in Financial Fraud Detection: A Multimodal Data Approach. *Journal of Computer and Communications*, **13**, 80-99.
https://doi.org/10.4236/jcc.2025.1312005

Received: November 9, 2025
Accepted: December 16, 2025
Published: December 19, 2025

Abstract

Improved accuracy in predicting corporate financial fraud significantly enhances regulatory efficiency and market stability. However, detecting increasingly sophisticated fraud patterns remains challenging due to data heterogeneity and model opacity. This study proposes an innovative, multimodal framework that integrates corporate financial indicators, organizational structures, and semantic features from annual reports. We employ BERT-based semantic extraction on the Management Discussion & Analysis (MD&A) section of an annual report, reduce dimensionality via PCA, and fuse features with corporate financial/organization structural metrics. Ensemble tree models (CatBoost/XGBoost/LightGBM) are optimized for fraud prediction, while SHAP values quantify the contributions of individual features. Experimental results demonstrate a peak ROC-AUC of 0.859, with key findings revealing that: 1) Significant asset transactions, future outlook, and operational overview are the most predictive MD&A contents; 2) Financial indicators dominate feature importance (50% of top predictors); 3) Annual report similarity and tone serve as critical textual red flags. This framework provides regulators with actionable insights through model interpretability, thereby advancing early-warning systems for financial misconduct.

Keywords

Financial Fraud Detection, Multimodal Data Fusion, Explainable AI, Textual Semantic Analysis, Ensemble Learning

1. Introduction

Listed companies constitute a critical component of the market economy, with their financial health and operational performance exerting a direct influence on the stability and development of a nation's economy. In recent years, however, the

frequent occurrence of financial fraud incidents among Chinese listed companies has imposed severe negative repercussions on the market. As the market economy grows increasingly complex and corporate fraud becomes more sophisticated and concealed, regulatory agencies face unprecedented challenges in detecting financial fraud. Traditional financial fraud detection (FFD) methods, which rely heavily on limited datasets, are proving inadequate in addressing the vast and heterogeneous data generated by modern business operations [1] [2]. Consequently, exploring novel methodologies to detect increasingly complex and hidden corporate fraud has become imperative.

Machine learning algorithms have emerged as a promising solution due to their high predictive accuracy and adaptability to diverse data types, distributions, and volumes. These algorithms are particularly well-suited for large-scale, complex data prediction tasks. However, the black-box nature of machine learning models has hindered their practical application in real-world scenarios. Current research on machine learning-based financial fraud prediction models primarily focuses on enhancing predictive performance, with limited attention paid to the role of predictive factors. As a result, the relationship between these factors and corporate financial fraud risk remains inadequately explored [3] [4]. In practice, interpretable models are often preferred due to their transparency and credibility, making them more readily accepted by stakeholders [3].

The rapid expansion of diverse data sources has pushed financial fraud prediction beyond traditional metrics. Predictive models increasingly incorporate external information—organizational charts, annual reports, financial news, and stock commentaries—to boost accuracy and reveal latent fraud signals, offering a fuller view of corporate operations. However, research has not yet fully realized the synergistic integration of external and financial data or exploited multimodal collective effects; analysis and use of textual information remain fragmented and underdeveloped.

This study addresses those gaps by applying multimodal data fusion and explainable machine learning to corporate financial fraud prediction. Using NLP, we analyze annual reports and combine report-derived features with financial indicators to predict fraud risk in listed companies. We also use explainable models to identify latent risk factors from publicly disclosed information, maximizing the value of multilayered data and offering new perspectives and technical solutions for multimodal fraud detection.

2. Literature Review

2.1. Indicators for Detecting Corporate Financial Fraud

Research on corporate financial fraud detection has established a comprehensive framework of indicators. Early studies primarily focused on financial metrics, including profitability, solvency, growth potential, operational efficiency, and cash flow [4] [5]. However, as market environments have grown more complex and fraudulent practices more sophisticated, reliance on financial metrics alone has

proven insufficient for effective risk identification. Consequently, researchers have expanded the scope of fraud detection to include non-financial indicators, like board structure or executive team characteristics [6] [7]. These indicators offer insights into the quality of corporate governance. For instance, Liu *et al.* [8] demonstrated that the shareholding ratio of major shareholders serves as a significant predictor of fraudulent behavior in Chinese firms.

Beyond the tabular-style metrics, textual corporate annual reports have emerged as a critical resource for fraud detection. Annual reports, as the primary medium for corporate disclosure, are instrumental for investors in assessing financial risks. Studies reveal that firms engaging in financial misconduct often employ a dual-fraud strategy, manipulating both financial data and report narratives [9]. For example, Xu & Zhang [10] found that high-risk firms tend to adopt an overly optimistic tone in their reports to obscure irregularities. Similarly, Wang *et al.* [11] found a negative correlation between a firm's financial status and report readability, suggesting that obfuscation is a common tactic used conceal financial distress. Zhang *et al.* [12] incorporated readability metrics into fraud prediction models, achieving a 26.33% improvement in F1-score, underscoring the predictive value of textual.

Furthermore, an increasing number of advancements have deepened the analysis of the report content. Brown *et al.* [13] and Craja *et al.* [14] extracted direct fraud signals from report paragraphs, while Wu and Du [1] input text feature vectors to LSTM models to achieve a 94.98% prediction accuracy. Liu *et al.* [15] further integrated latent semantic features of reports with accounting metrics, demonstrating the synergistic benefits of multimodal data fusion.

2.2. Models for Detecting Financial Fraud

Methodologies for fraud prediction can generally be classified into three categories: statistical models, machine learning approaches, and deep learning techniques. Statistical models, such as logistic regression (LR), are valued for their predictive power and strong interpretability; however, they often encounter difficulties when dealing with high-dimensional or nonlinear data relationships [16]. By contrast, machine learning models are well-suited for capturing complex patterns in data. Notably, ensemble methods—such as random forests (RF), XGBoost, and LightGBM—have delivered outstanding performance in fraud detection tasks. For instance, Ali *et al.* [17] optimized an XGBoost model to achieve an accuracy rate of 96.05%, while Zhao & Bai [18] combined LR with XGBoost to further enhance predictive performance. Additionally, Yadav [19] demonstrated that deep neural networks (DNNs), when applied to textual data with thorough feature engineering, can achieve accuracy of up to 95%. Even so, the black-box nature of machine learning models limits their adoption in practical applications, as most studies prioritize accuracy over interpretability [17]. Moreover, ensemble methods predominantly rely on structured data, leaving textual information underexplored [15].

2.3. Emerging Directions and Research Gaps

Breakthroughs in NLP have significantly expanded the possibilities for fraud detection research. For example, Wu & Du [1] utilized word embeddings to vectorize annual reports, providing a robust foundation for subsequent analyses. Craja *et al.* [14] applied hierarchical attention networks to extract nuanced semantic features from financial texts. Building on these advances, Bhattacharya & Mickovic [20] fine-tuned BERT for analyzing 10-K reports, achieving a 15% performance improvement over traditional models. Additionally, Wang *et al.* [21] proposed a multimodal model that incorporates attention mechanisms to jointly analyze textual and financial data, thereby offering deeper insights into the factors influencing fraud risk.

Despite these advancements, three critical gaps persist in the literature. First, regarding multimodal data fusion, although Wang *et al.* [21] demonstrated that multimodal models outperform unimodal approaches, there is a paucity of research examining the effects of data clustering within such frameworks. Second, in terms of model interpretability, previous studies have not systematically explored the specific mechanisms through which various predictors impact model outputs [22]. Third, regarding textual depth, current analyses of annual reports primarily focus on lexical, sentiment, and readability features, with limited attention to more nuanced semantic representations.

To address these gaps, this study integrates advanced deep learning models, multimodal data fusion, and explainable machine learning methodologies. This approach not only aims to improve the accuracy of fraud risk predictions but also seeks to elucidate the key textual and financial determinants underlying these risks. By enhancing both semantic modeling and interpretability, this research aspires to provide methodological innovations for the development of more effective fraud early-warning systems.

3. Research Design

3.1. Research Framework

This study aims to develop a multimodal, data-driven model for predicting corporate fraud by integrating predictive factors across three dimensions: financial indicators, organizational structure, and annual reports. While financial indicators and organizational structure are represented as tabular data, annual reports are processed as textual data. The primary objective is to uncover latent fraud-related signals embedded within annual reports and to improve prediction accuracy through the integration of these diverse data sources. To this end, we first perform a comprehensive analysis of annual report content to establish a semantic description framework. Next, we combine features derived from financial indicators and organizational structure to construct the predictive model. Furthermore, explainable machine learning techniques are applied to quantitatively evaluate the contribution of each feature to fraud risk, thereby clarifying the model's reasoning process and providing actionable guidance for regulatory authorities and stake-

holders.

The research workflow is depicted in Figure 1. Feature extraction is conducted in the initial stage: financial and organizational structure indicators are obtained from the CSMAR database (www.gtarsc.com). Meanwhile, annual reports are parsed to generate four categories of textual features: tone, readability, content similarity, and semantics. These multimodal features are subsequently integrated into the predictive modeling process. Focusing on ensemble tree models, we utilize three widely adopted algorithms—LightGBM, XGBoost, and CatBoost—alongside LR as a baseline. After model training, hyperparameter optimization, and evaluation, the model with the best performance is selected. Finally, the SHAP (Shapley Additive Explanations) method is employed to interpret the model, assessing feature importance and their relationships with fraud risk, with particular attention given to fraud-related elements identified within annual reports.

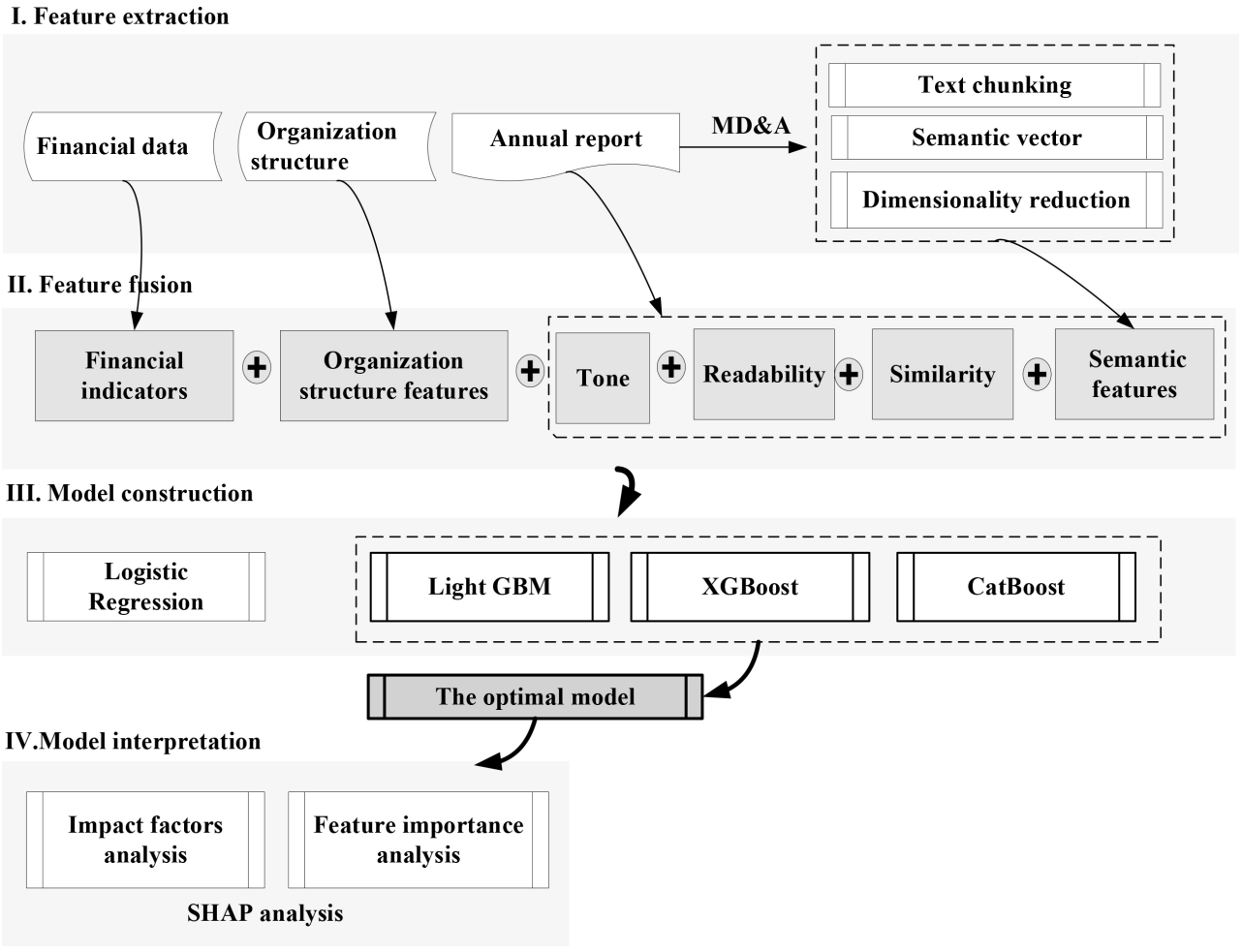

Figure 1. Research workflow for financial fraud detection.

3.2. Annual Report Text Analysis

3.2.1. Semantic Features of Annual Reports

Annual reports, required by regulators, reveal a company's operational status. Be-

cause worsening financial conditions often trigger corporate fraud, annual reports may contain signals of fraudulent behavior. To identify these signals systematically, we analyze the content of the report—focusing exclusively on the Management Discussion and Analysis (MD&A) section, which is authored by management and discusses causes, risks, and outlooks. The MD&A's language reflects management's mindset and strategy; under performance pressure or fraud incentives, it often grows more complex, ambiguous, or overly positive, indicating possible information manipulation. Therefore, the MD&A is a key textual source for detecting corporate fraud and management intent.

The MD&A section is lengthy but adheres to standardized disclosure rules, so we can split the section into nine definite thematic modules: 1) overall operational overview, 2) core business analysis, 3) non-core business analysis, 4) asset and liability status, 5) investment activities, 6) significant asset & equity transactions, 7) analysis of major subsidiaries and affiliates, 8) structured entities, and 9) future outlook. Each module is extracted using keyword-based regular expressions and truncated to fit the 512-character input limitation of the BERT model.

Subsequently, we conduct semantic encoding and dimensionality reduction for each module. For semantic encoding, we utilize the pre-trained BERT-Base-Chinese model, which features a 12-layer Transformer encoder with 12 self-attention heads per layer (Table 1 presents details on fine-tuning parameters). BERT generates a 768-dimensional semantic vector for each module, which is then reduced to between 3 and 20 dimensions through Principal Component Analysis (PCA) to enable integration with structured data features. The optimal dimensionality is determined by model performance across various settings. This reduction in feature dimensionality not only alleviates issues associated with high-dimensional data but also improves computational efficiency and enhances the model's interpretability.

Table 1. BERT fine-tuning parameters.

Parameters	Value	Interpretation
num_hidden_layers	12	number of hidden layers
hidden_size	768	hidden layer dimension
max_length	512	maximum length of input character sequence
num_attention_heads	12	number of attention heads
epoch	10	training epochs

3.2.2. Tone and Readability

The tone of an annual report corresponds to its textual sentiment features. In this study, we consider two indicators: the overall tone of the report and the tone of the MD&A section, both obtained from the CNRDS Annual Report Sentiment Database (www.cnrds.com). Consistent with the method proposed by Zeng *et al.* [23], the tone value is calculated using the Loughran-McDonald Financial Senti-

ment Dictionary. Specifically, as shown in Equation (1), the tone is determined based on the counts of positive words (*POSword*) and negative words (*NEGword*) within the text.

$$Tone = \frac{POSword - NEGword}{POSword + NEGword} \tag{1}$$

Readability, a linguistic feature that reflects text complexity, is typically inversely related to text length and the frequency of technical terms. While financial research uses various methods to assess readability, we simplify measurement by using the MD&A file size as a proxy, since text length correlates closely with cognitive load—longer texts require greater processing effort and thus indicate higher reading difficulty. In financial reports, which follow standardized formats, excessive length often signals structural complexity or redundancy, implying lower readability. Given this study's focus on textual content across a large sample of financial documents, file size is employed as a simple and accessible readability proxy.

3.2.3. Content Similarity

Content similarity measures the degree of change in annual reports across consecutive fiscal years and is typically calculated using cosine similarity (see Equation (2)). Higher similarity indicates fewer new disclosures [24], a pattern that previous research has associated with increased fraud risk (e.g., Qian & Zhu [25], 2020). In this study, cosine similarity is computed using document vectors generated by Doc2Vec, a widely used technique that effectively captures contextual information for representing texts in similarity analyses. In Equation (2), V_t represents the text vector for the annual report in year *t*.

$$\text{Similarity}\left(V_t, V_{t-1}\right) = \frac{V_t \cdot V_{t-1}}{\left\|V_t\right\| \cdot \left\|V_{t-1}\right\|} \tag{2}$$

3.3. Research Models

3.3.1. Ensemble Tree-Based ML Algorithms

In this study, financial fraud detection models are constructed using logistic regression (LR), XGBoost, LightGBM, and CatBoost. Logistic regression serves as the baseline model due to its simplicity and interpretability; however, it is limited in capturing complex, nonlinear relationships. In addition, the performance of gradient-boosted decision tree (GBDT) algorithms is evaluated. As noted by Chen & Guestrin [26], GBDT integrates a set of weak learners through iterative optimization to minimize loss functions. Previous studies have demonstrated that tree-based ensemble methods consistently outperform deep neural networks on structured tabular data [22], primarily due to their ability to model nonlinear relationships, robustness during training, and adaptability in terms of parameter sensitivity, computational efficiency, and handling of missing values.

The GBDT variants employed in this study each offer distinct advantages: XGBoost enhances generalization and computational speed through regulariza-

tion and parallelization; LightGBM reduces prediction error using histogram-based algorithms and leaf-wise tree growth; and CatBoost offers robust handling of categorical features while delivering high precision and stability.

3.3.2. Interpretable ML with SHAP

For the financial fraud detection model built using ensemble tree algorithms, we use the highly interpretable SHAP method to analyze the importance and mechanisms of various features. Rooted in Shapley value theory from game theory [27], SHAP decomposes model predictions into weighted contributions from input features, thereby enhancing the interpretability of complex machine learning models. Under the SHAP framework, the prediction output y_i for a given sample x_i can be expressed as a linear combination of all feature contributions, as shown in Equation (3):

$$y_i = y_{base} + \sum_{j=1}^{M} f\left(x_{ij}\right) \tag{3}$$

Here, y_{base} represents the global mean of the target variable (baseline value), x_{ij} denotes the *j*-th feature of sample x_i, and $f\left(x_{ij}\right)$ corresponds to the SHAP value of the feature, reflecting its marginal contribution to the prediction. A positive SHAP value indicates a promotive effect on fraud risk, while a negative value suggests an inhibitory effect. By quantifying these directional impacts, SHAP elucidates the model's decision logic. As illustrated in **Figure 2**, the baseline value y_{base} serves as the initial reference (e.g., average predicted probability). Arrows depict deviations from this baseline, e.g., feature x_{i1} exerts a positive contribution ($f\left(x_{i1}\right) > 0$), elevating the prediction above the baseline, while x_{i3} exhibits a negative contribution, reducing the prediction. The final prediction results from the algebraic sum of all feature contributions. SHAP provides post-hoc explanations applicable to various machine learning algorithms, particularly excelling in interpreting nonlinear decision processes of tree-based ensemble models.

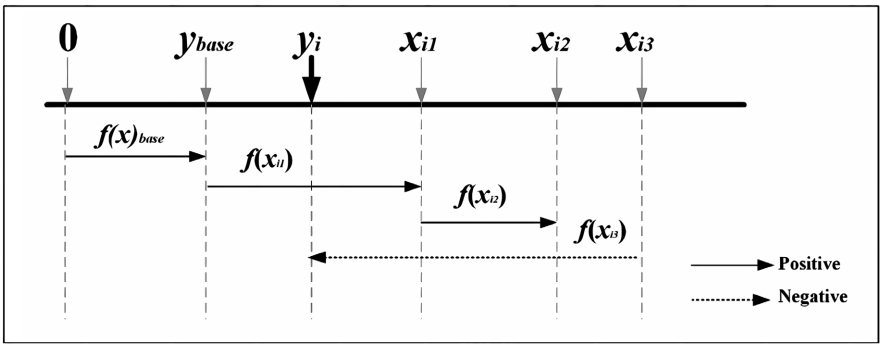

Figure 2. SHAP Feature attribution diagram.

The target variable in this study is corporate fraud risk, while the predictors include financial indicators, organizational structure features, and dimensions derived from annual reports. SHAP analysis is employed to quantify the marginal contributions of each predictor to fraud risk, identify early warning indicators sig-

nificantly associated with fraudulent activities, and provide deeper insights into the underlying mechanisms of corporate fraud.

4. Experiments and Results

4.1. Dataset Construction

4.1.1. Data Samples

This study utilizes violation records from the CSMAR database, focusing on firms listed on the Shanghai and Shenzhen Stock Exchanges between 2015 and 2020. Special attention is given to five prevalent forms of financial fraud: fictitious profits, inflated assets, false entries, major omissions, and disclosure misstatements. To ensure sample consistency and avoid confounding factors, financial firms are excluded due to their unique operational characteristics and higher financial risk. The resulting dataset comprises 1226 firms involved in financial misconduct, accounting for a total of 2652 fraud cases.

For the control group, non-fraudulent firms are selected from the CNRDS ESG-R database, which has provided ESG (Environmental, Social, and Governance) ratings for all Chinese A-share listed companies since 2007. Employing the ESG ratings as the criterion for non-fraud control firms has clear theoretical and empirical support. The governance dimension comprehensively reflects a company's standards in risk management, internal control, financial reporting quality, and information transparency. A higher governance score typically indicates a sound governance structure and effective oversight mechanisms, thereby reducing the likelihood of fraud. Cohen *et al.* [28] found that governance quality directly affects the reliability of financial reporting; Dechow *et al.* [29] showed that weak governance significantly increases fraud risk; and Tamimi and Sebastianelli [30] further demonstrated that higher scores under the ESG framework are associated with greater information transparency and lower fraud risk. Governance-related ESG indicators, including risk management capability, financial reporting quality, and disclosure transparency, have been proven to be relevant for fraud detection [31].

In the study, non-fraudulent firms are matched to fraudulent firms in a 1:1 ratio by year and industry. Additional criteria require non-fraudulent samples to have higher governance scores, no violation records from 2015 to 2020, and not be under ST (Special Treatment) status. These systematic matches yield 2652 non-fraudulent firms, thereby providing a robust foundation for comparative analysis.

4.1.2. Feature Variables

1) Financial Indicators

Nineteen financial indicators across five dimensions—solvency, profitability, growth potential, operational efficiency, and cash flow capacity—are used (see Table 2). Solvency indicators assess default risk, profitability measures earning ability, growth reflects expansion, operational efficiency captures asset utilization, and cash flow evaluates management of liquidity. Unusual variations in these indicators, such as sharp shifts in profitability or cash flow, can signal potential financial fraud.

Table 2. Financial indicators (19 Items).

Dimensions	Indicators	Definition and Operation
Solvency	*Aslbrt*	Asset Liability Ratio; Total Liabilities/Total Assets
	Curtrt	Flow Rate Ratio; Total Liability Ratio/Total Flow Asset
	Qikrt	Quick Ratio; (Current Assets - Inventory)/Current Liabilities
	Equrt	Equity Ratio; Total Liabilities/Total Shareholders' Equity
	Pmcptdbrt	Long-term Debt Ratio; Total Non-flowing Debt/(Total Shareholders' Equity + Total Non-flowing Liabilities)
Profitability	*Roe_1*	Return on Equity (ROE); Net Profit/Total Shareholders' Equity
	Salnpm	Net Profit margin; Net Profit/Revenue
	Salgm	Gross Profit Margin; Gross Profit/Revenue
	Salpm	Profit Margin; Profit/Revenue
Growth potential	*Atrt*	Total Assets Growth Rate; (Current Period Adjusted Figure - Prior Year Same Period Adjusted Figure)/Prior Year Same Period Adjusted Figure for ABS
	Opicrt	Operating Revenue Growth Rate; (Current Period Adjusted Figure - Prior Year Same Period Adjusted Figure)/Prior Year Same Period Adjusted Figure for ABS
	Oirt	Operating Profit Growth Rate, (Current Period Adjusted Figure - Prior Year Same Period Adjusted Figure)/Prior Year Same Period Adjusted Figure for ABS
Operational Efficiency	*Actrcbto*	Accounts Receivable Turnover; Revenue/((Beginning Net Accounts Receivable + Ending Net Accounts Receivable)/2)
	Fxastto	Fixed Asset Turnover; Total Revenue/((Beginning Fixed Assets + Ending Fixed Assets)/2)
	Totastto	Total Asset Turnover; Total Revenue/((Beginning Total Assets + Ending Total Assets)/2)
	Actpayto	Accounts Payable Turnover; Cost of Goods Sold (COGS)/((Beginning Accounts Payable + Ending Accounts Payable)/2)
Cash Flow Capacity	*Opncf_rev*	Net operating cash flow/Total Revenue
	Opncfrt	Percentage of net cash flow from operating activities; Net cash flow from operating activities/(Net cash flow from operating activities + Net cash flow from investing activities + Net cash flow from financing activities)
	Csopindex	Cash flow to sales ratio; Net cash flow from operating activities/Cash from operations

2) Organization Structure Features

Organizational structure related features are classified into two main categories: ownership concentration and top management characteristics (see Table 3). Ownership metrics evaluate the proportion of shares held by major shareholders, with particular emphasis on the concentration among the top five shareholders, a key indicator for fraud detection in China (Qian & Luo [7]). Additionally, a higher shareholding by the largest shareholder serves as an effective deterrent to fraud,

as it strengthens shareholder oversight and control [32]. Top management characteristics encompass variables such as dual roles, management shareholdings, and other related attributes.

Table 3. Features related to corporate organizational structure (11 Items).

Dimension	Indicators	Definition and Operation
Ownership Concentration	*ShrHolder1*	First Largest Shareholder Ownership Ratio
	ShrHolder3	Sum of Ownership Ratios of Top Three Shareholders
	ShrHolder5	Sum of Ownership Ratios of Top Five Shareholders
	ShrHolder10	Sum of Ownership Ratios of Top Ten Shareholders
	StOwRt	State-Owned Share Ratio
Top Management Characteristics	*Cmceo_Dum*	Whether serving as both Chairman and CEO (1 = Yes, 0 = No)
	Cmgm_Dum	Whether serving as both Chairman and General Manager (1 = Yes, 0 = No)
	MShrRat	Management Ownership Ratio: Percentage of company shares held by management
	BShrRat	Board of Directors Ownership Ratio: Percentage of company shares held by all board members
	SShrRat	Supervisory Board Ownership Ratio: Percentage of company shares held by all supervisory board members
	InDrcRat	Proportion of Independent Directors: Ratio of independent directors to the total number of directors

3) Annual Report Characteristics

Using the method from Section 3.2, we extract semantic features, tone, readability, and content similarity from the MD&A (see Table 4).

Table 4. Features related to corporate annual report.

Dimension	Indicators	Definition and Operation
Tone	*LM_tone2*	Annual report tone value
	LM_tone	MD&A section tone value
Readability	*FileSize*	MD&A text file size
Similarity	*Similarity*	Adjacent-year MD&A content similarity
Semantic Features	$TextFeatrue_1$ -$TextFeature_n$	MD&A semantic feature vector (n-dimensional)

4.2. Experimental Design

Following the overall research framework (see Figure 1), three main experiments were conducted. First, semantic features were extracted from MD&A texts using BERT-based classification models across nine thematic modules to predict corporate fraud risk. The resulting theme-specific semantic vectors were then subjected to dimensionality reduction, yielding compact representations for the entire MD&A section (denoted as $TextFeature_i$, where $i = 1 \cdots n$).

In the second experiment, feature fusion was performed by combining these semantic vectors with financial indicators, organizational structure metrics, and additional features related to tone, readability, and content similarity from annual reports. An ensemble tree-based model was optimized using grid search to achieve the best predictive performance.

The third experiment employed the interpretable SHAP method on the best-performing model to clarify the contribution of each feature to fraud risk, thereby identifying key indicators of fraudulent behavior. Model performance was evaluated using accuracy, F1-score, and ROC-AUC.

4.3. Results and Analysis

4.3.1. Semantic Features of MD&A

Table 5. MD&A content-based corporate fraud detection model performance.

MD&A-based thematic modules	Accuracy	F1	ROC-AUC
1 *Overview*	**0.654**	**0.654**	**0.695**
2 *Core business analysis*	0.615	0.611	0.659
3 *Non-core business analysis*	0.603	0.603	0.639
4 *Asset and liability status*	0.619	0.616	0.658
5 *Investment activities*	0.589	0.551	0.632
6 *Significant asset & equity transactions*	**0.634**	**0.631**	**0.697**
7 *Analysis of major subsidiaries and affiliates*	0.619	0.613	0.696
8 *Structured entities*	0.613	0.613	0.637
9 *Future outlook*	**0.626**	**0.626**	**0.670**

Table 5 presents the fraud prediction performance of BERT models across nine MD&A thematic modules. The *Significant asset & equity transactions* module achieves the highest predictive performance (ROC-AUC = 0.697), followed by *Overview* (ROC-AUC = 0.695) and *Future outlook* (ROC-AUC = 0.670). This performance hierarchy can be explained by the informational value of each section: 1) *Significant asset and equity transactions* documents substantial transactions with direct financial implications; 2) The *Overview* synthesizes the company's comprehensive status, effectively encapsulating management's core messaging; and 3) *Future Outlook* conveys development expectations and profitability

projections, offering critical insights into management intentions and potential fraud motives. Conversely, the *Investment activities* module demonstrates relatively weaker predictive power (ROC-AUC = 0.632), suggesting its limited relevance to fraud risk assessment.

Semantic vectors (CLS outputs) from the three top modules—*Overview*, *Significant Asset and Equity Transactions*, and *Future Outlook*—were dimensionally reduced to remove redundancy and emphasize key information. These reduced vectors were evaluated with XGBoost, LightGBM, CatBoost, and logistic regression (LR); CatBoost performed best and was selected for the final model. Finally, the reduced semantic vector from the best-performing module (*Significant Asset and Equity Transactions*) was concatenated with other annual report features and corporate financial and organizational structure indicators to form a multi-source feature vector, which was fed into CatBoost (see Table 6).

4.3.2. Integrated-Data Financial Fraud Detection Model

Table 6 presents the results of corporate fraud prediction using three ensemble tree-based algorithms and LR, which incorporate features from multiple levels, including financial and organizational structure indicators, semantic features derived from MD&A, as well as tone, similarity, and readability. The ensemble tree algorithms consistently outperform the baseline LR model across all evaluation metrics. CatBoost achieves the highest performance, with ROC-AUC scores of 0.859 for models built on the *Significant Asset & Equity Transactions* module. The corresponding dimension of semantic features for the module is 7, demonstrating that even low-dimensional semantic representations derived from the MD&A section provide valuable information for detecting financial fraud.

Table 6. Performance of integrated-data fraud detection models.

Algorithm	MD&A-based thematic modules	Reduced Semantic features dimension	Accuracy	F1	ROC-AUC
CatBoost	1 *Overview*	**3**	**0.786**	**0.780**	**0.858**
	6 *Significant asset & equity transactions*	**7**	**0.798**	**0.793**	**0.859**
	9 *Future outlook*	**6**	**0.784**	**0.778**	**0.857**
XGBoost	1 *Overview*	**4**	**0.781**	**0.779**	**0.856**
	6 *Significant asset & equity transactions*	7	0.784	0.779	0.850
	9 *Future outlook*	4	0.780	0.775	0.853
LightGBM	1 *Overview*	**3**	**0.786**	**0.782**	**0.852**
	6 *Significant asset & equity transactions*	5	0.780	0.773	0.847
	9 *Future outlook*	3	0.774	0.768	0.849
LR	1 *Overview*	**8**	**0.677**	**0.678**	**0.743**
	6 *Significant asset & equity transactions*	10	0.666	0.667	0.730
	9 *Future outlook*	15	0.677	0.675	0.734

4.3.3. Feature Importance Analysis

We utilized the optimal CatBoost model for SHAP analysis, incorporating semantic features from the *Significant Asset & Equity Transactions* module alongside 41 additional features. Figure 3 presents the top 20 features ranked by SHAP values, where the X-axis represents feature importance and different colors distinguish between feature categories.

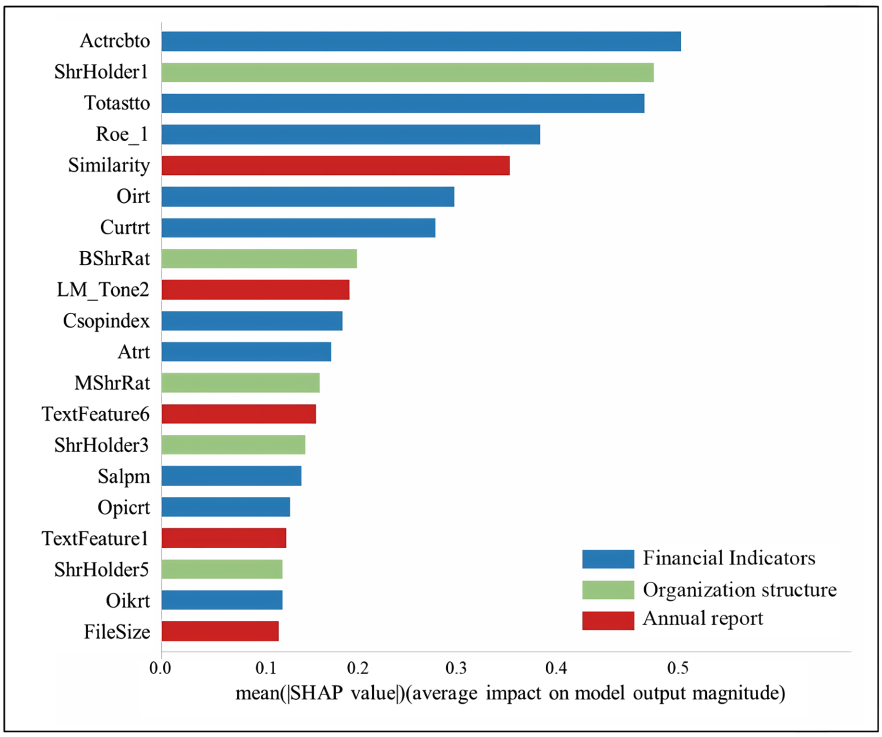

Figure 3. SHAP-based feature importance analysis (Top 20 features).

Among the top 20 features, financial indicators constitute 50%, with organization structure and annual report features each accounting for 25%. Financial indicators have a median rank of 7, significantly higher than the other two categories, underscoring their dominant role in fraud detection—particularly metrics related to operational efficiency (e.g., accounts receivable turnover, *Actrcbto*; total asset turnover, *Totastto*) and profitability (e.g., return on equity, *Roe_1*).

Within the realm of organization structure, the shareholding ratio of the largest shareholder (*ShrHolder1*) and the board's shareholding ratio (*BShrRat*) are particularly noteworthy. The former reflects the concentration of corporate power, while the latter indicates both the alignment of interests between the board and the company and the board's capacity for independent oversight and decision-making. These findings highlight the importance of power distribution and interest alignment in predicting corporate fraud risk.

Among annual report features, similarity (*Similariy*), tone (*LM_tone2*), and two semantic indicators ($TextFeature_6$ and $TextFeature_1$) rank among the most important variables. Similarity reveals year-over-year changes in report content,

while tone indicates management's confidence in performance and outlook. Although semantic features are less influential than tone or similarity, they still meaningfully aid in fraud detection, suggesting that annual reports encode information closely tied to corporate fraud.

4.3.4. Analysis of Impact Factors

The SHAP beeswarm plot (Figure 4) visually demonstrates the relationship between the top 20 features and financial fraud risk. The x-axis represents SHAP values, indicating both the magnitude and direction of each feature's impact. Positive SHAP values suggest an increased fraud risk, while negative values indicate a decreased risk.

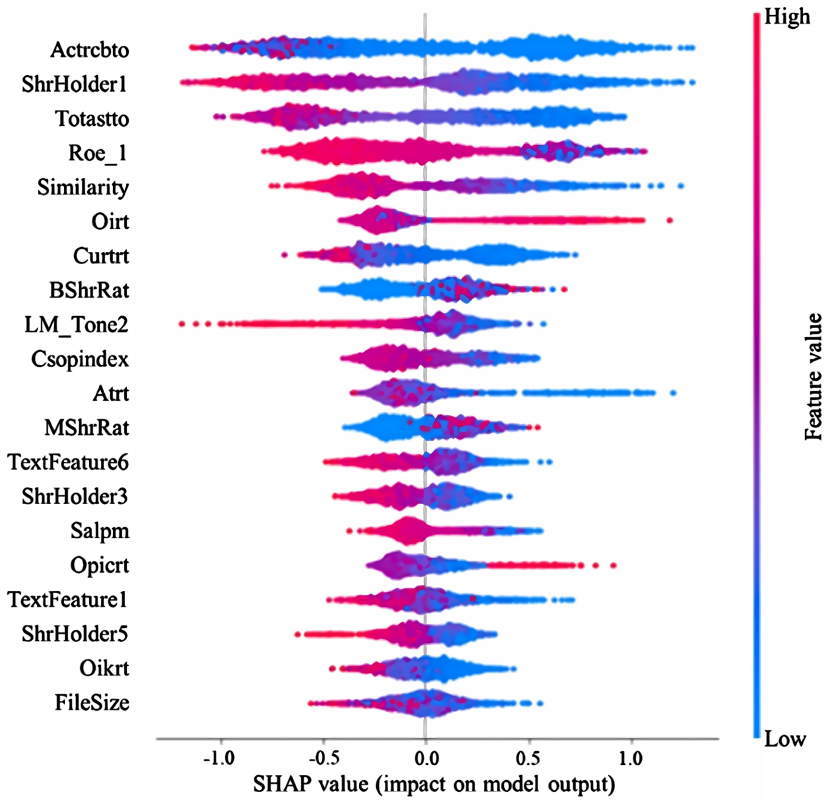

Figure 4. SHAP beeswarm plot of top 20 features.

5. Discussion

5.1. Integrated Modeling and Methodological Advances

This study demonstrates that integrating semantic extraction, data fusion, and multimodal feature incorporation can substantially improve corporate fraud prediction. The fusion-based model, which combines financial indicators, organizational structure variables, and annual report features, achieves a notable ROC-AUC of 0.859. While financial indicators remain the core predictors, the addition of organizational and textual information significantly enhances predictive accuracy and provides deeper insights into fraud detection.

Notably, ensemble tree models exhibit a clear advantage over traditional logistic regression, thanks to their ability to capture the complex, nonlinear patterns inherent in real-world fraud data. Consistently attaining ROC-AUC scores above 0.85, these ensemble approaches provide a more robust basis for fraud detection in complex environments.

The integration between textual and tabular data further strengthens model performance. Leveraging BERT for semantic extraction from MD&A sections, followed by dimensionality reduction via PCA, enhances both computational efficiency and the model's sensitivity to fraud-related signals. Analysis shows that disclosures in the *Overview*, *Significant asset & equity transactions*, and *Future outlook* sections of annual reports are particularly predictive of fraudulent activity, which suggests that companies engaged in fraud often focus manipulation on key narrative and transactional elements within their reporting.

5.2. Feature Interpretation and Forensic Insights

Beyond overall predictive success, model interpretability—enabled by SHAP analysis—provides important forensic insights into the relationship between individual features and fraud risk. Financial metrics stand out as primary indicators: reduced operating efficiency, profitability, and growth are all closely tied to a heightened risk of fraud. Regarding organizational structure, a higher concentration of shares held by major shareholders is negatively correlated with fraud risk, supporting the idea that concentrated ownership can deter fraudulent actions [33]. In contrast, board shareholding has a nuanced effect. Moderate ownership aligns directors' interests with shareholders', reducing agency problems and fraud risk. But excessive ownership can have the opposite effect: large stakes increase board control, weaken external oversight and the balancing role of independent directors, and raise information asymmetry and opportunities for manipulation. Directors with substantial holdings may conceal poor performance or manipulate financials to protect share prices or personal wealth. Empirical studies support this nonlinearity: Jensen and Meckling [34] argue managerial ownership reduces agency conflict only up to a point; Larcker *et al.* [35] find fraud risk rises when insider ownership exceeds an optimal range; and Yang *et al.* [36] report that excessive ownership alignment can introduce decision biases. Overall, board ownership and fraud risk appear to follow an inverted U-shaped relationship: moderate ownership improves governance, while excessive ownership concentrates power and increases fraud risk.

Linguistic and behavioral signals embedded in annual reports also add another layer of valuable insight. Key features such as year-over-year content similarity, tone, semantics, and readability help surface indications of hidden or manipulative practices. Large changes in report content or a shift away from an optimistic tone often accompany increased fraud risk, perhaps as management attempts to obscure irregularities. Module-level analysis highlights the particular importance of the *Overview*, *Significant asset & equity transactions* ,and *Future outlook*, while

variations in readability provide additional red flags. Overall, how management adjusts report language and complexity can reflect underlying intentions and psychological states. These patterns in annual report content—whether in narrative style, explanation, or structure—offer meaningful clues, allowing for earlier and more accurate identification of potential corporate fraud.

5.3. Research Contributions and Implications

This study presents new evidence supporting the effectiveness of machine learning ensemble tree algorithms in identifying corporate financial fraud. By employing an innovative method for semantic feature extraction, joined with feature dimension reduction, we modularized extensive annual report texts. This approach not only facilitated the seamless integration of textual and tabular information but also offered a practical solution for managing complex multimodal financial data.

Such an approach has meaningful implications for practice: extracting module-level risk signals lets regulators and practitioners efficiently screen many disclosures and target fraud-related content. For example, if the *Future Outlook* text closely repeats the prior year while financial analysis shows a sharp drop in profitability and cash flow, the system can auto-trigger a risk alert to prompt a focused investigation, enabling earlier detection and more precise allocation of regulatory resources. The study also extends risk identification to linguistic behavior, examining how tone, and readability reflect management's psychological and behavioral tendencies during report preparation. Quantifying these linguistic features aids investors and regulators in assessing corporate transparency and integrity culture and provides a basis for regulators to refine disclosure-review standards, making oversight more scientific, dynamic, and evidence-based.

6. Limitations and Future Directions

While this study provides insight into corporate fraud prediction, certain limitations remain that highlight potential areas for further investigation. The current model focuses solely on assessing the likelihood of fraudulent behavior, without distinguishing between specific types of fraud. However, in practice, corporate fraud takes diverse forms and exhibits considerable complexity. Future research could address this by developing automated methods to identify and classify different types of fraud, enabling more precise and actionable predictions.

Another important aspect concerns the processing of annual report texts. Although the study employs modularization and semantic modeling to preserve critical information, capturing the full depth of semantic meaning remains a challenge. More advanced approaches, such as the use of attention mechanisms, may help pinpoint keywords and phrases that are particularly relevant for fraud detection, thereby refining the model's sensitivity to subtle textual cues.

Additionally, the analysis presented here is based on data from a single year, which may not fully capture the dynamic and cumulative nature of financial fraud. Many fraudulent activities unfold over longer periods, making the exploration of

multi-year data and dynamic feature extraction an important direction for improving model accuracy and robustness.

Addressing these limitations will contribute to the practical advancement of corporate fraud detection models. By expanding fraud type classification, enhancing text analysis methods, and incorporating temporal changes, future research can provide stronger tools for identifying and mitigating financial fraud risks.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Wu, X. and Du, S. (2022) An Analysis on Financial Statement Fraud Detection for Chinese Listed Companies Using Deep Learning. *IEEE Access*, **10**, 22516-22532. https://doi.org/10.1109/access.2022.3153478

[2] Sun, J., Fujita, H., Chen, P. and Li, H. (2017) Dynamic Financial Distress Prediction with Concept Drift Based on Time Weighting Combined with Adaboost Support Vector Machine Ensemble. *Knowledge-Based Systems*, **120**, 4-14. https://doi.org/10.1016/j.knosys.2016.12.019

[3] Fukas, P., Rebstadt, J., Menzel, L. and Thomas, O. (2022) Towards Explainable Artificial Intelligence in Financial Fraud Detection: Using Shapley Additive Explanations to Explore Feature Importance. In: Franch, X., Poels, G., Gailly, F. and Snoeck, M., Eds., *Advanced Information Systems Engineering*, Springer, 109-126. https://doi.org/10.1007/978-3-031-07472-1_7

[4] Ashtiani, M.N. and Raahemi, B. (2022) Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review. *IEEE Access*, **10**, 72504-72525. https://doi.org/10.1109/access.2021.3096799

[5] Kirkos, E., Spathis, C. and Manolopoulos, Y. (2007) Data Mining Techniques for the Detection of Fraudulent Financial Statements. *Expert Systems with Applications*, **32**, 995-1003. https://doi.org/10.1016/j.eswa.2006.02.016

[6] Xiong, F.J. and Zhang, L.P. (2016) Risk Identification and Evidence Collection of Financial Fraud on the Listed Companies. *Research on Economics and Management*, **37**, 138-144.

[7] Qian, P. and Luo, M. (2015) Predicting Accounting Fraud in China. *Accounting Research*, **7**, 18-25, 96.

[8] Liu, Y.Q., Wu, B. and Zhang, M. (2022) Financial Fraud Recognition Model and Application. *Journal of Quantitative & Technological Economics*, **39**, 152-175.

[9] Tan, J.H. and Wang, X.Y. (2022) Corporate Fraud and Manipulation of Annual Report Text Information. *China Soft Science*, **3**, 99-111.

[10] Xu, C. and Zhang, Y.M. (2021) Can Managers' Tone Management Predict the Financial Fraud Risk: Based on MD&A Forward-looking Text Information. *Science and Management*, **41**, 73-81.

[11] Wang, K.M., Wang, H.J., Li, D.D. and Dai, X.Y. (2018) Complexity of Annual Report and Management Self-Interest: Empirical Evidence from Chinese Listed Firms. *Journal of Management World*, **34**, 120-132, 194.

[12] Zhang, Y., Liu, T. and Li, W. (2024) Corporate Fraud Detection Based on Linguistic Readability Vector: Application to Financial Companies in China. *International Review of Financial Analysis*, **95**, Article ID: 103405.

https://doi.org/10.1016/j.irfa.2024.103405

[13] Brown, N.C., Crowley, R.M. and Elliott, W.B. (2020) What Are You Saying? Using *topic* to Detect Financial Misreporting. *Journal of Accounting Research*, **58**, 237-291. https://doi.org/10.1111/1475-679x.12294

[14] Craja, P., Kim, A. and Lessmann, S. (2020) Deep Learning for Detecting Financial Statement Fraud. *Decision Support Systems*, **139**, Article ID: 113421. https://doi.org/10.1016/j.dss.2020.113421

[15] Liu, W., Wang, Z. and Zhang, X. (2025) Research on Financial Fraud Detection by Integrating Latent Semantic Features of Annual Report Text with Accounting Indicators. *Journal of Accounting & Organizational Change*, **21**, 841-866. https://doi.org/10.1108/jaoc-06-2024-0199

[16] Hajek, P. and Henriques, R. (2017) Mining Corporate Annual Reports for Intelligent Detection of Financial Statement Fraud—A Comparative Study of Machine Learning Methods. *Knowledge-Based Systems*, **128**, 139-152. https://doi.org/10.1016/j.knosys.2017.05.001

[17] Ali, A.A., Khedr, A.M., El-Bannany, M. and Kanakkayil, S. (2023) A Powerful Predicting Model for Financial Statement Fraud Based on Optimized XGBoost Ensemble Learning Technique. *Applied Sciences*, **13**, Article 2272. https://doi.org/10.3390/app13042272

[18] Zhao, Z. and Bai, T. (2022) Financial Fraud Detection and Prediction in Listed Companies Using SMOTE and Machine Learning Algorithms. *Entropy*, **24**, Article 1157. https://doi.org/10.3390/e24081157

[19] Yadav, A.K.S. (2024) Financial Statement Fraud Detection Using Optimized Deep Neural Network. In: Asirvatham, D., Gonzalez-Longatt, F.M., Falkowski-Gilski, P. and Kanthavel, R., Eds., *Evolutionary Artificial Intelligence*, Springer, 131-141. https://doi.org/10.1007/978-981-99-8438-1_10

[20] Bhattacharya, I. and Mickovic, A. (2024) Accounting Fraud Detection Using Contextual Language Learning. *International Journal of Accounting Information Systems*, **53**, Article ID: 100682. https://doi.org/10.1016/j.accinf.2024.100682

[21] Wang, G., Ma, J. and Chen, G. (2023) Attentive Statement Fraud Detection: Distinguishing Multimodal Financial Data with Fine-Grained Attention. *Decision Support Systems*, **167**, Article ID: 113913. https://doi.org/10.1016/j.dss.2022.113913

[22] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., *et al.* (2020) From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, **2**, 56-67. https://doi.org/10.1038/s42256-019-0138-9

[23] Zeng, Q.S., Zhou, B., Zhang, C. and Chen, X.Y. (2018) Annual Report Tone and Insider Trading: "Consistent" or "Deceptive"? *Journal of Management World*, **34**, 143-160.

[24] Li, J., Li, N., Xia, T. and Guo, J. (2023) Textual Analysis and Detection of Financial Fraud: Evidence from Chinese Manufacturing Firms. *Economic Modelling*, **126**, Article ID: 106428. https://doi.org/10.1016/j.econmod.2023.106428

[25] Qian, A. and Zhu, D. (2020) Financial Report Textual Similarity and Likelihood of Regulatory Penalties: Based on the Empirical Evidence of Textual Analysis. *Accounting Research*, **9**, 44-58.

[26] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[27] Shapley, L.S. (1953) 17. A Value for n-Person Games. In: Kuhn, H.W. and Tucker, A.W., Eds., *Contributions to the Theory of Games* (*AM*-28), *Volume II*, Princeton University Press, 307-318. https://doi.org/10.1515/9781400881970-018

[28] Cohen, J.R., Krishnamoorthy, G. and Wright, A. (2004) The Corporate Governance Mosaic and Financial Reporting Quality. *Journal of Accounting Literature*, **23**, 87-152.

[29] Dechow, P.M., Ge, W., Larson, C.R. and Sloan, R.G. (2011) Predicting Material Accounting Misstatements. *Contemporary Accounting Research*, **28**, 17-82. https://doi.org/10.1111/j.1911-3846.2010.01041.x

[30] Tamimi, N. and Sebastianelli, R. (2017) Transparency among S&P 500 Companies: An Analysis of ESG Disclosure Scores. *Management Decision*, **55**, 1660-1680. https://doi.org/10.1108/md-01-2017-0018

[31] Yu, E.P., Luu, B.V. and Chen, C.H. (2020) Greenwashing in Environmental, Social and Governance Disclosures. *Research in International Business and Finance*, **52**, Article ID: 101192. https://doi.org/10.1016/j.ribaf.2020.101192

[32] Chen, G.J., Lin, H. and Wang, L. (2005) Corporate Governance, Reputation Mechanism and the Behavior of Listed Firms in Committing in Fraud. *Nankai Business Review*, **6**, 35-40.

[33] Pan, W.B. and Hong, Y. (2017) Characteristics of Companies with Financial Fraud based on Time-Dependent COX Model. *Journal of University of Science and Technology of China*, **47**, 255-261.

[34] Jensen, M.C. and Meckling, W.H. (1976) Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure. *Journal of Financial Economics*, **3**, 305-360. https://doi.org/10.1016/0304-405x(76)90026-x

[35] Larcker, D.F., Richardson, S.A. and Tuna, I. (2007) Corporate Governance, Accounting Outcomes, and Organizational Performance. *The Accounting Review*, **82**, 963-1008. https://doi.org/10.2308/accr.2007.82.4.963

[36] Yang, Q., Yu, L. and Chen, N. (2009) Board Characters and Financial Fraud: Empirical Evidence from Chinese Listed Companies. *Accounting Research*, **7**, 64-70, 96.