Journal of
Knowledge Economy Studies

https://kes.hmu.ac.ir

Hazrat-e Masoumeh University

Developing a Deep Learning–Based Model for Predicting and Detecting Fraud in Financial Statements

Article Type:
Research Article

Narges Mehrabi Hashtchin[1]
Faculty of Social Sciences and Economics, Alzahra University, Tehran, Iran
E-mail: n.mehrabi@alzahra.ac.ir

Gholamreza Soleymani Amiri[2*]
Faculty of Social Sciences and Economics, Alzahra University, Tehran, Iran
Corresponding Author
E-mail: gh.soleymani@alzahra.ac.ir

Received: 27 January 2026
Revised: 17 February 2026
Accepted: 17 March 2026
Available Online: 18 March 2026

Abstract

This study developed a data-driven framework for financial statement fraud detection by benchmarking machine learning, deep learning, and hybrid classifiers under a unified, leakage-resistant evaluation protocol. The fraud cases were identified from the U.S. Securities and Exchange Commission's Accounting and Auditing Enforcement Releases (AAERs) and matched with Compustat data over 1991–2014, producing 122,526 firm-year observations, including 902 confirmed fraud cases. Four structured-input configurations were evaluated: 28 raw financial statement items, 14 financial ratios, their combined set (28+14), and a parsimonious seven-feature subset (six ratios plus Altman's Z-score). The features were selected using minimum redundancy–maximum relevance (mRMR), class imbalance was addressed via cost-sensitive learning, and performance was assessed with a firm-level 80/20 split and stratified group-based five-fold cross-validation within training. The empirical results indicated that deep and hybrid models consistently outperform classical tabular baselines, reflecting non-linear and interaction-driven fraud signals. The Transformer achieved the most stable and highest overall performance, reaching 0.98898 accuracy and a 0.51087 F1-score under the seven-feature configuration. The combined raw-item and ratio inputs outperformed the ratios alone, implying incremental predictive value in raw accounting items, while the best overall outcomes were obtained with parsimonious seven-feature subset. Collectively, the findings supported the study's hypotheses and demonstrated the effectiveness of attention-based modeling for financial statement fraud detection.

Keywords

Artificial intelligence, Fraudulent financial statements, Machine learning, Deep learning, Imbalanced datasets.

Cite this article: Mehrabi Hashtchin, N. & Soleymani Amiri, Gh., (2026). Developing a deep learning–based model for predicting and detecting fraud in financial statements. *Journal of Knowledge Economy Studies (JKES)*, 3(1), 7-24. DOI: http//doi.org/10.22034/kes.2026.2083828.1098

Online ISSN: 3060-7329

Introduction

Financial statement fraud remains a persistent concern for capital markets and financial oversight because intentional misreporting can distort valuation, impair resource allocation, and erode confidence in audited reporting. In practice, fraud detection is difficult because fraudulent behavior is typically adaptive, multi-faceted, and intentionally concealed within otherwise legitimate reporting processes, while audit procedures operate under time and evidence constraints. Consequently, an active research agenda has emerged around data-driven fraud-risk screening systems that can help auditors and regulators prioritize firm-year investigations and allocate attention to the highest-risk cases.

Recent research in fraud analytics has advanced along three complementary directions: (1) improved learning architectures for complex, non-linear patterns, (2) richer representations and modalities beyond conventional tabular inputs, and (3) increased emphasis on explainability and operational deployability. First, a growing body of evidence supported using modern machine learning ensembles and meta-classifiers for financial fraud prediction, where combining heterogeneous learners can yield stronger and more stable performance than individual models (Achakzai & Juan, 2022; Azim Mim et al., 2024; Cao et al., 2023).

This perspective is consistent with broader fraud-detection works showing that ensemble strategies, including soft-voting and boosting variants, can improve robustness under noisy signals and rare-event settings (Ahmed et al., 2025; Azim Mim et al., 2024).

Second, the scope of information sources used for fraud detection has broadened beyond the handcrafted ratio sets. In financial reporting contexts, recent studies demonstrated that unstructured disclosures contain incremental predictive content. Contextual language learning approaches based on Transformer-style architectures (e.g., BERT) can extract deception-related signals from narrative sections of annual reports, thereby improving accounting fraud detection performance and increasing the yield of detected fraudulent observations under limited investigative capacity (Bhattacharya & Mickovic, 2024). In parallel, work that constructs multi-dimensional firm portraits by combining structured indicators with broader firm characteristics and unstructured data further supports the view that richer representations can improve the fraud prediction and, importantly, can be paired with post-hoc explanation methods to identify salient risk drivers (Zhang et al., 2025). These developments collectively suggest that fraud-risk systems should be evaluated with attention to both predictive performance and the evidentiary value of the underlying signals.

Third, explainability has become a central requirement in high-stakes financial applications. Several recent contributions explicitly argued that many prior fraud-detection models remain difficult to interpret, which limits their acceptance in audit and regulatory environments. To address this, explainable frameworks have been proposed that embed accounting-relevant structure into the modeling pipeline. For instance, a two-layer knowledge-graph approach for financial statement fraud detection models

semantic and articulation relationships among statement items to enable interpretable pattern mining and credible fraud assertions (Cai & Xie, 2024). Complementary research proposed interpretable graph-learning formulations that jointly address severe imbalance and provide built-in interpretability outputs, reflecting a broader movement toward accountable and deployable fraud analytics (Lu et al., 2026). Related work on graph-based fraud detection further indicated that relational structure and dependency modeling can improve detection capability in fraud settings characterized by complex interactions (Shao et al., 2026).

Across these lines of research, two technical challenges recur. The first is extreme class imbalance. Fraud is typically rare, so high overall accuracy can coexist with limited minority-class detection quality. Prior work in financial misstatement settings highlighted the importance of cost-sensitive learning and imbalance-aware modeling to reduce the risk of accuracy inflation and to improve performance on fraud-related outcomes (Kim et al., 2016; Lu et al., 2026). The second challenge is operational realism. Fraud tactics evolve and detection systems may need to function under limited labels, streaming data and scale constraints. Studies on transaction fraud increasingly emphasized online or adaptive frameworks, distributed learning, unsupervised anomaly detection, and hybrid sampling to improve practicality under real-world conditions (Ahmed et al., 2025; Karnavou et al., 2025; Lei et al., 2023; Narayana Gorle & Panigrahi, 2026). Although these works often targeted transaction-level fraud, they reinforced methodological expectations that are increasingly relevant for financial statement fraud. Stable generalization, careful evaluation under realistic splits, and transparent reasoning or explanation mechanisms were given due attention.

Despite substantial progress, an open and practically relevant question remains in financial statement fraud research regarding which structured inputs are most informative and deployable: raw financial statement items, derived financial ratios, or their combination. In addition, the literature increasingly calls for solutions that not only achieve strong predictive discrimination but also remain usable for practitioners through parsimonious feature sets and interpretable outputs (Cai & Xie, 2024; Zhang et al., 2025). Motivated by these directions, this study developed and evaluated a data-driven framework for financial statement fraud detection that systematically compares classical machine learning methods with modern deep and hybrid architectures under firm-level splitting and stratified group-based cross-validation. The researchers examined four input scenarios including raw financial variables, financial ratios, their combination, and a compact seven-feature subset to assess the incremental value of raw versus ratio-based information and to identify a parsimonious configuration suitable for fraud-risk monitoring and audit planning. To address the severe rarity of fraud cases, we incorporated imbalance-aware learning through cost-sensitive training, and we emphasized stability across validation folds to support robust model comparisons in applied settings (Kim et al., 2016; Lu et al., 2026).

Literature Review

Financial statement fraud is a major topic in accounting and auditing and refers to deliberate and deceptive actions undertaken to obtain benefits (Arboleda et al., 2018). Fraud in financial reporting refers to the misstatement of financial reports and the presentation of a distorted picture of the business entity (Vakilifard et al., 2009).

The aim of this study was to apply deep learning to predict and detect fraud in financial statements based on financial ratios and to compare its effectiveness with raw financial statement data. The fraud cases were extracted from material misstatements reported in the U.S. Securities and Exchange Commission's AAERs, because these releases rely on publicly available, low-cost data and enable fair comparison with prior studies. Drawing on the literature (Ahmed & Courtis, 2015; Kanapickienė & Grundienė, 2015), financial ratios due to their grounding in accounting expertise, simplicity, and widespread use can serve as more effective indicators for fraud detection. This study employed the validated dataset of Bao et al.'s study (2020), which includes both raw features and financial ratios, and by comparing these inputs, identified the most effective feature set for improving the fraud detection accuracy.

Table 1 synthesizes representative prior studies and highlights heterogeneity in data domains, input representations, model families, and evaluation protocols, which complicates direct cross-study inference.

Table 1.
Representative Studies and Design Choices in Fraud Detection Research

Study	Data domain	Inputs	Model family	Validation design
Dechow et al. (2011)	Financial statements	Ratios / engineered	Benchmark statistical	Holdout
Cecchini et al. (2010)	Financial statements	Raw items	Classical ML (SVM)	Holdout
Bao et al. (2020)	Financial statements	Raw items (+ ratio benchmarks)	ML / ensemble	Holdout
Bhattacharya & Mickovic (2024)	Disclosures	Text sequences	Transformer/NLP (BERT)	Holdout
Cai & Xie (2024)	Financial statements	Graph	Explainable KG + pattern mining	Benchmark evaluation
Zhang et al. (2025)	Financial statements	Structured + broader signals	ML + explainability (SHAP)	Holdout
Lu et al. (2026)	Financial fraud	Graph-structured features	Interpretable, imbalance-aware graph learning	Benchmark evaluation
Karnavou et al. (2025)	Transactions	Transaction variables	Unsupervised anomaly detection + SHAP	Operational/benchmark; not firm-year splitting
Lei et al. (2023)	Transactions	Transaction variables	Distributed DNN	Holdout

(Source: The Researcher's Findings)

Despite growing interest in financial statement fraud detection, the literature still lacks a rigorous, apples-to-apples comparison of classical, deep, and hybrid models under a shared experimental protocol and leakage-resistant validation (e.g., the firm-level splitting

and group-based cross-validation). Moreover, prior studies rarely isolated the incremental value of raw accounting items versus financial ratios or their combination making it difficult to disentangle algorithmic gains from feature-construction effects. Finally, evidence remains limited on whether a parsimonious feature subset can retain minority-class detection performance under extreme imbalance while improving practical deployability for audit planning and continuous monitoring.

Deep Learning and Machine Learning Methods for Fraud Detection

In recent years, the adoption of machine learning and advanced models particularly in the field of financial fraud detection has led to remarkable progress. In this study, a broad set of classical machine learning methods and advanced deep learning architectures were employed to analyze and accurately classify financial data. First, classical models including logistic regression, decision tree, random forest, and support vector machine were applied to establish baseline performance and evaluate traditional approaches.

In the deep learning component, a spiking classifier was first used, which due to its dynamic temporal nature can identify behavioral patterns that vary across financial periods (Jeyasothy et al., 2024). Next, residual networks ResNet-18 and ResNet-15 were employed. Then, a one-dimensional capsule network, as well as an efficient capsule network architecture, and a highly efficient and low-parameter variant of capsule networks were utilized for analyzing the financial data (Mazzia et al., 2021). In addition, the CAT-Net model was applied, integrating convolutional neural network layers, channel attention, and Transformers, enabling it to capture both local features of financial statements and long-term dependencies among different items (Islam et al., 2024). Finally, an advanced ensemble learning approach based on a Type-3 fuzzy decision-making system was used to intelligently and integratively combine the outputs of different models (Mehrabi Hashjin et al., 2024). Unlike simple aggregation methods, this system models the inherent uncertainty in financial data and fraud consequences, thereby enabling more accurate final decisions. To optimally tune parameters and fuzzy rules, an Improved Chaos Game Optimization algorithm which is an enhanced version of the Chaos Game Optimization algorithm was employed.

In detecting fraudulent financial transactions, Sai et al. (2023) evaluated LightGBM and XGBoost alongside deep neural networks. By applying measures such as hyperparameter tuning and imbalance-handling techniques, they showed that LightGBM achieved the best performance (accuracy ≈ 98.3%, F1 ≈ 0.70, AUC ≈ 0.96). Bao et al. (2020) developed a model for predicting fraud in U.S. publicly listed companies by using raw accounting numbers (rather than financial ratios) and applying ensemble learning. Their model significantly outperformed two widely used benchmarks: (1) the financial ratio–based logistic regression model proposed by Dechow et al. (2011), and (2) the SVM model based on ratios derived from raw data in Cecchini et al.'s study (2010).

Research Hypotheses

Based on the theoretical foundations presented, the research hypotheses were formulated as follows:

- **Hypothesis 1:** Deep learning models outperform classical machine learning baselines in detecting financial statement fraud.
- **Hypothesis 2:** Ensemble learning improves performance relative to single-model baselines under severe class imbalance.
- **Hypothesis 3:** A parsimonious feature subset can achieve performance comparable to or better than full feature sets, particularly in terms of F1-score.

Method

The present study was conducted to develop and evaluate methods for detecting fraud in financial statements. In terms of purpose, it is an applied study, and in terms of nature, it is descriptive. Moreover, because it uses historical data, it falls within the category of empirical and quasi-experimental research. The dependent variable is financial statement fraud. To enable comparability with prior studies, material misstatements reported in the U.S. SEC's AAERs were used as fraud cases, and the independent variables are financial ratios based on the dataset of Bao et al. (2020). Finally, by integrating the data, a new model with 24 features (23 financial ratios and one control variable) was developed for predicting and detecting fraud, allowing the performance of machine learning and deep learning methods to be evaluated in a manner comparable to previous research.

Research Variables

The dependent variable of the study is financial statement fraud, which is qualitative and nominal in nature and was coded as a binary variable (fraud firm = 1, non-fraud firm = 0). The fraud cases were identified based on material misstatements reported in the SEC's AAERs and in accordance with the dataset of Bao et al. (2020). The model inputs included 24 features, consisting of 23 financial ratios as independent variables and one control variable.

The independent variables were adopted from prior studies (e.g., Bao et al., 2020). After aligning the selection criteria with manifestations of fraudulent financial reporting, 23 financial ratios were selected and used, including: current ratio; current assets to total assets; fixed assets to total assets; working capital to total assets; total liabilities to total assets; retained earnings to total assets; sales to total assets; net income to total assets; gross profit to total assets; total liabilities to shareholders' equity; long-term debt to shareholders' equity; cost of goods sold to sales; gross profit to sales; net income to sales; operating expenses to sales; operating income to sales; net income to shareholders' equity; financial expenses to total liabilities; accounts receivable to sales; inventories to sales; cash and cash equivalents to total assets; inventories to total liabilities; and net income to gross profit.

Due to the role of financial distress in creating incentives and weakening the control

environment, the Altman Z-score (1983) was included as a control variable in order to control for and examine the association between fraud and financial distress (Etemadi & Zolghi, 2013).

$$Z = 0.717x_1 + 0.847x_2 3.1x_3 + 0.42x_4 + 0.998x_5 \quad (1)$$

In Equation (1), x_1 denotes working capital to total assets, x_2 retained earnings to total assets, x_3 earnings before interest and taxes (EBIT) to total assets, x_4 book value of shareholders' equity to total liabilities, and x_5 sales to total assets.

In this study, advanced Shannon cross-entropy–based approach was employed for optimal feature selection in order to eliminate redundant information, improve model accuracy, and enhance computational efficiency. Ultimately, seven final features were selected. Moreover, using the dataset comprising 28 raw accounting variables and 14 financial ratios from Bao et al. (2020), the ratio-based results of this study were compared with the findings reported in that work. This dataset combines the raw data of Cecchini et al.'s study (2010) with financial ratios adapted from Dechow et al.'s study (2011) and Cecchini et al.'s study (2010). The full list of input variables is provided in Table 2.

Table 2.
The List of Input Variables

Input data used in Bao et al.'s study (2020, p. 213)			**Input data used in the present study**
The 28 raw financial variables adopted from Cecchini et al.'s study (2010):		Fourteen financial ratios (financial changes): including 11 ratios derived from Dechow et al.'s study (2011) and three ratios derived from Cecchini et al.'s study (2010).	The seven features used in this study were selected based on the theoretical literature and consisted of six financial ratios plus one control variable, and the Altman Z-score (1983).
- Short-Term Investments - Current Liabilities - Total Liabilities - Net Income - Property, Plant and Equipment - Preferred Stock - Retained Earnings - Accounts Receivable - Sales (Revenue) - Sale of Stock (Equity Issuance) - Taxes Payable - Total Taxes - Interest Expense - Inventory - Other Investments	- Current Liabilities - Total Assets - Common Shareholders' Equity - Cash and Short-Term Investments - Cost of Goods Sold - Common Shares Outstanding - Current Liabilities - Long-Term Debt Issuance - Total Long-Term Debt - Depreciation - Income Before Extraordinary Items	- Change in Common Stock - Change in Cash Margin - Change in Return on Assets - Securities Issuance - Book-to-Market Ratio - Discretionary Accruals Index - Adjusted Return on Equity - Soft Assets (excluding tangible assets) - Earnings Before Interest and Taxes Change in Free Cash Flow	- Current Ratio - Total Debt-to-Equity Ratio - Cost of Goods Sold to Sales Ratio - Operating Profit to Sales Ratio - Net Income to Shareholders' Equity Ratio - Financial Expenses to Total Debt Ratio - Financial Distress (Altman Z-score, 1983)

(Source: The Researcher's Findings)

Data Collection Method

The data used in this study were obtained from the Compustat accounting database via the dataset employed by Bao et al. (2020), which is publicly available on GitHub under the directory JarFraud/FraudDetection[1].

Statistical Sample

The statistical sample of this study included all firms listed on U.S. stock exchanges over the period 1991–2014, comprising a total of 146,045 observations. After removing missing data and applying preprocessing procedures, the final sample was reduced to 122,526 observations, including 121,624 observations in the non-fraud class and 902 observations in the fraud class.

Research Procedure

After collecting the financial ratios from the database, a preprocessing stage was conducted by removing missing values and preparing the data (including transforming them into images). The dataset was then split into training and testing sets. Next, machine learning–based and deep learning–based models were trained, and their hyperparameters were tuned to minimize error. Their performance was evaluated using the specified assessment metrics. Finally, the model with the highest accuracy was identified as the best-performing model for predicting and detecting the fraud.

Data Preprocessing Process

Before feeding the data into the machine learning models, a standardized preprocessing pipeline was applied. Specifically, observations containing missing values in any input variable were removed via listwise deletion. All predictors were then z-score standardized to place them on a common scale prior to model training and evaluation. Finally, we employed the mRMR feature selection method (Peng et al., 2005) to retain variables with the highest relevance to fraud identification while mitigating redundancy among predictors.

Converting Data into Images for Deep Learning

To enable the use of image-based deep architectures (e.g., ResNet-18, ResNet-50, and capsule networks), we transformed the tabular inputs into a two-dimensional time–frequency representation using the continuous wavelet transform (CWT) implemented in MATLAB. For each firm-year observation, the feature vector was treated as a short one-dimensional sequence. We computed the CWT under MATLAB's default parameterization, which employed the analytic Morse wavelet (symmetry parameter $\gamma = 3$, time-bandwidth product $P2 = 60$) with 10 voices per octave. The admissible scale range was selected automatically, and coefficients were L1-normalized. The transform yielded a complex coefficient matrix, from which we constructed the input image as the magnitude scalogram, $S = |cwt(x)|$. To ensure a uniform input resolution across observations and compatibility with standard CNN backbones, we resized the scalogram via interpolation (imresize) to a fixed image size (e.g., 224 × 224).

1. https://github.com/JarFraud/FraudDetection

Importantly, we did not replicate (tile) the original feature values, as repetition may introduce artificial periodicity and does not add information. Instead, resizing the scalogram preserved the original signal content while providing a consistent image representation for downstream training and evaluation.

Data Splitting

In this study, the data were first split into 80% for training and 20% for testing. To prevent information leakage and ensure a fair evaluation, the split was performed using a random, group-based partitioning strategy (based on the firm identifier), such that all observations for each firm were placed entirely in either the training set or the test set. Then, to reduce the risk of overfitting and obtain a more reliable estimate of the model performance, five-fold stratified, group-based cross-validation was applied within the training set; that is, while maintaining group independence in each fold, the class distribution (fraud/non-fraud) was kept as consistent as possible across folds.

Modeling Methods

To ensure that the empirical comparisons are conceptually motivated rather than a purely technical benchmark, we evaluated a set of model families that represent distinct learning mechanisms relevant to financial statement fraud detection. First, the classical tabular baselines logistic regression, the decision tree, random forest, and the support vector machine were included as established reference methods for structured accounting data. Second, deep feature-extraction models, a spiking neural network, and a one-dimensional convolutional neural network (1D-CNN) were considered to capture the non-linear patterns and local dependencies in the input series. Third, to examine whether the hierarchical representation learning benefits the fraud detection under an image-based representation, we evaluated CWT-derived image models, including ResNet-18, ResNet-50, a one-dimensional capsule network, and an efficient capsule network. Fourth, an attention-based Transformer architecture was included to model the global interactions across the input dimensions. Finally, we employed an ensemble learning strategy to aggregate complementary learners and improve robustness and performance stability under severe class imbalance.

Model Evaluation

The evaluation metrics used in this study included accuracy, precision, recall (sensitivity), specificity, and the F1-score.

Findings

Addressing the Class Imbalance Using Cost-Sensitive Learning

To address the severe class imbalance (the small number of fraud cases relative to non-fraud cases), a cost-sensitive learning approach, or class weighting, was applied throughout the entire modeling process. Accordingly, for the deep learning models, a weighted binary cross-entropy loss function was used:

$$L = -\left(w_{pos}\, y \log(p) + w_{neg}\,(1-y)\log(1-p)\right) \quad (2)$$

where $y \in \{0,1\}$ is the true label and p is the predicted probability for the fraud class (the positive class). The positive-class weight was determined based on the class imbalance ratio:

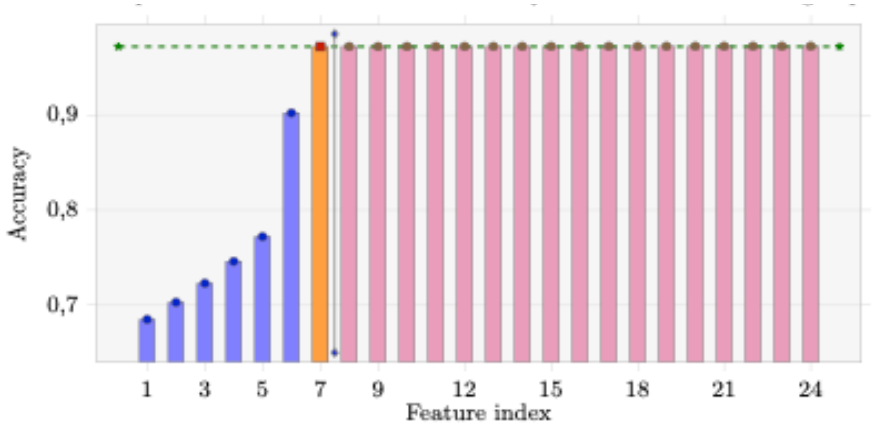
$$r = \frac{N_{neg}}{N_{pos}} \approx 134.8, \quad w_{pos} = \sqrt{r} \approx 11.6, \quad w_{neg} = 1 \tag{3}$$

Choosing $\sqrt{r}$ instead of r was intended to avoid overweighting the minority class and to maintain the stability of the learning process. For classical models, whose loss functions are not necessarily based on binary cross-entropy, the same idea was applied in an equivalent and implementable manner by incorporating class weights directly into the learning mechanism. Specifically, for logistic regression and other probabilistic models, a weighted cost function (i.e., class weighting) was used. For the support vector machine, the class weights were introduced into the error-penalty term so that misclassifications of the fraud class incurred a higher cost. For the decision tree and random forest models, the class weights were incorporated into split criteria and node impurity calculations (e.g., the Gini index or entropy), such that errors associated with the minority class exerted greater influence on split selection and the final model structure.

Feature Selection

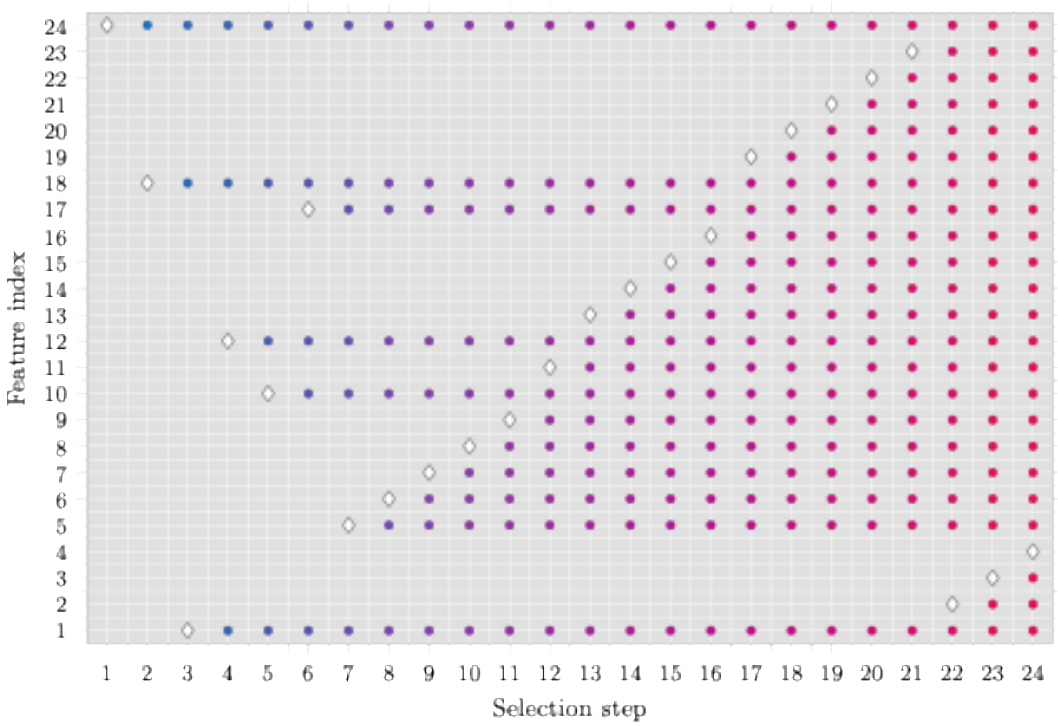
In this study, feature selection was performed using the mRMR method proposed by Peng et al. (2005). This method is based on Shannon mutual information and is designed to maximize the relevance of features to the target variable while minimizing redundancy among predictors. To determine the optimal number of input features, we conducted an incremental SVM-based evaluation, where features were added sequentially according to the mRMR ranking and the corresponding classification performance was tracked. Figure 1 reports the SVM classification accuracy as a function of the number of selected features, which provides an empirical basis for selecting a compact yet informative subset. Figure 2 visualizes the mRMR selection order across 24 steps and the cumulative selection path, illustrating how the candidate variables progressively enter the feature set. Based on the trade-off between predictive performance and parsimony, seven features were ultimately retained as the final input feature set.

Figure 1.
The Bar Chart of SVM Classification Accuracy versus Feature Index



(Source: The Researcher's Findings)

Figure 2.
The Feature Selection Order across 24 Steps

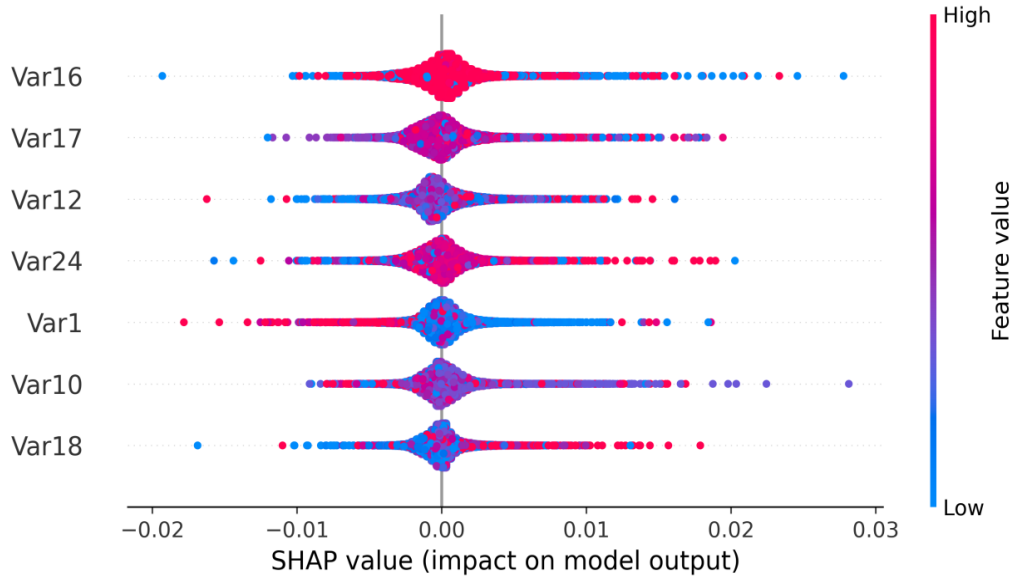


(Source: The Researcher's Findings)

The Model Interpretability and SHAP-based Explanation

To enhance the interpretability in a high-stakes setting such as financial statement fraud detection, we provided a model-based explanation of the Transformer's decisions using SHAP analysis under the selected-feature setting. The SHAP summary (Figure 3) indicated that the model's outputs are primarily driven by profitability-related measures most notably Operating Profit to Sales Ratio (Var16) and Net Income to Shareholders' Equity (Return on Equity; Var17) which show the largest dispersion of SHAP values and therefore the greatest potential to shift predictions away from the baseline across observations. Financial health/distress, proxied by Altman's Z-score (Var24), exhibited a comparatively clearer directional pattern, with higher values generally associated with positive SHAP contributions, whereas higher levels of the Current Ratio (Var1) more frequently align with negative contributions, suggesting that stronger short-term liquidity reduces the likelihood of the positive class in the learned decision function. The remaining indicators including Total Debt-to-Equity Ratio (Var10), Cost of Goods Sold to Sales Ratio (Var12), and Financial Expenses to Total Debt Ratio (Var18) displayed mixed contributions around zero, consistent with context-dependent and non-monotonic effects arising from non-linear interactions among profitability, leverage, liquidity, and distress signals captured by the attention-based architecture.

Figure 3.
The SHAP Beeswarm Summary for the Transformer



(Source: The Researcher's Findings)

Descriptive Statistics

Some descriptive statistics including mean, median, maximum, minimum, standard deviation, skewness, and kurtosis, are reported in Table 3. For example, the mean of the current ratio is 11.81. The standard deviation of this variable is 20.52, indicating that the average dispersion of the observations around the mean is of this magnitude. The median of the current ratio is 5.53, implying that 50% of the observations lie above and 50% lie below this value.

Table 3.
The Descriptive Statistics of the Study Dataset

Kurtosis	Skewness	Std.	Minimum	Maximum	Median	Mean	Feature
25.58	4.42	20.52	0.31	145.51	5.53	11.81	Current Ratio
36.12	5.44	5.35	0.05	41.43	1.07	2.45	Total Debt-to-Equity Ratio
58.59	7.33	2.38	0.07	21.06	0.67	1.00	Cost of Goods Sold to Sales Ratio
61.47	-7.55	1.43	-12.46	0.04	0	-0.23	Operating Profit to Sales Ratio
19.10	-0.76	1.29	-7.07	6.27	0.07	-0.03	Net Income to Shareholders' Equity
14.09	2.73	0.03	0	0.25	0.03	0.37	Financial Expenses to Total Debt Ratio
41.23	-5.83	10.58	-80.10	12.10	2.10	0.32	Financial Distress (Altman Z-score, 1983)

(Source: The Researcher's Findings)

Discussion

The dataset was first partitioned into an 80% training set and a 20% held-out test set using a firm-level group split to prevent the information leakage (i.e., all observations from a given firm were assigned exclusively to either training or test). The model development and hyperparameter tuning were conducted exclusively on the training set

via five-fold stratified, firm-level group cross-validation. In each cross-validation iteration, the model was trained on four folds and validated on the remaining fold. After selecting the final configuration, the performance was assessed on the held-out test set. The results reported in Tables 4–7 summarize the test-set performance, aggregated over the five cross-validation–trained models.

According to the results reported in Tables 4–7, deep learning–based approaches clearly outperformed classical models. Among them, the Transformer neural network consistently achieved the best performance across all four datasets and provided the best balance between detecting fraud cases and controlling the false positive rate. After the Transformer, ensemble learning and the efficient capsule network ranked second and third, respectively (Tables 4–7). Among the convolutional models, the ResNet-50 showed the strongest performance within this family.

Finally, comparing the four datasets based on the best-performing model indicated that the highest F1-score was obtained with the 7 selected features (Table 7), followed by the combined 28+14 dataset (Table 6), the 28 raw variables (Table 4), and lastly the 14 financial ratios (Table 5). This pattern underscored that using raw data either alone or combined with financial ratios provides richer and complementary information for extracting the fraud-related patterns, leading to greater improvements in F1 compared with using financial ratios alone.

Practical Implications and Implementation in Auditing

From a practical standpoint, the proposed model should be viewed as a risk-screening tool that can be integrated into audit planning and continuous monitoring rather than as a stand-alone decision system. In an audit workflow, practitioners can compute the seven input features from routinely available financial statement data, generate a fraud-risk score using the trained model, and then use this score to (1) prioritize the firm-year engagements for enhanced procedures, (2) tailor the nature, timing, and extent of substantive testing in high-risk areas, and (3) support triage in regulatory surveillance. Implementation, however, requires explicit attention to operational constraints, including data quality and mapping consistency across reporting periods, threshold calibration to manage the cost of false positives, and periodic re-validation to mitigate the performance drift over time.

A further practical challenge is label and outcome latency in financial reporting fraud, ground-truth confirmation often arrives months or years after the reporting period (e.g., via enforcement actions), which complicates the timely model recalibration and performance monitoring.

Table 4.
The Performance of Twelve Methods on 28 Raw Input Variables Used in Bao et al. (2020)

Row	Method	Sensitivity	Precision	Specificity	Accuracy	F1-score
1	Decision Tree	0.28	0.06	0.96	0.96	0.09
2	Random Forest	0.34	0.08	0.97	0.96	0.13
3	Support Vector Machine	0.35	0.09	0.97	0.97	0.15
4	Spiking Neural Network	0.39	0.12	0.97	0.97	0.18
5	One-Dimensional Capsule Network	0.49	0.17	0.98	0.97	0.25
6	One-Dimensional Convolutional Neural Network	0.54	0.19	0.98	0.98	0.27
7	ResNet-18	0.52	0.18	0.98	0.97	0.26
8	ResNet-50	0.60	0.24	0.98	0.98	0.34
9	Efficient Capsule Network	0.67	0.29	0.98	0.98	0.40
10	Ensemble Learning	0.70	0.30	0.98	0.98	0.42
11	Transformer Neural Network	0.74	0.34	0.98	0.98	0.46
12	Logistic Regression	0.23	0.05	0.96	0.96	0.08

(Source: The Researcher's Findings)

Table 5.
The Performance of 12 Methods Used in This Study on 14 Financial Ratios from Bao et al. (2020)

Row	Method	Sensitivity	Precision	Specificity	Accuracy	F1-score
1	Decision Tree	0.25	0.05	0.96	0.95	0.08
2	Random Forest	0.32	0.07	0.97	0.96	0.12
3	Support Vector Machine	0.34	0.08	0.97	0.96	0.13
4	Spiking Neural Network	0.38	0.11	0.97	0.97	0.17
5	One-Dimensional Capsule Network	0.48	0.16	0.98	0.97	0.24
6	One-Dimensional Convolutional Neural Network	0.52	0.18	0.98	0.97	0.26
7	ResNet-18	0.49	0.17	0.98	0.97	0.25
8	ResNet-50	0.58	0.22	0.98	0.98	0.31
9	Efficient Capsule Network	0.65	0.27	0.98	0.98	0.38
10	Ensemble Learning	0.67	0.29	0.98	0.98	0.40
11	Transformer Neural Network	0.70	0.30	0.98	0.98	0.42
12	Logistic Regression	0.22	0.04	0.96	0.95	0.07

(Source: The Researcher's Findings)

Table 6.
The Performance of 12 Methods Used in This Study on Combined Set of 28 Raw Variables and 14 Financial Ratios from Bao et al. (2020)

Row	Method	Sensitivity	Precision	Specificity	Accuracy	F1-score
1	Decision Tree	0.29	0.06	0.96	0.96	0.10
2	Random Forest	0.35	0.09	0.97	0.96	0.15
3	Support Vector Machine	0.38	0.10	0.97	0.97	0.16
4	Spiking Neural Network	0.41	0.13	0.97	0.97	0.20
5	One-Dimensional Capsule Network	0.52	0.19	0.98	0.98	0.27
6	One-Dimensional Convolutional Neural Network	0.55	0.21	0.98	0.98	0.30
7	ResNet-18	0.54	0.2	0.98	0.98	0.29
8	ResNet-50	0.61	0.25	0.98	0.98	0.35
9	Efficient Capsule Network	0.70	0.30	0.98	0.98	0.42
10	Ensemble Learning	0.71	0.32	0.98	0.98	0.45
11	Transformer Neural Network	0.76	0.36	0.98	0.98	0.48
12	Logistic Regression	0.24862	0.05	0.96	0.96	0.09016

(Source: The Researcher's Findings)

Table 7.
The Performance of 12 Methods Used in This Study on Seven Features

Row	Method	Sensitivity	Precision	Specificity	Accuracy	F1-score
1	Decision Tree	0.32	0.06	0.96	0.96	0.11
2	Random Forest	0.38	0.1	0.97	0.97	0.15
3	Support Vector Machine	0.39	0.10	0.97	0.97	0.17
4	Spiking Neural Network	0.44	0.13	0.97	0.97	0.21
5	One-Dimensional Capsule Network	0.55	0.2	0.98	0.98	0.29
6	One-Dimensional Convolutional Neural Network	0.60	0.22	0.98	0.98	0.32
7	ResNet-18	0.58	0.21	0.98	0.98	0.31
8	ResNet-50	0.65	0.27	0.98	0.98	0.38
9	Efficient Capsule Network	0.71	0.32	0.98	0.98	0.44
10	Ensemble Learning	0.75	0.34	0.98	0.98	0.46
11	Transformer Neural Network	0.77	0.38	0.99	0.98	0.51
12	Logistic Regression	0.28	0.06	0.96	0.96	0.09

(Source: The Researcher's Findings)

These findings were derived from U.S. listed firms with AAER-identified misstatements; therefore, the implications should be interpreted within the U.S. regulatory and reporting environment.

Table 8.
The McNemar Paired Comparisons versus the Transformer

Model	Difference (%)	95% CI for Difference (%)	p_value
Decision Tree	-2.54	[-2.634, -2.456]	0
Random Forest	-1.88	[-1.965, -1.811]	0
Support Vector Machine	-1.72	[-1.794, -1.647]	0
Spiking Neural Network	-1.33	[-1.403, -1.271]	0
One-Dimensional Capsule Network	-0.87	[-0.927, -0.818]	4.3757e-263
One-Dimensional Convolutional Neural Network	-0.69	[-0.742, -0.643]	1.452e-197
ResNet-18	-0.76	[-0.819, -0.716]	9.5197e-225
ResNet-50	-0.45	[-0.497, -0.414]	2.9818e-116
Efficient Capsule Network	-0.23	[-0.268, -0.195]	4.2567e-37
Ensemble Learning	-0.15	[-0.197, -0.120]	3.2608e-16
Logistic Regression	-2.68	[-2.779, -2.597]	0

(Source: The Researcher's Findings)

Table 8 reports paired, two-sided exact McNemar tests comparing each model with the Transformer under the seven-feature setting. Difference (%) represents the net paired difference in instance-level correctness derived from the discordant pairs. Negative values indicate that the Transformer correctly classifies more cases than the comparator model. Across all baselines, the estimated differences were negative and the 95% confidence intervals lay entirely below zero, and all exact tests were significant ($p<0.001$), indicating a consistent instance-level advantage for the Transformer. The magnitude of the advantage was largest for classical tabular baselines and narrowed for stronger deep/hybrid models, yet remained statistically significant throughout.

Practical Integration into Audit Planning and Regulatory Surveillance

The proposed Transformer model can be integrated as a risk-scoring layer in audit planning to rank firms by predicted fraud likelihood, thereby improving the resource allocation and

the design of targeted substantive procedures (e.g., focusing on revenue-related accounts and accrual-intensive items). The parsimonious feature subset enables fast, low-cost screening and allows the score to be mapped into the operational risk tiers (low/medium/high) that trigger predefined audit responses and documentation requirements. For regulators, the same score supports continuous surveillance, early-warning alerts, and prioritization of issuers for review under constrained supervisory capacity. Coupling the risk score with explainability outputs (e.g., feature-attribution summaries) further enhances the interpretability and practitioner trust in the resulting flags.

Conclusion

This study developed and evaluated a data-driven framework for detecting financial statement fraud using a unified, leakage-resistant evaluation protocol and a broad benchmark set of classical, deep, and hybrid classifiers. Using AAER-identified fraud cases matched to Compustat accounting data for U.S. listed firms (1991–2014), we compared four input scenarios including raw accounting items, financial ratios, their combination, and a compact seven-feature subset (six ratios plus Altman's Z-score). Across all scenarios, deep and hybrid models consistently outperformed the classical baselines, underscoring the non-linear and interaction-driven structure of fraud signals in financial reporting data. The Transformer model delivered the strongest and most stable performance, achieving the best overall balance between fraud detection capability and false-positive control, with its highest F1-score obtained under the parsimonious seven-feature configuration. Importantly, the combined raw-item and ratio representation outperformed the ratios-only configuration, indicating that raw financial statement items provide incremental discriminatory information beyond the engineered ratios.

Beyond the predictive accuracy, the study emphasized deployability in audit and regulatory settings. The selected seven-feature subset enabled low-cost screening based on routinely available financial statement information, while the SHAP-based interpretability analysis provided transparency on which accounting signals most strongly influence the Transformer's decisions. Together, these results supported the use of the proposed model as a risk-screening layer to prioritize firm-year engagements for further investigations and to inform audit planning and supervisory surveillance.

Limitations and Future Research Directions

The limitations of this paper include missing or inconsistent data as well as class imbalance between the fraudulent and non-fraudulent classes. In addition, a systematic comparison between the imbalance-handling strategy adopted in this study, cost-sensitive learning and alternative data-imbalance techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN), as well as potential hybrid schemes combining SMOTE/ADASYN with cost-sensitive learning, would be a valuable extension for future research in financial statement fraud detection.

REFERENCES

Achakzai, M. A. K., & Juan, P. (2022). Using machine learning Meta-Classifiers to detect financial frauds. *Finance Research Letters*, *48*, 102915. https://doi.org/10.1016/J.FRL.2022.102915.

Ahmed, K., & Courtis, J. K. (2015). The determinants of financial ratio disclosures and quality: Evidence from an emerging market. *British Accounting Review*, *31*(1), 35–61. https://doi.org/10.1006/BARE.1998.0082.

Ahmed, K. H., Axelsson, S., Li, Y., & Sagheer, A. M. (2025). A credit card fraud detection approach based on ensemble machine learning classifier with hybrid data sampling. *Machine Learning with Applications*, *20*, 100675. https://doi.org/10.1016/J.MLWA.2025.100675.

Altman, E.I. (1983) Corporate Financial Distress: A Complete Guide to Predicting, Avoiding, and Dealing with Bankruptcy. Wiley, New York (1983), 368.

Arboleda, F. J. M., Guzman-Luna, J. A., & Torres, I. D. (2018). Fraud detection-oriented operators in a data warehouse based on forensic accounting techniques. *Computer Fraud & Security*, *2018*(10), 13–19. https://doi.org/10.1016/S1361-3723(18)30098-8.

Azim Mim, M., Majadi, N., & Mazumder, P. (2024). A soft voting ensemble learning approach for credit card fraud detection. *Heliyon*, *10*(3), e25466. https://doi.org/10.1016/j.heliyon.2024.e25466.

Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. *Journal of Accounting Research*, *58*(1), 199–235. https://doi.org/10.1111/1475-679X.12292.

Bhattacharya, I., & Mickovic, A. (2024). Accounting fraud detection using contextual language learning. *International Journal of Accounting Information Systems*, *53*, 100682. https://doi.org/10.1016/J.ACCINF.2024.100682.

Cai, S., & Xie, Z. (2024). Explainable fraud detection of financial statement data driven by two-layer knowledge graph. *Expert Systems with Applications*, *246*, 123126. https://doi.org/10.1016/J.ESWA.2023.123126.

Cao, R., Wang, J., Mao, M., Liu, G., & Jiang, C. (2023). Feature-wise attention based boosting ensemble method for fraud detection. *Engineering Applications of Artificial Intelligence*, *126*, 106975. https://doi.org/10.1016/J.ENGAPPAI.2023.106975.

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Detecting management fraud in public companies. *Management Science*, *56*(7), 1146–1160. https://doi.org/10.1287/MNSC.1100.1174.

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting material accounting misstatements. *Contemporary Accounting Research*, *28*(1), 17–82. https://doi.org/10.1111/J.1911-3846.2010.01041.X.

Etemadi, H., & Zolghi, H. (2013). Application of logistic regression in detecting fraudulent financial reporting. *Danesh-e Hesabresi (Auditing Knowledge)*, *13*(51) 145-163.

Islam, M. R., Qaraqe, M., Qaraqe, K., & Serpedin, E. (2024). CAT-Net: Convolution, attention, and transformer based network for single-lead ECG arrhythmia classification. *Biomedical Signal Processing and Control*, *93*, 106211. https://doi.org/10.1016/J.BSPC.2024.106211.

Jeyasothy, A., Suresh, S., Ramasamy, S., & Sundararajan, N. (2024). Development of a novel transformation of spiking neural classifier to an interpretable classifier. *IEEE Transactions on Cybernetics*, *54*(1), 3–12. https://doi.org/10.1109/TCYB.2022.3181181.

Kanapickienė, R., & Grundienė, Ž. (2015). The model of fraud detection in financial statements by means of financial ratios. *Procedia - Social and Behavioral Sciences*, *213*, 321–327.

https://doi.org/10.1016/J.SBSPRO.2015.11.545.

Karnavou, E., Cascavilla, G., Marcelino, G., & Geradts, Z. (2025). I know you're a fraud: Uncovering illicit activity in a Greek bank transactions with unsupervised learning. *Expert Systems with Applications*, *288*, 128148. https://doi.org/10.1016/J.ESWA.2025.128148.

Kim, Y. J., Baik, B., & Cho, S. (2016). Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert Systems with Applications*, *62*, 32–43. https://doi.org/10.1016/J.ESWA.2016.06.016.

Lei, Y. T., Ma, C. Q., Ren, Y. S., Chen, X. Q., Narayan, S., & Huynh, A. N. Q. (2023). A distributed deep neural network model for credit card fraud detection. *Finance Research Letters*, *58*, 104547. https://doi.org/10.1016/J.FRL.2023.104547.

Lu, J., Xu, Q., & Hu, J. (2026). A novel graph learning framework for interpretable and imbalance financial fraud detection. *Engineering Applications of Artificial Intelligence*, *167*, 113709. https://doi.org/10.1016/J.ENGAPPAI.2025.113709.

Mazzia, V., Salvetti, F., & Chiaberge, M. (2021). Efficient-CapsNet: capsule network with self-attention routing. *Scientific Reports, 11*(1), 14634. https://doi.org/10.1038/s41598-021-93977-0.

Mehrabi Hashjin, N., Amiri, M. H., Mohammadzadeh, A., Mirjalili, S., & Khodadadi, N. (2024). Novel hybrid classifier based on fuzzy type-III decision maker and ensemble deep learning model and improved chaos game optimization. *Cluster Computing*, *27*(7), 10197–10234. https://doi.org/10.1007/S10586-024-04475-7/METRICS.

Narayana Gorle, V. L., & Panigrahi, S. (2026). An efficient heuristic optimization-based fraudulent activities detection in the financial sector using adaptive machine learning and deep learning system. *Expert Systems with Applications*, *302*, 130551. https://doi.org/10.1016/J.ESWA.2025.130551.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159.

Sai, C. V., Das, D., Elmitwally, N., Elezaj, O., & Islam, M. B. (2023). *Explainable ai-driven financial transaction fraud detection using machine learning and deep neural networks*. https://doi.org/10.2139/SSRN.4439980.

Shao, Z., Yu, H., Wen, J., Liu, Z., & Qi, P. (2026). A graph fraud detection model based on mutual information. *Neurocomputing*, *663*, 131972. https://doi.org/10.1016/J.NEUCOM.2025.131972.

Vakilifard, H. R., Jabarzadeh Kangarlouei, S., & Pourreza Sultan Ahmadi, A. (2009). An investigation of the characteristics of fraud in financial statements. *Monthly Magazine of the Iranian Association of Certified Public Accountants*, (210), 26–41.

Zhang, Z., Wang, Z., & Cai, L. (2025). Predicting financial fraud in Chinese listed companies: An enterprise portrait and machine learning approach. *Pacific-Basin Finance Journal*, *90*, 102665. https://doi.org/10.1016/J.PACFIN.2025.102665.