World Journal of Economics and Business Research
Print ISSN: 2960-0073
Online ISSN: 2960-0081
DOI: https://doi.org/10.61784/wjebr4126

BEYOND ISOLATED RATIOS: ACCRUAL–REVENUE–CASH FLOW INDICATORS AND INTERACTION-AWARE LEARNING FOR DETECTING FINANCIAL REPORTING IRREGULARITIES IN U.S. PUBLIC COMPANIES

WenJie Ping[1*], HuiJie Fan[2]
[1]*College of the Holy Cross, Worcester 01610, Massachusetts, USA.*
[2]*Hunan Agricultural University, Changsha 410128, Hunan, China.*
[*]*Corresponding Author: WenJie Ping*

Abstract: Financial reporting irregularities are rare, evolve over time, and surface only after costly enforcement, which makes their detection from public filings a persistent challenge for regulators, auditors, and investors. This paper examines whether accounting-based indicators—accruals, revenue growth, operating cash flow, financial ratios, and the relations among them—identify misreporting risk out of sample. Using 146,045 firm-years of 10-K fundamentals of U.S. public companies (fiscal years 1990–2014) labeled with SEC Accounting and Auditing Enforcement Releases (964 misstated firm-years from 412 enforcement cases), we construct an Accrual–Revenue–Cash (ARC) indicator system of 53 indicators in six families, including cross-statement divergence and multi-year accrual accumulation, complemented by eight anomaly-lens scores based on ECOD. Unsupervised detectors, the F-score logit, RUSBoost, gradient boosting, a multilayer perceptron with piecewise linear encoding (PLE-MLP), and their rank ensemble (ARC-Lens) are evaluated under a rolling protocol with a 24-month gap and serial-fraud correction. ARC indicators raise the AUC of a logit model from 0.666 to 0.702 and of gradient boosting from 0.697 to 0.714. ARC-Lens reaches a mean AUC of 0.717 in 2003–2008, significantly above the F-score (+0.051) and the 42-input boosting benchmark (+0.020), on par with RUSBoost, and captures 14.9% of misstated firm-years in the top 5% of scores (10.7% for RUSBoost). Unsupervised outlier detectors perform at chance level (AUC 0.48–0.50): misreporting firms are statistically typical rather than extreme. Ablation and TreeSHAP analyses identify accrual accumulation as the dominant signal and attribute 58.9% of model attribution to pairwise interactions, chiefly between accrual measures and firm scale and financing.
Keywords: Financial reporting irregularities; Accruals; SEC enforcement releases; Fraud detection; Gradient boosting; Anomaly detection; Explainable artificial intelligence

1 INTRODUCTION

Misstated financial statements impose large losses on investors and erode confidence in capital markets. In the United States, the Securities and Exchange Commission (SEC) documents the most serious cases in its Accounting and Auditing Enforcement Releases (AAERs), which constitute the most reliable public evidence that reported numbers were misstated [1-2]. Because enforcement typically arrives years after the misreporting has occurred, regulators, auditors, and investors depend on screening tools that flag elevated risk from information already contained in annual 10-K filings [3-4].

Accounting research has developed a rich set of such signals. Accruals—the non-cash component of earnings—are less persistent than cash flows and are the principal vehicle of earnings manipulation [5-6]; discretionary-accrual models isolate the part of accruals that is not explained by revenue changes and capital intensity [7-9]; and composite scores such as the M-score and the F-score combine accrual, growth, and performance ratios into a probability of manipulation [3-4]. Machine learning (ML) extracts additional signal from the same filings: ensemble learners applied to raw financial-statement items outperform ratio-based logit models [10-13].

Three questions nevertheless remain open. First, indicators are usually evaluated in isolation, whereas misreporting manifests through the relations among statements—revenue that is not converted into cash, receivables that grow faster than sales, or earnings that diverge from operating cash flow. Whether explicitly engineered relational and multi-year indicators add detection power beyond established inputs has not been settled under a leakage-free protocol. Second, irregularities are often framed as anomalies, which invites modern unsupervised outlier detectors [14-16]; whether enforcement-labeled misreporting is statistically anomalous is an empirical question. Third, recent progress in tabular learning—strong gradient-boosting baselines [17], numerical-feature embeddings for neural networks [18], and evidence that boosting and deep models are complementary [19]—has not been assessed against the best-documented benchmark of this domain [11].

This paper addresses these questions with the Accrual–Revenue–Cash (ARC) framework outlined in Figure 1, built on 146,045 firm-years of U.S. 10-K fundamentals and SEC enforcement labels. The contributions are as follows.

1) An indicator system of 53 accounting indicators in six economically motivated families—accruals, revenue growth,

© By the Author(s) 2023, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

operating cash flow, financial ratios, cross-statement divergence, and temporal accumulation—together with family-wise ECOD anomaly lenses and a digit-level Benford measure.
2) A rolling out-of-sample evaluation with a 24-month reporting gap and serial-fraud correction that compares unsupervised, parsimonious, and state-of-the-art supervised detectors, including ARC-Lens, a heterogeneous ensemble of gradient boosting and a piecewise-linear-encoded multilayer perceptron.
3) Evidence that indicator design drives detection: ARC indicators improve both linear and tree-based learners, ARC-Lens matches the strongest published benchmark with higher top-5% sensitivity, and unsupervised outlier detection performs at chance level.
4) An interaction analysis based on TreeSHAP values, conditional enforcement rates, and an interaction logit that characterizes how accruals, revenue growth, cash flow, and firm characteristics jointly shape misreporting risk.

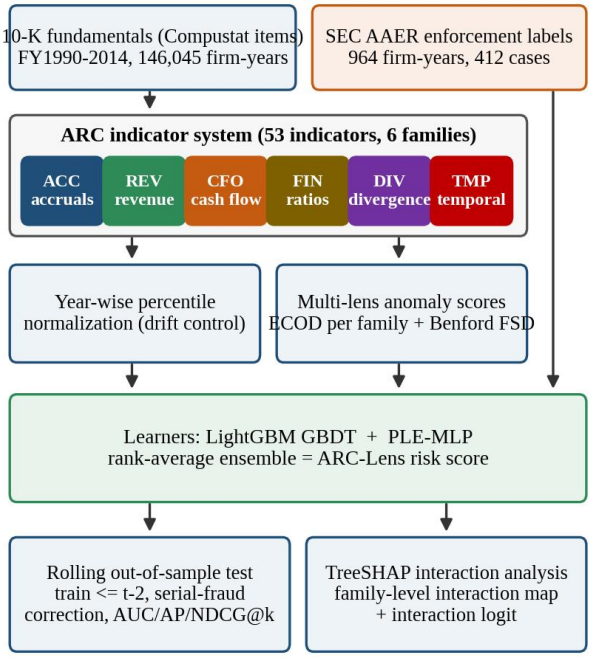


Figure 1 Overview of the ARC-Lens Framework: Indicator Construction from 10-K Fundamentals, Year-Wise Normalization, Multi-Lens Anomaly Scoring, Heterogeneous Learners, Leakage-Free Rolling Evaluation, and Interaction Analysis

Section 2 reviews related work, Section 3 describes the data and the indicator system, Section 4 presents the methodology, Section 5 reports the results, Section 6 discusses implications, and Section 7 concludes.

2 RELATED WORK

2.1 Accounting-Based Indicators of Misreporting

Sloan established that the accrual component of earnings is less persistent than the cash component [5], and Richardson et al. linked this property to the reliability of individual accrual categories [6]. Discretionary-accrual models estimate the abnormal part of accruals from revenue changes and capital intensity [7-8], with a performance control to account for profitability [9]; balance-sheet estimates of accruals contain measurement error that is relevant for their interpretation [20]. Firms subject to enforcement actions tend to raise external financing and exhibit weak governance [1]. Beneish and Dechow et al. condensed these insights into the M-score and the F-score [3-4], the latter estimated on AAER firms and widely used as a benchmark. Distributional measures based on Benford's law have been proposed as error indicators [21], while their incremental value over the F-score and simple growth screens has been contested [22]. Karpoff et al. show that conclusions depend on the misconduct database [2], which motivates the use of AAERs as the label source of this study.

2.2 Machine Learning for Misreporting Detection

Perols compared statistical and ML classifiers on fraud samples and found logistic regression and support vector machines competitive [10]. Bao et al. showed that RUSBoost applied to 28 raw financial-statement items outperforms ratio-based models [11, 23], and released the AAER-labeled dataset used here together with a protocol that accounts for detection lags and serial frauds. Bertomeu et al. applied gradient-boosted trees to accounting, market, governance, and audit variables and found that accounting variables become informative mainly through interactions [12]. Brown et al. demonstrated that the topic content of 10-K narratives adds incremental signal [24], and Achakzai and Peng introduced

dynamic ensemble selection to fraud detection [13]. These studies focus on predictive accuracy; the role of relational and temporal accounting indicators and of their interactions has received less systematic attention.

2.3 Tabular Learning and Anomaly Detection

Gradient-boosted decision trees such as XGBoost and LightGBM are reference models for tabular data [25-26], and a large benchmark confirmed that tree ensembles outperform deep networks on typical tabular tasks [17]. Shwartz-Ziv and Armon found that ensembles of boosting and deep models outperform either component [19], and Gorishniy et al. showed that piecewise linear encodings (PLE) of numerical features substantially improve multilayer perceptrons [18]. For unsupervised detection, isolation forests and the parameter-free ECOD detector are strong baselines [14-15], and ADBench documents the value of supervision when labels are available [16]. TreeSHAP interaction values provide exact attributions of pairwise effects in tree ensembles and underpin the interaction analysis of this paper [27].

3 DATA AND THE ARC INDICATOR SYSTEM

3.1 Sample and Enforcement Labels

The sample combines annual financial-statement data of U.S. listed firms with AAER-based labels, as released in the public replication package of Bao et al. [11]. The accounting data comprise 28 items reported in 10-K filings and standardized in Compustat (e.g., total assets, receivables, sales, cost of goods sold, income before extraordinary items, current assets and liabilities) and the 14 financial ratios of Dechow et al. [4]. A firm-year is labeled as misstated when an AAER identifies it as affected by a material misstatement. The releases originate from the database of the Center for Financial Reporting and Management, extended with releases issued until December 2018, which leaves at least ten years between the end of the main test window and the last enforcement release. The data cover fiscal years 1990–2014; 1990 serves only to construct lagged indicators. Table 1 summarizes the sample windows.

Table 1 Sample Composition and Evaluation Windows

Window	Fiscal years	Firm-years	AAER firm-years
Full sample	1990–2014	146,045	964
Training (expanding)	1991–($t-2$)	67,166–96,919	499–781†
Validation (tuning)	2000–2001	13,114	167
Validation (ensemble)	1999–2002	26,009	317
Main test	2003–2008	35,166	261
Extended test	2009–2014	33,064	112

†Counts before serial-fraud relabeling for test years 2003 and 2008. The 964 misstated firm-years stem from 412 AAER cases.

3.2 Indicator Families

Table 2 lists the indicator families. Flow variables are scaled by average total assets, growth rates use the arc growth rate, and ratio indicators are winsorized at the 1st and 99th percentiles within each fiscal year. Total accruals and the operating cash-flow proxy follow the balance-sheet approach [5, 20]:

$$TACC_t = (\Delta WC_t - DEP_t) / \bar{A}T_t \tag{1}$$

$$CFO_t = IB_t / \bar{A}T_t - TACC_t \tag{2}$$

where $\Delta WC_t = (\Delta CA_t - \Delta Cash_t) - (\Delta CL_t - \Delta STD_t - \Delta TP_t)$ is the change in non-cash working capital; *CA*, *Cash*, *CL*, *STD*, and *TP* denote current assets, cash and short-term investments, current liabilities, debt in current liabilities, and income taxes payable, *DEP* is depreciation and amortization, *IB* is income before extraordinary items, and $\bar{A}T_t$ is average total assets. Growth is measured by the arc growth rate

$$g(x_t) = 2(x_t - x_{t-1}) / (|x_t| + |x_{t-1}|) \tag{3}$$

which is bounded in [−2, 2] and remains defined when the base value is zero or negative. The divergence family (DIV) encodes relations across statements. The earnings–cash divergence ratio

$$ECR_t = (IB_t - OCF_t) / (|IB_t| + |OCF_t|) \tag{4}$$

measures the share of earnings not backed by operating cash flow *OCF*, and gap indicators contrast the growth of receivables with that of sales, sales with cash sales, inventory with cost of goods sold, assets with sales, and sales and earnings with operating cash flow. Abnormal accruals are the residuals of a modified Jones model with a performance control [8-9], estimated by least squares for each fiscal year:

$$ta_t = \alpha_0 + \alpha_1/A_{t-1} + \alpha_2\Delta cr_t + \alpha_3 ppe_t + \alpha_4 ROA_t + \varepsilon_t \tag{5}$$

where ta_t, Δcr_t, and ppe_t denote total accruals, the change in cash revenue ($\Delta REV_t - \Delta REC_t$), and gross property, plant, and equipment, each scaled by lagged total assets A_{t-1}. The temporal family (TMP) captures accumulation and persistence through lagged and two-year cumulative accruals, accrual changes, lagged divergence measures, three-year volatilities of cash flow and sales growth, and the number of positive-accrual years among the last three. The anomaly-lens family (LENS) contains ECOD scores per family and for all 53 ARC indicators within each fiscal year (Section 4.2) and the mean absolute deviation of first significant digits from Benford's law for the 28 items of the current year and of the current and previous year [21].

Table 2 ARC Indicator Families

Family	No.	Indicators
ACC	6	WC accruals; RSST accruals; Δreceivables; Δinventory; soft assets; total accruals *TACC* (1)
REV	8	Arc sales growth and its lag; growth acceleration; %Δcash sales; Δ(receivables/sales); Δgross margin; asset turnover and its change
CFO	6	*CFO* (2) and its change; *CFO*/sales; Δcash margin; Δfree cash flow; cash/assets
FIN	16	ROA and ΔROA; EBIT/TA; RE/TA; book-to-market; depreciation index; actual issuance; equity and debt issuance/TA; leverage and its change; current ratio; log TA; log market value; asset growth; interest/TA
DIV	8	Abnormal accruals (5); *ECR* (4); growth gaps: receivables–sales, sales–cash sales, inventory–COGS, assets–sales, sales–*CFO*, earnings–*CFO*
TMP	9	Lagged and two-year cumulative *TACC*; Δ*TACC*; lagged abnormal accruals, receivable gap, and *ECR*; three-year volatility of *CFO* and sales growth; positive-accrual streak
LENS	8	ECOD scores for ACC, REV, CFO, FIN, DIV, and all ARC indicators; Benford first-digit deviation (one and two years)

Numbers in parentheses refer to equations. ACC–TMP form the 53 ARC indicators; LENS scores are unsupervised summaries computed within each fiscal year.

3.3 Univariate Evidence

Figure 2 reports the 20 indicators with the highest direction-adjusted univariate AUC in the pooled 1991–2008 sample, computed on within-year percentile ranks. Firm size dominates (market value 0.693, total assets 0.668), consistent with the enforcement focus on economically significant firms. Among accounting signals, the share of soft assets (0.638) and the two-year cumulative accruals (0.627) are the strongest; both the lagged accruals (0.611) and the cumulative accruals carry more information than current accruals (0.586), indicating that misstatements build up over several reporting periods. Growth indicators—asset growth (0.608), cash-sales growth (0.606), and lagged sales growth (0.601)—follow closely. Averaged by family, accruals lead (mean AUC 0.588), followed by temporal and financial-ratio indicators (0.567 each), revenue (0.547), cash flow (0.528), and divergence indicators (0.522). The ECOD lenses are individually weak (0.50–0.55), and the Benford deviation is lower, not higher, for misstated firm-years (AUC 0.561 in the negative direction).

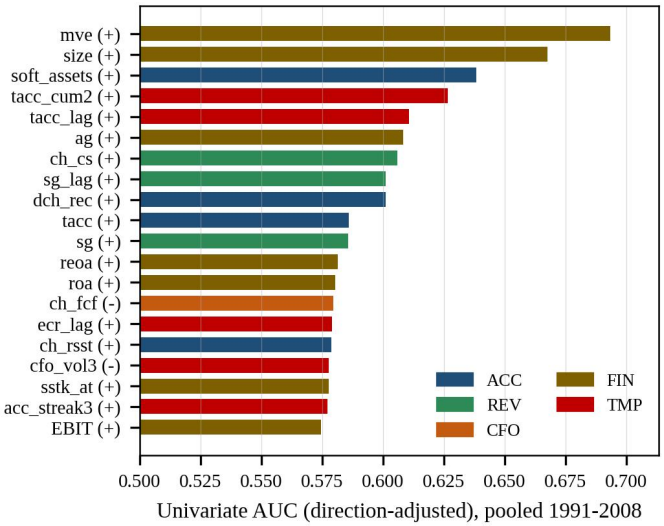


Figure 2 Top-20 Indicators by Direction-Adjusted Univariate AUC (Pooled 1991–2008, Within-Year Percentile Ranks)
The sign in parentheses gives the direction of association with misstatement; colors indicate families.

4 METHODOLOGY

4.1 Problem Formulation and Normalization

Let x_{it} denote the indicator vector of firm i in fiscal year t and $y_{it} \in \{0, 1\}$ its AAER label. The task is to produce a risk score s_{it} whose ranking places misstated firm-years at the top of each year's cross-section, matching the use of such scores for allocating investigative and audit resources. To remove level shifts across years, every indicator is mapped to its within-year percentile rank

$$r_{itj} = \text{rank}(x_{itj}) / N_t \tag{6}$$

where N_t is the number of firms in year t. The transformation requires only the cross-section of filings available at the end of each filing season.

4.2 Multi-Lens Anomaly Scoring

ECOD scores a firm-year by the tail probabilities of its indicators under the empirical cumulative distribution functions (ECDFs) of the cross-section [15]:

$$O_i = \max\{U^L_i, U^R_i, U^S_i\} \quad (7)$$

where $U^L_i = -\Sigma_j \log F^L_j(x_{ij})$ and $U^R_i = -\Sigma_j \log F^R_j(x_{ij})$ aggregate left- and right-tail probabilities obtained from the ECDFs $F^L_j(x) = P(X_j \le x)$ and $F^R_j(x) = P(X_j \ge x)$, and U^S_i uses for each dimension the tail indicated by its sample skewness. ARC-Lens computes O_i separately for the ACC, REV, CFO, FIN, and DIV families and for all ARC indicators within each fiscal year. The lens scores serve two purposes: as stand-alone unsupervised detectors and as conditioning inputs that summarize how unusual a firm is along each economic dimension.

4.3 Learners and the ARC-Lens Ensemble

The supervised component of ARC-Lens combines two learners trained on 89 inputs (28 raw items, 53 ARC indicators, and 8 lens scores) in percentile-rank form. The first is a LightGBM ensemble with 15 leaves [26], at least 200 observations per leaf, learning rate 0.03, 200 trees, column subsampling 0.5, row subsampling 0.8, and L2 regularization 5. The second is a multilayer perceptron with piecewise linear encoding (PLE-MLP) [18]. Each input is encoded with $T = 8$ quantile bins estimated on the training data,

$$e_t(x) = \min\{1, \max\{0, (x - b_{t-1}) / (b_t - b_{t-1})\}\} \quad (8)$$

for $t = 1, \ldots, T$, where $b_0 < \ldots < b_T$ are the bin edges; missing values receive a zero encoding and a missingness indicator. The encoded vector feeds an MLP with two hidden layers of 128 and 64 ReLU units, trained with Adam (learning rate 10^{-3}, batch size 512, 10 epochs, L2 penalty 10^{-4}) and a positive-class weight of 10. Each learner is averaged over three random seeds, and the scores are combined by rank averaging,

$$s_i = w \cdot \rho^G_i + (1 - w) \cdot \rho^M_i \quad (9)$$

where ρ^G_i and ρ^M_i are the normalized within-year ranks of the GBDT and PLE-MLP scores and $w = 0.5$ is selected on validation years. The combination follows the evidence that tree and neural learners make complementary errors on tabular data [19].

4.4 Benchmarks

All benchmarks follow the identical protocol. Two unsupervised detectors, ECOD-ARC and IForest-ARC [14], are applied to the 53 ARC indicators of each test year. Logit-FScore is a logistic regression on the seven variables of the F-score model [4]. RUSBoost-28 reproduces the model of Bao et al. on the 28 raw items with AdaBoost.M2 pseudo-loss [11], random under-sampling to a 1:1 ratio, CART base learners with a minimum leaf size of 5, learning rate 0.1, and 1,000 learners selected on validation [23]. LGB-42 is LightGBM on the 28 raw items and 14 ratios used by Bao et al. Logit-ARC is an L2-regularized logistic regression on the 53 ARC indicators. The individual ARC-Lens components, ARC-GBDT and PLE-MLP, are reported as well.

4.5 Evaluation Protocol

Following Bao et al. [11], models for test year t are trained on fiscal years 1991 to $t - 2$. The 24-month gap reflects the delay between fiscal year-end and the availability of complete financial statements and enforcement information. Training firm-years that belong to an AAER case also present in the test year are relabeled as non-misstated, because the misstatement had not been revealed at prediction time. Hyperparameters were selected on validation years 2000–2001 and the ensemble composition and weight on 1999–2002, where the selected ensemble attains an AUC of 0.763 (RUSBoost-28: 0.702); no test-year information enters model design. The main test window is 2003–2008 (35,166 firm-years, 261 misstated); 2009–2014 serves as an extended window. Performance is measured by the area under the ROC curve (AUC), average precision (AP), the normalized discounted cumulative gain at the top 1% of firm-years (NDCG@1%), and the sensitivity at the top 5%. Differences in year-averaged AUC are tested with a paired stratified bootstrap with 1,000 resamples of firm-years within each test year.

4.6 Interaction Analysis

An XGBoost model [25] with 300 trees of depth 4 is fitted on the 53 ARC indicators of the pooled 1991–2008 sample, and TreeSHAP interaction values φ_{jk} [27] are computed for all 837 misstated firm-years and 4,000 randomly drawn other firm-years. Interaction strength is aggregated to family pairs (a, b) as

$$I_{ab} = \Sigma_{j \in a,\, k \in b} E|\varphi_{jk}| \,/\, \Sigma_{j \ne k} E|\varphi_{jk}| \quad (10)$$

with the sum taken over unordered indicator pairs. Conditional AAER rates over accrual and growth terciles and a logit model with interaction terms and firm-clustered standard errors complement the attribution analysis.

5 RESULTS

5.1 Out-of-Sample Detection

Table 3 reports out-of-sample performance. The two unsupervised detectors rank misstated firm-years at chance level (AUC 0.499 and 0.483 in 2003–2008), far below every supervised model. The F-score logit attains an AUC of 0.666.

Replacing its seven ratios by the 53 ARC indicators lifts a linear model to 0.702, and adding the ARC indicators and lens scores to the 42 established inputs raises LightGBM from 0.697 to 0.714. ARC-Lens records the highest mean AUC in the main window (0.717) and the highest pooled AUC (0.7223 versus 0.7219 for RUSBoost-28). Its yearly AUC varies less than that of RUSBoost-28 and LGB-42 (standard deviation 0.026 versus 0.039 and 0.037), and it captures 14.9% of misstated firm-years within the top 5% of scores, compared with 10.7% for RUSBoost-28 and 10.8% for the F-score; Logit-ARC reaches 16.8% on this criterion. At the narrow top-1% budget the F-score logit attains the highest NDCG; with about 58 firm-years per year in that segment, this metric rests on one or two cases per year and fluctuates strongly between years.

Table 3 Out-of-Sample Detection Performance

Model	AUC	SD	AP (%)	NDCG @1%	Sens. @5% (%)	AUC 2009–14	AUC 2003–14
ECOD-ARC [15]	0.499	0.055	0.73	0.002	2.9	0.467	0.483
IForest-ARC [14]	0.483	0.056	0.69	0.000	3.6	0.483	0.483
Logit-FScore [4]	0.666	0.026	1.73	**0.038**	10.7	0.715	0.691
RUSBoost-28 [11]	0.715	0.039	1.77	0.021	10.6	**0.739**	**0.727**
LGB-42 [26]	0.697	0.037	1.67	0.020	10.1	0.725	0.711
Logit-ARC	0.702	**0.025**	1.86	0.019	**16.8**	0.718	0.710
PLE-MLP [18]	0.709	0.028	**2.05**	0.032	13.5	0.702	0.705
ARC-GBDT	0.714	0.030	1.80	0.013	14.2	0.712	0.713
ARC-Lens (ours)	**0.717**	0.026	1.84	0.020	14.9	0.714	0.715

Columns 2–6 refer to the main window 2003–2008 (means over test years; SD is the standard deviation of yearly AUC). Best value per column in bold.

Figure 3 shows the yearly AUC. In the main window ARC-Lens stays at or above 0.680 in every year, while RUSBoost-28 ranges from 0.663 (2008) to 0.770 (2004). In the extended window RUSBoost-28 attains the highest AUC (0.739), whereas ARC-Lens attains the highest NDCG@1% (0.036 versus 0.030) and top-1% sensitivity (5.8% versus 4.4%). Over 2003–2014 ARC-Lens has the lowest year-to-year standard deviation of all models (0.023 versus 0.040 for RUSBoost-28).

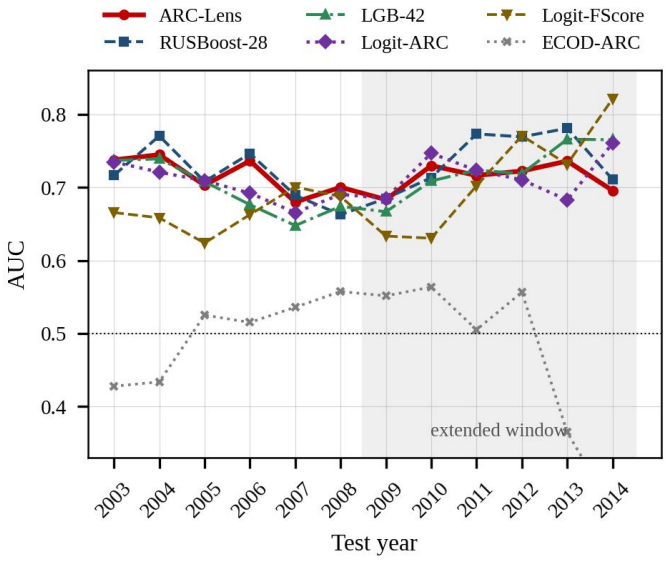


Figure 3 Out-of-Sample AUC by Test Year

The shaded area marks the extended window (2009–2014); the dotted line indicates chance level.

5.2 Statistical Significance

Table 4 reports paired bootstrap tests. In the main window ARC-Lens significantly outperforms the F-score (ΔAUC = 0.051, $p \leq 0.001$), LGB-42 (0.020, $p = 0.026$), its PLE-MLP component (0.008, $p = 0.042$), and both unsupervised detectors (≥0.218, $p \leq 0.001$); the difference to RUSBoost-28 is not significant (0.001, $p = 0.886$). Over 2003–2014 the advantage over the unsupervised detectors and PLE-MLP persists, the difference to the F-score is marginal (0.025, $p = 0.056$), and ARC-Lens and RUSBoost-28 remain statistically on par (−0.012, $p = 0.324$).

Table 4 Paired Bootstrap Tests of ΔAUC (ARC-Lens Minus Benchmark)

Benchmark	ΔAUC 2003–08	95% CI	*p*	ΔAUC 2003–14	*p*
Logit-FScore	+0.050	[+0.020, +0.083]	≤0.001	+0.025	0.056
RUSBoost-28	+0.001	[−0.021, +0.025]	0.886	−0.012	0.324
LGB-42	+0.020	[+0.003, +0.036]	0.026	+0.005	0.630
Logit-ARC	+0.015	[−0.006, +0.035]	0.174	+0.005	0.472

Benchmark	ΔAUC 2003–08	95% CI	*p*	ΔAUC 2003–14	*p*
PLE-MLP	+0.008	[+0.000, +0.016]	0.042	+0.010	0.006
ARC-GBDT	+0.003	[−0.005, +0.010]	0.496	+0.002	0.548
ECOD-ARC	+0.218	[+0.175, +0.260]	≤0.001	+0.232	≤0.001
IForest-ARC	+0.234	[+0.190, +0.278]	≤0.001	+0.232	≤0.001

Stratified resampling of firm-years within each test year (1,000 replications); positive values favor ARC-Lens.

5.3 Contribution of Indicator Families

Figure 4 summarizes the ablation study, run with the GBDT component of ARC-Lens and identical seeds. Removing the accrual family causes the largest loss (−0.010 AUC), followed by the raw items (−0.006), the revenue family (−0.005), the temporal family (−0.003), the anomaly lenses (−0.003), and the divergence family (−0.001); replacing within-year ranks by raw values costs 0.002. Removing the cash-flow or financial-ratio family slightly increases AUC (+0.001 and +0.004), so their information is largely contained in the accrual, growth, and size-related inputs once these are present. Adding the temporal family to the 42 established inputs raises AUC from 0.697 to 0.705, whereas the divergence family alone adds little (0.697). Single-family models confirm the ranking: accruals (0.659) and revenue growth (0.658) are the most informative families, followed by temporal (0.652), financial-ratio (0.644), cash-flow (0.613), and divergence indicators (0.585), and the 53 ARC indicators jointly reach 0.708.

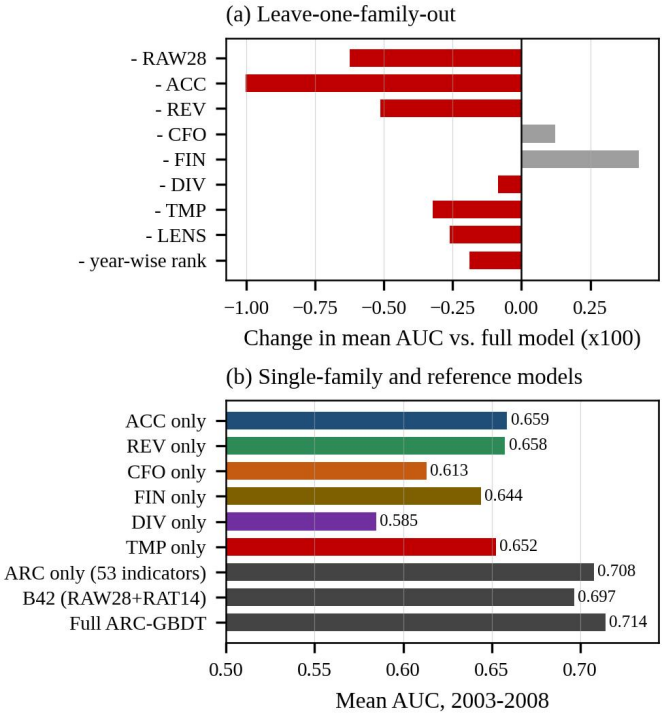


Figure 4 Indicator-Family Ablation with the GBDT Component of ARC-Lens (2003–2008): (a) Change in Mean AUC When a Family Is Removed; (b) AUC of Single-Family and Reference Models

5.4 Misreporting Is Not Statistical Outlyingness

The failure of ECOD and isolation forests is informative. Both detectors rank firms by their distance from the cross-sectional bulk, which places small, distressed, and volatile firms at the top. AAER firm-years, in contrast, belong to larger, growing, and financing-active firms whose ratios lie inside the typical range (Figure 2); their irregularity resides in conditional patterns—how much accruals accumulate given growth and scale—rather than in extreme values. The same lens scores add 0.003 AUC when used as conditioning inputs of the supervised model, which is consistent with the decisive role of label information documented by ADBench and with the weak stand-alone performance of digit-level measures [16, 22].

5.5 Interrelationships Among Indicators

Pairwise interactions account for 58.9% of the TreeSHAP attribution, confirming that the indicators act jointly rather than additively. Figure 5 aggregates interaction strength to family pairs. Pairs involving the financial-ratio family, which contains firm size, market value, equity issuance, and liquidity, account for 62.0% of the total, led by ACC–FIN (13.4%) and REV–FIN (9.8%); among the remaining pairs, ACC–REV (4.7%), ACC–TMP (3.7%), and REV–TMP (3.6%) are the largest. Table 5 lists the strongest individual pairs: the soft-asset share interacts with retained earnings, asset turnover, market value, equity issuance, and liquidity, and total accruals interact with changes in leverage. Accrual

signals therefore carry different meaning for firms of different scale and financing needs, in line with the financing motive for manipulation [1].

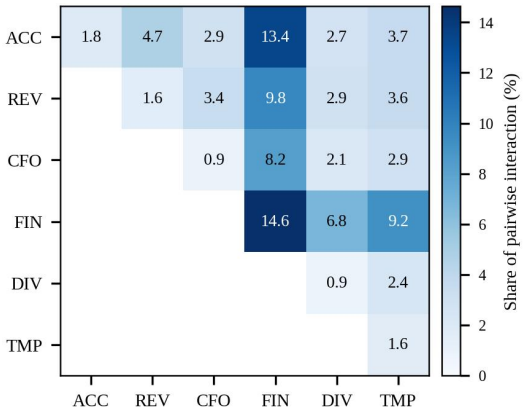


Figure 5 Family-Level TreeSHAP Interaction Map

Share (%) of total pairwise interaction strength attributable to each family pair (upper triangle; the diagonal shows within-family interactions).

Table 5 Strongest Pairwise TreeSHAP Interactions

Indicator pair	Families	Share (%)
Soft assets × RE/TA	ACC–FIN	0.95
Soft assets × Asset turnover	ACC–REV	0.88
Current ratio × log TA	FIN–FIN	0.87
Soft assets × log MVE	ACC–FIN	0.85
Soft assets × Equity issuance/TA	ACC–FIN	0.65
Soft assets × Current ratio	ACC–FIN	0.64
Cash/TA × log MVE	CFO–FIN	0.61
Soft assets × log TA	ACC–FIN	0.60
TACC × ΔLeverage	ACC–FIN	0.51
EBIT/TA × log MVE	FIN–FIN	0.51

Share of total pairwise interaction strength among the 1,378 pairs of the 53 ARC indicators.

Figure 6 displays AAER rates across terciles of two-year cumulative accruals and sales growth, split at the median cash flow. The base rate is 0.83%. Firm-years in the top tercile of both accruals and growth reach 1.43% (low cash flow) and 1.84% (high cash flow), 1.7 and 2.2 times the base rate and five to seven times the rate of firm-years in the bottom tercile of both (0.25–0.28%). Within the high-accrual tercile the rate increases monotonically with growth in both cash-flow groups. Table 6 shows that in a logit model with firm-clustered standard errors, cumulative accruals ($z = 3.41$) and sales growth ($z = 2.85$) are significant, while their interaction terms with each other and with cash flow are not significant at the 5% level (sales growth × cash flow: $z = 1.71$, $p = 0.088$). Accruals and growth thus compound multiplicatively in the odds of misreporting, whereas the non-linear interactions that matter most involve firm scale and financing, which tree ensembles capture and linear indicator models miss.

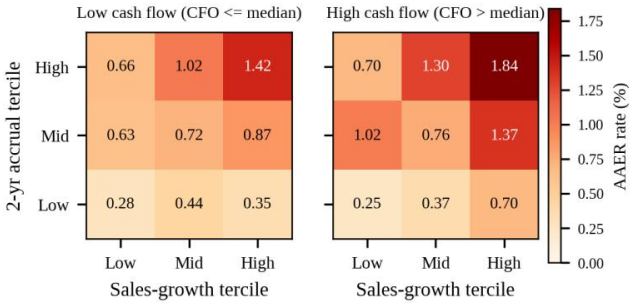


Figure 6 AAER Rate (%) by Terciles of Two-Year Cumulative Accruals and Sales Growth for Firm-Years with Low and High Operating Cash Flow (1991–2008, Base Rate 0.83%)

Table 6 Logit Model with Interaction Terms

Variable	Coef.	*z*	*p*
Intercept	-5.419	-50.69	<0.001
Cumulative accruals (*A*)	0.243	3.41	<0.001
Sales growth (*G*)	0.183	2.85	0.004
Cash flow (*C*)	-0.106	-1.32	0.188
A × *G*	0.040	0.67	0.504

Variable	Coef.	z	p
$A \times C$	0.084	1.48	0.140
$G \times C$	0.103	1.71	0.088
$A \times G \times C$	-0.048	-0.98	0.326
Soft assets	0.560	5.87	<0.001
log market value	0.883	9.36	<0.001

n = 72,972 firm-years (1991–2008), pseudo-R^2 = 0.083. Regressors are standardized within-year percentile ranks; standard errors are clustered by firm.

6 DISCUSSION AND IMPLICATIONS

Three implications follow for the design of accounting-based screening tools. First, the information content of 10-K fundamentals is concentrated in accrual behavior over time: two-year accumulation, lagged accruals, and the soft-asset share consistently outrank contemporaneous ratios, and the temporal family delivers the largest increment to the established inputs. Screening systems should therefore track the trajectory of accruals rather than a single year's level. Second, indicator design matters at least as much as model choice: a regularized logit on ARC indicators exceeds the AUC of LightGBM on the 42 established inputs and attains the highest top-5% sensitivity in the main window, which offers regulators a transparent and auditable model. Third, unsupervised outlier detection is not a substitute for supervision. Because misreporting firms resemble healthy growth firms, anomaly scores are most useful as conditioning features inside supervised learners, as in ARC-Lens.

Operationally, ARC-Lens provides a stable ranking across years and captures about one in seven misstated firm-years within the top 5% of scores, which supports the prioritization of reviews when investigative capacity is fixed. The framework relies solely on standardized annual filing data and therefore extends directly to the XBRL financial-statement data distributed through SEC EDGAR, to quarterly filings, and to additional label sources such as intentional restatements; combining ARC indicators with textual signals from 10-K narratives [24] is a natural direction for further gains.

7 CONCLUSION

This paper examined whether accruals, revenue growth, operating cash flow, financial ratios, and their interrelationships identify financial reporting irregularities in U.S. public companies. Using 146,045 firm-years of 10-K fundamentals labeled with SEC enforcement releases and a rolling protocol free of look-ahead and serial-fraud leakage, the study shows that (i) a structured ARC indicator system improves both linear and tree-based detectors, with the largest gains from accrual accumulation; (ii) the ARC-Lens ensemble of gradient boosting and a PLE-encoded MLP matches the strongest published benchmark while capturing more misstatements within the top 5% of scores; (iii) unsupervised outlier detection fails because misstated firm-years are statistically typical; and (iv) the indicators operate jointly, with more than half of the model attribution arising from interactions, chiefly between accrual signals and firm scale and financing. These findings support accounting-based screening that tracks accrual trajectories and models their conditional relations rather than isolated ratios.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Dechow P M, Sloan R G, Sweeney A P. Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the SEC. Contemporary Accounting Research, 1996, 13(1): 1-36.
[2] Karpoff J M, Koester A, Lee D S, et al. Proxies and databases in financial misconduct research. The Accounting Review, 2017, 92(6): 129-163.
[3] Beneish M D. The detection of earnings manipulation. Financial Analysts Journal, 1999, 55(5): 24-36.
[4] Dechow P M, Ge W, Larson C R, et al. Predicting material accounting misstatements. Contemporary Accounting Research, 2011, 28(1): 17-82.
[5] Sloan R G. Do stock prices fully reflect information in accruals and cash flows about future earnings? The Accounting Review, 1996, 71(3): 289-315.
[6] Richardson S A, Sloan R G, Soliman M T, et al. Accrual reliability, earnings persistence and stock prices. Journal of Accounting and Economics, 2005, 39(3): 437-485.
[7] Jones J J. Earnings management during import relief investigations. Journal of Accounting Research, 1991, 29(2): 193-228.
[8] Dechow P M, Sloan R G, Sweeney A P. Detecting earnings management. The Accounting Review, 1995, 70(2): 193-225.
[9] Kothari S P, Leone A J, Wasley C E. Performance matched discretionary accrual measures. Journal of Accounting and Economics, 2005, 39(1): 163-197.

[10] Perols J. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. Auditing: A Journal of Practice & Theory, 2011, 30(2): 19-50.
[11] Bao Y, Ke B, Li B, et al. Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. Journal of Accounting Research, 2020, 58(1): 199-235.
[12] Bertomeu J, Cheynel E, Floyd E, et al. Using machine learning to detect misstatements. Review of Accounting Studies, 2021, 26(2): 468-519.
[13] Achakzai M A K, Peng J. Detecting financial statement fraud using dynamic ensemble machine learning. International Review of Financial Analysis, 2023, 89: 102827.
[14] Liu F T, Ting K M, Zhou Z H. Isolation forest. In: Proceedings of the 8th IEEE International Conference on Data Mining (ICDM), 2008: 413-422.
[15] Li Z, Zhao Y, Hu X, et al. ECOD: Unsupervised outlier detection using empirical cumulative distribution functions. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(12): 12181-12193.
[16] Han S, Hu X, Huang H, et al. ADBench: Anomaly detection benchmark. In: Advances in Neural Information Processing Systems (Datasets and Benchmarks Track), 2022, 35.
[17] Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? In: Advances in Neural Information Processing Systems (Datasets and Benchmarks Track), 2022, 35.
[18] Gorishniy Y, Rubachev I, Babenko A. On embeddings for numerical features in tabular deep learning. In: Advances in Neural Information Processing Systems, 2022, 35.
[19] Shwartz-Ziv R, Armon A. Tabular data: Deep learning is not all you need. Information Fusion, 2022, 81: 84-90.
[20] Hribar P, Collins D W. Errors in estimating accruals: Implications for empirical research. Journal of Accounting Research, 2002, 40(1): 105-134.
[21] Amiram D, Bozanic Z, Rouen E. Financial statement errors: Evidence from the distributional properties of financial statement numbers. Review of Accounting Studies, 2015, 20(4): 1540-1593.
[22] Walker S. Compared to what? Does Benford's law really detect corporate fraud? Econ Journal Watch, 2022, 19(1): 30-46.
[23] Seiffert C, Khoshgoftaar T M, Van Hulse J, et al. RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 2010, 40(1): 185-197.
[24] Brown N C, Crowley R M, Elliott W B. What are you saying? Using topic to detect financial misreporting. Journal of Accounting Research, 2020, 58(1): 237-291.
[25] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 785-794.
[26] Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems, 2017, 30: 3146-3154.
[27] Lundberg S M, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2020, 2(1): 56-67.