

EXPLAINABLE MACHINE LEARNING EARLY-WARNING SYSTEM FOR CORPORATE GOVERNANCE FAILURE AND FINANCIAL RESTATEMENT DETECTION

Assaf Filfilan¹

Received 29.04.2026.
Revised 05.08.2026.
Accepted 17.08.2026.

Keywords:

Corporate governance failure, Financial restatement, Accounting fraud detection, Early-warning system, Explainable artificial intelligence, SHAP, Gradient boosting; AAER

ABSTRACT

Corporate governance failures that end in financial restatement are costly, slow to surface and difficult for boards, auditors and regulators to anticipate. This study develops and evaluates an explainable machine learning early-warning system (EWS) that ranks listed firms by their risk of a material misstatement that is later the subject of a Securities and Exchange Commission (SEC) Accounting and Auditing Enforcement Release (AAER). Using a public benchmark panel of 108,399 US firm-years (15,320 firms, 1991-2008) with 837 misstatement firm-years, we train models only on information available two years before each test year and evaluate them out of sample over 2003-2008 (35,166 firm-years, 261 misstatements). A gradient-boosted tree model (XGBoost) built on 28 raw financial statement items achieves a mean out-of-sample AUC of 0.718 (pooled AUC 0.722, 95% CI 0.698-0.748), outperforming a logistic regression on the 14 financial ratios of Dechow et al. (2011) by 0.040 AUC points (95% CI 0.009-0.070; bootstrap $p = .004$). Translated into a four-tier EWS, the 20% of firms flagged Red, Amber or Watch contain 44.1% of all subsequently enforced misstatements, while the Green tier carries a misstatement rate 30% below the base rate. Shapley additive explanations (SHAP) show that the model relies on a small, stable set of signals (low property, plant and equipment intensity, high share price, recent equity issuance, large total and current assets), with year-to-year rank correlations of 0.84-0.98. The ratio-based logit remains stronger at the extreme top 1% of the ranking, which supports a layered design in which a flexible model screens broadly and a transparent model prioritises. The paper contributes a reproducible, explainable and temporally honest EWS template, and documents its limits: modest absolute precision, label incompleteness and the absence of board-level governance variables.



© 2026 Published by Faculty of Engineering

¹ Corresponding author: Assaf Filfilan
E-mail: afilfilan@uj.edu.sa

1. INTRODUCTION

Financial misreporting is one of the most visible symptoms of corporate governance failure. When management misstates earnings or the balance sheet, and the board, audit committee and external auditor fail to prevent or detect it, the consequences are severe. Firms caught misrepresenting their accounts lose a large share of their market value, much of it from reputational penalties rather than legal fines (Karpoff et al., 2008). Restatement announcements trigger sharp negative returns, especially where the restatement involves fraud rather than error (Palmrose et al., 2004; Hennes et al., 2008). Governance research has long linked these events to weak oversight, including boards with fewer outside directors (Beasley, 1996), audit committees lacking independence or financial expertise (Klein, 2002; Agrawal & Chadha, 2005), and firms whose governance only improves after detection (Farber, 2005).

Detection, however, is slow and uneven. Dyck et al. (2010) show that fraud is frequently exposed by employees, media and industry regulators rather than by the gatekeepers formally responsible for oversight. By the time an SEC enforcement release is issued, the misstatement has often persisted for several years. This lag creates a practical need for early-warning tools that can rank firms by risk using information that is already available, so that scarce audit, regulatory and board attention can be directed where it is most likely to matter. The accounting literature has produced a sequence of such tools. Beneish (1999) proposed the M-score, Dechow et al. (2011) developed the F-score from a set of accrual, performance and financing variables, and Bao et al. (2020) showed that an ensemble learning model applied to raw financial statement items outperforms ratio-based models out of sample. Machine learning has since become the leading approach to misstatement prediction (Perols, 2011; Brown et al., 2020; Bertomeu et al., 2021). Two problems remain. First, flexible models are often opaque. A board chair, audit partner or regulator cannot act on a score without understanding why a firm was flagged, and opacity is a serious concern in high-stakes settings (Rudin, 2019). Second, many studies report in-sample or randomly cross-validated performance, which overstates what an EWS could deliver in real time because misstatements are only revealed years later.

This paper addresses both problems. We build an EWS that (a) is trained only on data that would have been available, with a two-year gap to allow for detection lags; (b) is evaluated on a rolling out-of-sample basis across six test years; (c) converts continuous scores into an operational four-tier alert framework; and (d) explains both global model behaviour and individual alerts using Shapley additive explanations (SHAP; Lundberg & Lee, 2017; Lundberg et al., 2020).

The study is guided by four research questions:

- RQ1. How accurately can machine learning models applied to publicly reported financial statement data rank firms by their risk of a future enforced misstatement, relative to an established ratio-based benchmark?
- RQ2. How much of the misstatement population can a tiered early-warning framework capture at realistic screening budgets?
- RQ3. Which financial signals drive the model's alerts, and are those signals stable over time?
- RQ4. Can the model's individual alerts be explained in a form usable by boards, auditors and regulators?

The paper makes three contributions. Methodologically, it combines a strict temporal design with model-agnostic, game-theoretic explanations, and shows that the explanation layer is stable enough across years to support governance use. Empirically, it confirms on public benchmark data that a gradient-boosted model of raw financial items ranks enforced misstatements significantly better than the ratio benchmark, while showing that the ratio model retains an advantage at the very top of the ranking. Practically, it proposes a tiered alert design and an explanation format that can be embedded in audit planning, audit committee oversight and regulatory surveillance.

The remainder of the paper is organised as follows. Section 2 reviews the literature and theoretical framework. Section 3 describes the data and method. Section 4 reports the results. Section 5 discusses implications and limitations, and Section 6 concludes.

2. LITERATURE REVIEW AND THEORETICAL FRAMEWORK

2.1 Governance failure and financial misreporting

Agency theory provides the core rationale for treating misreporting as a governance failure. Where ownership and control are separated, managers have incentives and opportunities to act in their own interest, and governance mechanisms exist to bound those agency costs (Jensen & Meckling, 1976). Financial reporting is one of the main channels through which principals monitor agents, so a material misstatement signals that the monitoring system itself has broken down. Empirical work supports this view. Firms subject to SEC enforcement for earnings manipulation have weaker boards and are more likely to have a chief executive who also founded the firm or chairs the board (Dechow et al., 1996). Fraud firms have a lower share of outside directors (Beasley, 1996), and independent directors with financial expertise reduce the probability of restatement (Agrawal & Chadha, 2005). After fraud is revealed, firms tend to strengthen their

boards, which implies that the earlier governance arrangement was deficient (Farber, 2005).

The fraud triangle (Cressey, 1953) complements agency theory by describing the conditions under which misreporting occurs: pressure, opportunity and rationalisation. Financial statement data can proxy for the first two. Pressure appears as deteriorating performance, a need to raise external finance or high market expectations embedded in the share price. Opportunity appears as a large share of assets whose values depend on judgement, such as receivables and inventory, rather than hard assets such as property, plant and equipment. These constructs underpin most ratio-based detection models and provide a theoretical lens for interpreting machine learning explanations.

2.2 Measuring the outcome: restatements and enforcement

Researchers measure misreporting using restatement databases, securities class actions or SEC AAERs. Each proxy has trade-offs (Karpoff et al., 2017). Restatements include many minor technical errors, while AAERs capture a narrower set of serious, intentional or highly material misstatements that the SEC chose to pursue (Hennes et al., 2008). Because this study is concerned with governance failure rather than routine accounting error, AAER-linked misstatements are the more appropriate outcome. Every AAER case involves a restatement or correction of previously issued financial statements and a finding that internal and external oversight did not prevent the misstatement. The trade-off is that AAERs are an incomplete record of all misconduct, a point we return to in Section 5.

2.3 From ratio models to machine learning

Early detection models used a small number of theoretically motivated ratios in a linear or logistic specification. The Beneish (1999) M-score uses eight indices, including receivables growth, asset quality and accruals. Dechow et al. (2011) assembled a large AAER sample and developed the F-score from accrual quality, performance, non-financial and market-based variables, establishing a widely used benchmark. Perols (2011) compared several statistical and machine learning classifiers and found that relative performance depends heavily on class imbalance and misclassification costs.

Bao et al. (2020) changed the benchmark by showing that an ensemble learner (RUSBoost) applied to 28 raw financial statement items outperformed both the Dechow et al. (2011) logit and a support vector machine on financial ratios, reporting an out-of-sample AUC of about 0.73. Their key insight is that predefined ratios discard information that a flexible learner can recover from raw items. Subsequent work has extended detection to text (Brown et al., 2020) and to a broader set of accounting and market variables (Bertomeu et al., 2021). Bao et al.

(2020) also highlighted a methodological trap: because fraud often spans several consecutive years, a random split can place years of the same fraud in both training and test sets, inflating performance. They addressed this with a time gap and by recoding serial frauds, a design we adopt.

2.4 Explainable artificial intelligence in assurance and oversight

Flexible models such as gradient-boosted trees (Chen & Guestrin, 2016) and random forests (Breiman, 2001) capture non-linearities and interactions but do not offer coefficients that users can read directly. In governance settings this is more than an inconvenience. An auditor needs to document why additional procedures were planned, an audit committee needs to know which accounts to question, and a regulator needs a defensible basis for selecting firms for review. Rudin (2019) argues that, in high-stakes decisions, inherently interpretable models should be preferred where they perform comparably. Where they do not, post hoc explanation methods offer a bridge. SHAP decomposes each prediction into additive contributions from each feature, grounded in Shapley values from cooperative game theory, and satisfies consistency and local accuracy properties (Lundberg & Lee, 2017). TreeSHAP computes these values exactly and efficiently for tree ensembles (Lundberg et al., 2020). Broader reviews of explainable AI stress that explanations should be faithful, stable and meaningful to the intended audience (Barredo Arrieta et al., 2020).

2.5 Research gap and framework

Prior work establishes that machine learning improves misstatement detection, but three gaps remain. Few studies (a) pair strict out-of-sample evaluation with explanation, (b) test whether explanations are stable enough over time to be trusted, or (c) translate scores into an alert structure that governance actors can operationalise. The framework for this study therefore links agency theory and the fraud triangle (why misreporting occurs) to observable financial signals (what can be measured), to a predictive model (how risk is estimated), to SHAP explanations (why a firm is flagged), and finally to tiered alerts that inform audit, board and regulatory action.

3. DATA AND METHODOLOGY

3.1 Data source and sample

We use the publicly released replication dataset of Bao et al. (2020), which links Compustat annual financial data for US listed firms to misstatement labels derived from SEC AAERs. Using a public benchmark has two advantages: results are fully reproducible, and they can be compared directly with an established study. The sample covers fiscal years 1991-2008 and contains

108,399 firm-year observations for 15,320 distinct firms, of which 837 firm-years (0.77%) are identified as misstated. The outcome variable equals 1 if the firm-year's financial statements were materially misstated and the misstatement was later the subject of an AAER, and 0 otherwise. Consistent with Section 2.2, we treat this as an observable marker of governance failure: a material

misstatement that internal controls, the board, the audit committee and the external auditor did not prevent, and that required regulatory enforcement to correct.

The out-of-sample test period is 2003-2008, the same window used by Bao et al. (2020). Table 1 reports the test sample.

Table 1. Out-of-sample test years

Test year	Training window	Firm-years	Misstated firm-years	Misstatement rate (%)
2003	1991-2001	5,981	69	1.15
2004	1991-2002	5,934	58	0.98
2005	1991-2003	5,863	45	0.77
2006	1991-2004	5,908	33	0.56
2007	1991-2005	5,868	30	0.51
2008	1991-2006	5,612	26	0.46
Total		35,166	261	0.74

The declining rate in later years reflects the detection lag: misstatements from recent years had less time to be uncovered and enforced before the dataset was compiled. This is a known feature of AAER data and one reason why ranking metrics are more informative than accuracy at a fixed threshold.

3.2 Predictors

Two feature sets are used. The raw feature set comprises 28 financial statement items reported in Compustat, following Bao et al. (2020): current assets, accounts payable, total assets, common equity, cash and short-term investments, cost of goods sold, common shares outstanding, debt in current liabilities, long-term debt issued, long-term debt, depreciation and amortisation, income before extraordinary items, inventories, investments and advances (other), short-term investments, current liabilities, total liabilities, net income, gross property, plant and equipment (PP&E), preferred stock, retained earnings, receivables, sales, sale of common and preferred stock, income taxes payable, income taxes, interest expense and fiscal year-end share price.

The ratio feature set comprises the 14 financial ratios used by Dechow et al. (2011) and Bao et al. (2020): working capital accruals, RSST accruals, change in receivables, change in inventory, percentage of soft assets, change in cash sales, change in cash margin, change in return on assets, actual issuance of securities, book-to-market, depreciation index, retained earnings over assets, EBIT over assets and change in free cash flow. These ratios operationalise the pressure and opportunity dimensions of the fraud triangle. Missing values are set to zero, matching the source data's convention. For the logistic model only, ratios are winsorised at the 1st and 99th percentiles of each training window and standardised.

3.3 Temporal validation design

An EWS must work in real time, so the validation design mirrors the information available to a decision-maker. For each test year t , models are trained on all firm-years from 1991 to $t - 2$. The two-year gap allows for the time it takes misstatements to be detected; training on year $t - 1$ would use labels that a user in year t could not yet know. Following Bao et al. (2020), where a fraud episode spans both the training and test periods, its training-period observations are recoded to 0. This prevents the model from recognising a specific firm's ongoing fraud rather than learning generalisable risk patterns. Models are retrained for each of the six test years.

3.4 Models

Four model specifications are compared:

- 1) Logit (financial ratios). A class-weighted logistic regression on the 14 Dechow et al. (2011) ratios. This is the transparent benchmark.
- 2) Random forest (raw + ratios). 500 trees, minimum leaf size of five, balanced subsample class weights (Breiman, 2001).
- 3) XGBoost (raw + ratios). Gradient-boosted trees (Chen & Guestrin, 2016) on all 42 features.
- 4) XGBoost (raw only). The same learner on the 28 raw items, directly testing the Bao et al. (2020) proposition that raw data outperform predefined ratios.

Hyperparameters for the boosted models were fixed in advance and not tuned on test data: 400 trees, maximum depth of four, learning rate of 0.05, row and column subsampling of 0.8, and a positive-class weight equal to the ratio of negatives to positives in each training window. To reduce dependence on a single random seed, boosted-model scores are averaged across five seeds. A rank-average hybrid of the logit and XGBoost (raw + ratios) scores was also evaluated as a pre-specified sensitivity check.

3.5 Evaluation metrics

Because misstatements occur in fewer than 1% of firm-years and an EWS is used to rank rather than to classify, evaluation focuses on ranking quality. We report the area under the ROC curve (AUC); average precision (AP), which is more informative under severe class imbalance (Saito & Rehmsmeier, 2015); sensitivity and precision in the top 1% and top 5% of the ranking; and normalised discounted cumulative gain in the top 1% (NDCG@1%; Järvelin & Kekäläinen, 2002), as used by Bao et al. (2020):

$$NDCG@k = \frac{1}{IDCG@k} \sum_{i=1}^k \frac{y_{(i)}}{\log_2(i+1)} \quad (1)$$

where $y_{(i)}$ is the outcome of the firm ranked i -th and $IDCG@k$ is the maximum achievable value. Metrics are computed for each test year and averaged. To compare models statistically, scores are converted to within-year percentile ranks and pooled across 2003-2008. We then compute a pooled AUC with a 95% bootstrap confidence interval (300 resamples) and a paired bootstrap test of the AUC difference between XGBoost (raw only) and the logit benchmark (1,000 resamples).

3.6 Explainability

Model explanations are produced with TreeSHAP (Lundberg et al., 2020). For a model f and firm x , the prediction is decomposed as:

$$f(x) = \varphi_0 + \sum_{j=1}^M \varphi_j(x) \quad (2)$$

where φ_0 is the expected model output (in log-odds) and φ_j is the Shapley contribution of feature j . Global importance is measured as the mean absolute SHAP value of each feature in each test year, averaged across the five seeds. The direction of each effect is summarised by the correlation between a feature's rank and its SHAP value. Temporal stability of explanations is assessed using Spearman rank correlations between importance vectors in adjacent test years. Local explanations are illustrated for the highest-scoring true misstatement in the 2008 test year.

3.7 Early-warning tiers

Scores are converted to four alert tiers within each test year, reflecting typical resource constraints of audit firms and regulators: Red (top 1%), Amber (next 4%, to the top 5%), Watch (next 15%, to the top 20%) and Green (bottom 80%). For each tier we report the misstatement rate, its share of all misstatements and its lift relative to the base rate.

4. RESULTS

4.1 Predictive performance (RQ1)

XGBoost on raw financial items is the strongest overall ranker. Table 2 reports performance averaged across the six test years.

Table 2. Out-of-sample performance, mean of test years 2003-2008

Model	AUC	Pooled AUC (95% CI)	AP	NDCG @1%	Sensitivity top 1% (%)	Precision top 1% (%)	Sensitivity top 5% (%)	Precision top 5% (%)
Logit (financial ratios)	0.688	0.682 (0.654-0.709)	0.017	0.035	4.6	2.5	11.6	1.5
Random forest (raw + ratios)	0.690	-	0.016	0.016	2.2	2.0	10.1	1.7
XGBoost (raw + ratios)	0.682	0.690 (0.656-0.717)	0.017	0.024	2.6	2.2	10.9	1.8
XGBoost (raw only)	0.718	0.722 (0.698-0.748)	0.017	0.015	1.9	1.4	12.4	1.9
Hybrid (rank average)	0.701	0.703 (0.671-0.730)	0.016	0.011	1.3	0.9	11.1	1.5

Notes. Bold marks the best value in each column. Base rate is 0.74%, so a random ranking yields AUC 0.5, AP of about 0.007 and precision of 0.74% at any cut-off. XGBoost scores are averaged over five seeds; the standard deviation of AUC across seeds was about 0.007. Pooled AUC uses within-year percentile ranks; a dash indicates that a pooled AUC was not computed. AP = average precision; NDCG = normalised discounted cumulative gain.

XGBoost (raw only) achieves a mean AUC of 0.718, close to the AUC of about 0.73 reported by Bao et al. (2020) for their RUSBoost model on the same data. Its pooled AUC exceeds that of the ratio-based logit by 0.040 (95% bootstrap CI 0.009-0.070; $p = .004$), so the improvement is statistically reliable. Adding the 14 ratios to the boosted model did not help and slightly reduced AUC (0.682), consistent with the view that predefined ratios add noise once a flexible learner has access to the underlying items. The random forest performed similarly to the logit.

Table 3 shows that the advantage of XGBoost (raw only) holds in five of the six test years, although margins narrow in 2007-2008, when very few misstatements remain in the test data.

Table 3. AUC by test year

Test year	Logit (ratios)	XGBoost (raw + ratios)	XGBoost (raw only)	Hybrid
2003	0.681	0.707	0.744	0.713
2004	0.675	0.724	0.723	0.720
2005	0.645	0.661	0.699	0.667
2006	0.698	0.707	0.756	0.720
2007	0.713	0.656	0.715	0.698
2008	0.713	0.638	0.670	0.689

Note. Bold marks the highest AUC in each test year

The picture is different at the extreme top of the ranking. The logit benchmark places 4.6% of misstatements in its top 1%, compared with 1.9% for XGBoost (raw only), and has the highest NDCG@1% (0.035). The boosted model, in contrast, spreads its gains across the upper 5-20% of the distribution. This pattern matters for the design of the EWS and is discussed in Section 5.

Table 4. Early-warning tier performance, XGBoost (raw only), 2003-2008

Tier	Ranking band	Firm-years	Misstatements	Misstatement rate (%)	Share of all misstatements (%)	Lift vs base rate
Red	Top 1%	349	4	1.15	1.5	1.54
Amber	1-5%	1,407	29	2.06	11.1	2.78
Watch	5-20%	5,274	82	1.55	31.4	2.09
Green	Bottom 80%	28,136	146	0.52	55.9	0.70
Total		35,166	261	0.74	100.0	1.00

Note. Lift is the tier misstatement rate divided by the overall base rate of 0.74%. Shares may not sum exactly to 100 because of rounding

The top 5% (Red and Amber) captures 12.6% of misstatements, and the top 20% captures 44.1%, more than twice what a random selection of the same size would find. Firms in the Green tier are 30% less likely than average to have a subsequently enforced misstatement. Under the logit benchmark the Red tier is more productive (9 misstatements in 349 firm-years, a rate of 2.6%), but its top 20% captures 37.9% of cases, against 44.1% for XGBoost.

4.3 Global explanations and stability (RQ3)

SHAP analysis shows that the boosted model relies on a compact set of economically interpretable signals. Figure 1 shows the ten most important raw features, averaged across test years and seeds.

4.2 Early-warning tiers (RQ2)

Table 4 applies the four-tier framework to the XGBoost (raw only) scores, pooling the six test years.

The most important feature is gross PP&E, and its effect is strongly negative (rank-SHAP correlation of -0.91 in the 2003 test year; directions below are from the same year): firms with more hard, verifiable assets are less likely to misstate. This mirrors the soft-assets argument of Dechow et al. (2011) and the opportunity element of the fraud triangle. When the 14 ratios are added to the model, the percentage of soft assets becomes the single most important feature, which confirms that the raw-item model is learning the same construct. The next most important signals are a higher year-end share price (+0.92) and recent sale of common and preferred stock (+0.87). Both capture pressure: high market expectations and active capital raising create incentives to present favourable results. Larger total assets (+0.79) and receivables (+0.72) also raise risk, while larger short-term investment holdings (-0.81) and depreciation (-0.59) lower it.

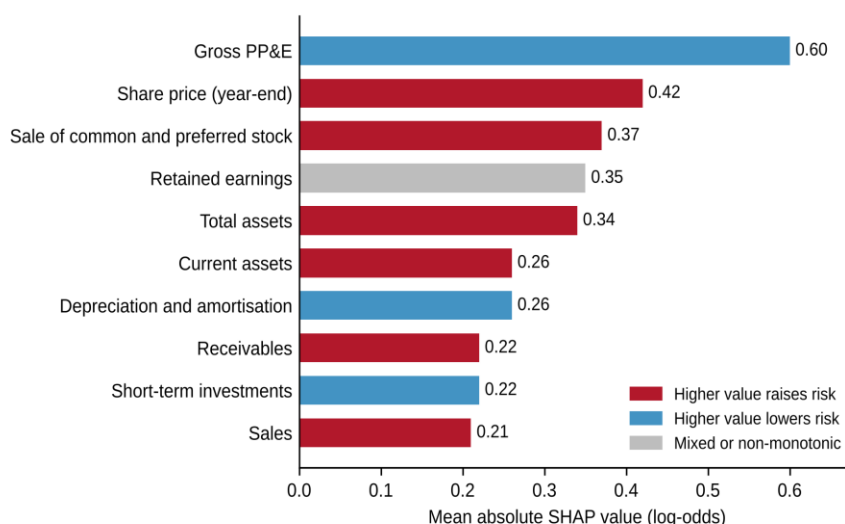


Figure 1. Global SHAP importance, top 10 of 28 raw features. Mean absolute SHAP value (log-odds), XGBoost on raw items, averaged over test years 2003-2008 and five seeds. Bar shading shows the direction of each effect, taken from the rank-SHAP correlation in the 2003 test year

Source: author's analysis of Bao et al. (2020) data

These explanations are stable over time. Spearman rank correlations between adjacent years' importance vectors range from 0.84 (2003 vs 2004) to 0.98 (2007 vs 2008), and the same five features (PP&E, share price, stock issuance, retained earnings and total assets) occupy the top five positions in every test year, with PP&E ranked first in all six. Stability is a precondition for governance use, because users need to trust that an alert reflects persistent risk patterns rather than idiosyncrasies of a single training window.

4.4 Local explanation of an individual alert (RQ4)

To illustrate how an alert would be presented to an audit committee or engagement partner, Table 5 decomposes the score of the highest-ranked true misstatement in the 2008 test year. The firm ranked in the top 0.2% of 5,612 firms, with a score of 0.87; its SHAP contributions sum to +1.67 log-odds above the model baseline. The firm is not identified.

Table 5. SHAP decomposition of a Red-tier alert, 2008

Feature	Firm value (\$m)	Sample median (\$m)	SHAP contribution (log-odds)	Effect
Gross PP&E	27,130	58	-1.38	Lowers risk
Current assets	26,563	74	+0.67	Raises risk
Shares outstanding (m)	458	41	+0.65	Raises risk
Current liabilities	25,436	35	+0.60	Raises risk
Accounts payable	18,919	9	+0.55	Raises risk
Receivables	18,580	16	+0.50	Raises risk
Depreciation and amortisation	2,351	6	-0.48	Lowers risk
Interest expense	1,570	2	-0.34	Lowers risk

Note. The eight features with the largest absolute contributions are shown; the remaining features account for the balance of the +1.67 log-odds total. Shares outstanding are reported in millions of shares rather than US\$ millions

The explanation shows a very large firm whose balance sheet is dominated by working capital. Receivables, payables and current assets each exceed \$18 billion, and together they push the score up, partly offset by a large fixed asset base. For a governance user, the actionable message is specific: review revenue recognition, receivables valuation and the working capital cycle, which are the account areas most often implicated in AAER cases (Dechow et al., 2011). The explanation does not prove misconduct; it directs attention.

5. DISCUSSION

5.1 Interpretation of findings

The results support three conclusions. First, flexible learning on raw financial statement data improves the ranking of firms by governance failure risk. The 0.040 gain in AUC over the ratio benchmark is modest in absolute terms but statistically reliable and consistent with Bao et al. (2020). It implies that the information content of financial statements is not fully captured by the handful of ratios that have dominated the literature.

Second, the explanations are theoretically coherent. The model did not discover arbitrary correlations. Its dominant signals map onto the opportunity and pressure elements of the fraud triangle: hard assets reduce the scope for judgement-based manipulation, while high valuations and equity issuance increase the incentive to present favourable results (Cressey, 1953; Dechow et al., 2011). This alignment between data-driven explanations and prior theory is a form of external validation and supports the use of the EWS by governance professionals who reason in these terms.

Third, no single model is best at every point of the ranking. The transparent logit concentrates more misstatements in the top 1%, while the boosted model is better across the top 5-20%. A practical EWS should therefore be layered. The boosted model is suited to broad screening, for example to set risk ratings across an audit firm's client portfolio or a regulator's population of listed entities. The ratio model, whose coefficients are directly interpretable, can then prioritise within the highest-risk group. The simple rank-average hybrid did not achieve this balance, which suggests that layering should be sequential rather than blended.

5.2 Implications for governance actors

Boards and audit committees. Tier assignments and SHAP explanations can be included in audit committee packs as a structured prompt for questioning management. A Watch or Amber rating driven by receivables and working capital, as in Table 5, points directors toward revenue recognition and credit policy. The EWS supplements, rather than replaces, the committee's own judgement.

External and internal auditors. Risk-based audit planning requires documented reasons for assessing a client as higher risk. SHAP contributions provide that documentation at the level of account areas, which supports both the planning of procedures and later quality review.

Regulators. Securities and corporate regulators that select filings for review could use the top 20% of the ranking to more than double their hit rate relative to random selection, while the explanation layer offers a transparent basis for selection that can withstand scrutiny.

Investors and lenders. Because the inputs are public financial statement items, the approach can be replicated by investors and lenders as part of due diligence, subject to the caveats below.

5.3 Limitations

Several limitations should be noted:

- 1) Absolute precision is low. Even in the best tier, only about 2% of flagged firm-years were later enforced. Most alerts are false positives in the sense of not leading to an AAER. The EWS is a prioritisation tool, not an accusation, and should be presented as such.
- 2) Label incompleteness. AAERs capture only misstatements the SEC detected and chose to pursue (Karpoff et al., 2017). Some Green-tier firms may have undetected problems, and some false positives may be true but undetected misstatements. Declining misstatement rates in later test years illustrate this censoring.
- 3) No direct governance variables. The benchmark dataset does not include board independence, audit committee composition, CEO duality, auditor tenure or ownership structure. Governance failure is therefore measured by its outcome rather than modelled through its causes. Adding these variables is the most important extension.
- 4) Time period and jurisdiction. The data end in 2008 and cover US firms only. Post-Sarbanes-Oxley reforms, changes in accounting standards and differences in enforcement intensity across jurisdictions may alter both the base rate and the relevant signals.
- 5) Explanations are associational. SHAP values describe how the model uses features, not causal effects. A large receivables balance raises a firm's score, but this does not mean receivables cause misreporting.
- 6) Size effects. Several important features are unscaled raw items, and the local example is a very large firm. Part of the model's signal may reflect firm size and the SEC's greater attention to larger firms, which an operational EWS should monitor.

References:

- Agrawal, A., & Chadha, S. (2005). Corporate governance and accounting scandals. *Journal of Law and Economics*, 48(2), 371-406. <https://doi.org/10.1086/430808>
- 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>
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115. <https://doi.org/10.1016/j.inffus.2019.12.012>

6. CONCLUSION AND FUTURE RESEARCH

This study developed an explainable machine learning early-warning system for corporate governance failure, measured as material financial misstatements subsequently enforced by the SEC. Using a strict temporal design on 108,399 public firm-years, a gradient-boosted model of 28 raw financial statement items ranked future enforced misstatements significantly better than an established ratio-based benchmark (AUC 0.718 vs 0.688). A four-tier alert framework placed 44.1% of misstatements in the top 20% of firms, and SHAP explanations revealed a small, theoretically coherent and highly stable set of drivers. Individual alerts can be decomposed into account-level contributions that direct audit and oversight attention.

The broader lesson is that accuracy and explainability are complements in governance analytics, not substitutes. A model that cannot explain its alerts is unlikely to be acted on, and an explanation that changes from year to year is unlikely to be trusted. The design proposed here, with temporal validation, stability-tested explanations and tiered alerts, offers a template for EWS development in other governance domains.

Future research should pursue five directions:

1. Integrate board, audit committee, auditor and ownership variables to model governance failure through its causes as well as its outcome.
2. Extend the analysis to more recent data and to non-US markets, including Australia and the Gulf Cooperation Council, where enforcement regimes and disclosure practices differ.
3. Combine financial data with textual signals from annual reports and audit opinions (Brown et al., 2020).
4. Evaluate cost-sensitive tier thresholds that reflect the relative costs of missed misstatements and unnecessary reviews for each user group.
5. Conduct field studies with auditors and audit committee members to test whether SHAP-based explanations improve decision quality in practice.

- Beasley, M. S. (1996). An empirical analysis of the relation between the board of director composition and financial statement fraud. *The Accounting Review*, 71(4), 443-465.
- Beneish, M. D. (1999). The detection of earnings manipulation. *Financial Analysts Journal*, 55(5), 24-36. <https://doi.org/10.2469/faj.v55.n5.2296>
- Bertomeu, J., Cheynel, E., Floyd, E., & Pan, W. (2021). Using machine learning to detect misstatements. *Review of Accounting Studies*, 26(2), 468-519. <https://doi.org/10.1007/s11142-020-09563-8>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- Brown, N. C., Crowley, R. M., & Elliott, W. B. (2020). What are you saying? Using topic to detect financial misreporting. *Journal of Accounting Research*, 58(1), 237-291. <https://doi.org/10.1111/1475-679X.12294>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). <https://doi.org/10.1145/2939672.2939785>
- Cressey, D. R. (1953). *Other people's money: A study in the social psychology of embezzlement*. Free Press.
- 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>
- Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1996). Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the SEC. *Contemporary Accounting Research*, 13(1), 1-36. <https://doi.org/10.1111/j.1911-3846.1996.tb00489.x>
- Dyck, A., Morse, A., & Zingales, L. (2010). Who blows the whistle on corporate fraud? *The Journal of Finance*, 65(6), 2213-2253. <https://doi.org/10.1111/j.1540-6261.2010.01614.x>
- Farber, D. B. (2005). Restoring trust after fraud: Does corporate governance matter? *The Accounting Review*, 80(2), 539-561. <https://doi.org/10.2308/accr.2005.80.2.539>
- Hennes, K. M., Leone, A. J., & Miller, B. P. (2008). The importance of distinguishing errors from irregularities in restatement research: The case of restatements and CEO/CFO turnover. *The Accounting Review*, 83(6), 1487-1519. <https://doi.org/10.2308/accr.2008.83.6.1487>
- Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4), 422-446. <https://doi.org/10.1145/582415.582418>
- Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4), 305-360. [https://doi.org/10.1016/0304-405X\(76\)90026-X](https://doi.org/10.1016/0304-405X(76)90026-X)
- Karpoff, J. M., Koester, A., Lee, D. S., & Martin, G. S. (2017). Proxies and databases in financial misconduct research. *The Accounting Review*, 92(6), 129-163. <https://doi.org/10.2308/accr-51766>
- Karpoff, J. M., Lee, D. S., & Martin, G. S. (2008). The cost to firms of cooking the books. *Journal of Financial and Quantitative Analysis*, 43(3), 581-611. <https://doi.org/10.1017/S0022109000004221>
- Klein, A. (2002). Audit committee, board of director characteristics, and earnings management. *Journal of Accounting and Economics*, 33(3), 375-400. [https://doi.org/10.1016/S0165-4101\(02\)00059-9](https://doi.org/10.1016/S0165-4101(02)00059-9)
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56-67. <https://doi.org/10.1038/s42256-019-0138-9>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765-4774).
- Palmrose, Z.-V., Richardson, V. J., & Scholz, S. (2004). Determinants of market reactions to restatement announcements. *Journal of Accounting and Economics*, 37(1), 59-89. <https://doi.org/10.1016/j.jacceco.2003.06.003>
- Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory*, 30(2), 19-50. <https://doi.org/10.2308/ajpt-50009>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215. <https://doi.org/10.1038/s42256-019-0048-x>
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432. <https://doi.org/10.1371/journal.pone.0118432>

Assaf Filfilan

Accounting Department,
College of Business,
University of Jeddah,
Jeddah,
Kingdom of Saudi Arabia
afilfilan@uj.edu.sa
ORCID 0009-0002-5578-866X
