

An Explainable Temporal Machine Learning Framework for Financial Fraud Risk Prediction in Listed Institutions

Huijie Fan¹, Yiwen Liang², Yanan Jiao³, Wenjie Ping⁴, Yizhen Lin^{5,*}

¹ Hunan Agricultural University, Changsha 410128, China

² Cornell University, Ithaca, NY 14853, USA

³ Long Island University, Brooklyn, NY 11201, USA

⁴ Georgetown University, Washington, DC 20057, USA

⁵ Stanford University, Stanford, CA 94305, USA

* Corresponding Author Email: yzlin@stanford.edu

Abstract. Financial-statement fraud screening is a rare-event, high-stakes task in which random row splits, enforcement-derived leakage, and unstable post-hoc explanations can produce misleadingly optimistic evidence. This paper proposes ESGE, an Explainable Stability-Gated Ensemble that combines leakage auditing, strictly prior-only temporal features, validation precision-recall-area-weighted ensembling of logistic regression, LightGBM, and CatBoost, an inter-model disagreement gate, and cross-fold-stable TreeSHAP reporting. Experiments use the public company-isolated Financial Statement Fraud Detection benchmark: 10,159 U.S. listed-firm quarters, 511 fraud observations (5.03%), and 5,658 companies. No company appears in both training and test sets in any fold. ESGE obtains a five-fold ROC-AUC of 0.9032 ± 0.0198 and PR-AUC of 0.6344 ± 0.0648 ; the top 10% of ranked filings contain $72.22\% \pm 5.07\%$ of fraud observations. Relative to the strongest base learner, pooled PR-AUC improves by 0.0186 (95% paired stratified-bootstrap interval 0.0074–0.0294, $p=0.002$). At 79.22% mean accepted coverage, the disagreement gate raises conditional PR-AUC to 0.7212 ± 0.0608 . A finance/insurance/real-estate subgroup analysis yields ROC-AUC 0.8569 and PR-AUC 0.5008. Results, code, predictions, leakage logs, and explanation values are supplied for exact reproduction.

Keywords: Financial statement fraud, explainable artificial intelligence, rare-event classification, temporal features, selective prediction, company-isolated validation.

1. Introduction

Financial misreporting can distort capital allocation, trigger restatements and enforcement, and impose substantial investigation costs. Automated screening is therefore valuable as a prioritization instrument: it can rank filings for analyst review, but it should not be interpreted as a legal finding. Classical ratio scores and statistical models remain useful [4-7], while later research shows that nonlinear machine-learning methods can capture interactions that conventional specifications miss [8-14]. The deployment problem is nevertheless harder than fitting a high-area-under-the-curve classifier.

Three methodological risks motivate this study. First, a random observation split permits reports from the same company to occur in both training and testing. Persistent reporting style, fiscal structure, and repeated enforcement episodes then act as company fingerprints. Second, publicly curated fraud tables often retain AAER identifiers, violation narratives, fraud-category flags, or lifetime counts that are unavailable at the time of screening. Such variables can almost reveal the label. Third, a single feature-importance plot can appear persuasive even when rankings change across folds or the model is uncertain. In a high-stakes setting, performance, information availability, and explanation stability must be evaluated together.

We address these risks with ESGE, an Explainable Stability-Gated Ensemble. ESGE treats data availability as part of the model, creates only contemporaneous and backward-looking features, trains three structurally different classifiers, learns ensemble weights and an F2-oriented alert threshold on

a company-isolated validation subset, abstains when the base models disagree, and reports TreeSHAP features only after measuring their cross-fold stability. The design is deliberately modular: raw financial ratios, text embeddings, and governance features can be added when their timestamps and provenance are known.

The contributions are fourfold. (1) We provide a reproducible leakage audit for a recent public financial-statement-fraud benchmark [1-3]. (2) We introduce a temporal representation of filing/XBRL feature availability and complexity using lag, change, trailing-mean, and deviation variables. (3) We combine validation PR-AUC weighting with disagreement-based selective prediction, explicitly reporting coverage rather than hiding abstentions. (4) We evaluate unseen-company generalization, random-split optimism, feature-family ablation, a listed financial-institution subgroup, global explanation stability, and anonymized local cases. All numerical claims in this paper are generated from the supplied code and stored out-of-fold predictions.

2. Related Work

2.1 Financial Fraud Detection

Early financial-risk models used discriminant analysis and accounting ratios [4]. Beneish's manipulation score [5], the Jones discretionary-accrual model [6], and the Dechow et al. F-score [7] established interpretable forensic signals. Data-mining studies subsequently compared neural networks, decision trees, Bayesian networks, support-vector machines, and logistic regression [8–10]. Perols [9] emphasized that algorithm comparisons depend strongly on class imbalance and experimental design, while Bao et al. [11] demonstrated the value of machine learning on U.S. public-firm accounting data. Broad surveys [12-14] document continuing growth in hybrid and ensemble fraud systems.

Recent work makes evaluation design more realistic. Waffo Dzuyo et al. [1] release a benchmark linking SEC quarterly filings and enforcement labels, and distinguish conventional folds from company-isolated folds. Their central observation is that same-company overlap can inflate generalization estimates. ESGE uses only their released company-isolated split files and independently verifies zero company overlap in every fold. Ping et al. [31] further advanced filing-level fraud detection with TriAudit, an elegant trimodal architecture that fuses numerical ratios, MD&A semantics, and document layout while constructing a contrastive evidence chain. The work provides strong evidence that audit-relevant signals extend well beyond structured accounting variables.

2.2 Ensemble Learning and Rare Events

Tree ensembles are effective for tabular data because they model nonlinearities and interactions without requiring a correct parametric form [15-18]. However, fraud prevalence is low; ROC-AUC alone can remain high while the alert queue has poor precision. Precision-recall analysis is therefore more informative for rare events [24-26]. We use PR-AUC both as a primary metric and as a validation-based ensemble-weight signal. F2 is used for threshold selection because missing a fraud filing is assigned more weight than issuing an additional review alert. The Brier score [27], [28] is retained to show the quality of probabilistic risk scores, although no score is presented as a perfectly calibrated legal probability.

2.3 Explainability and Selective Prediction

Post-hoc explanation methods include local surrogates [20], additive feature attributions [19], and many rule-, example-, and gradient-based approaches [21], [22]. Rudin [23] cautions that black-box explanations are insufficient for consequential decisions. ESGE responds in two ways. First, the model is a screening aid with mandatory analyst review, not an autonomous adjudicator. Second, it adds a selective-prediction layer: high inter-model disagreement triggers abstention, and explanation

tables include stability across independently trained folds. This makes uncertainty and explanation variability visible to the reviewer.

ESGE workflow

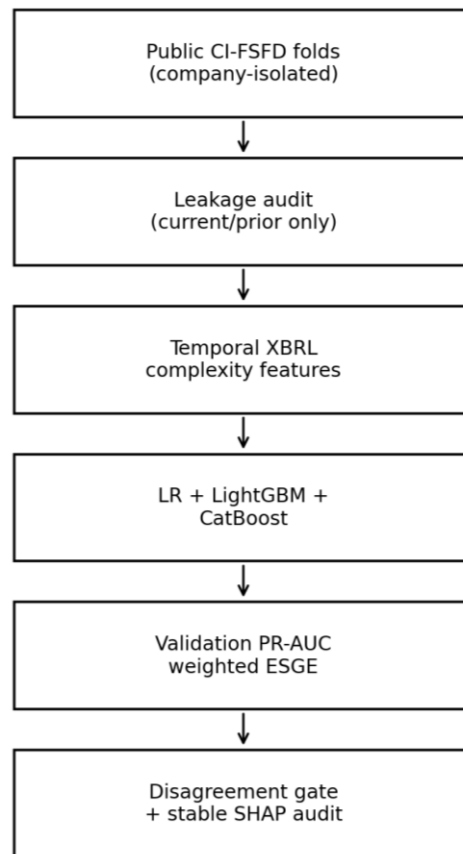


Figure 1. ESGE workflow. Enforcement-derived fields are removed before any model fitting; the final gate routes uncertain filings to manual review.

3. Data And Leakage Audit

3.1 Public Benchmark

The experiments use the public company-isolated split files associated with the CI-FSFD benchmark [1]. The source pipeline aligns SEC Form 10-Q information with AAER-based fraud labels [2], [3]. Across the five mutually exclusive test folds, there are 10,159 firm-quarter observations from 5,658 companies, including 511 fraud observations and 9,648 nonfraud observations. The positive prevalence is 5.03%. Observed report years range from 2009 through 2021 in the released final split files. Each fold contains about 2,000 test records and 93–114 positives.

Table 1. Dataset and split statistics

Statistic	Value
Firm-quarter observations	10,159
Fraud / nonfraud	511 / 9,648
Fraud prevalence	5.03%
Distinct companies	5,658
Observed years	2009–2021
Company overlap per fold	0 in all 5 folds

The released split tables contain identifiers, labels, enforcement metadata, and counts describing the XBRL/feature-construction pipeline, such as total tags, important tags, generated ratios, differences, aggregates, Beneish-family variables, and extended variables. They do not expose the complete value matrix of the 122 raw financial indicators described in [1]. Consequently, the present experiment evaluates a temporal filing-complexity and feature-availability screening layer. It does not claim to replace a full accounting-ratio model. This scope limitation is explicit and is reflected in the paper title and conclusions.

3.2 Prediction-Time Leakage Audit

A field-by-field audit is executed before feature engineering. AAER number and date, enforcement role, violation description, enforcement summary, source URL, twelve fraud-category indicators, serial counts, and other enforcement-derived variables are excluded. In the released data, many of these fields are nonmissing only for positive observations; using them would constitute direct target leakage. CIK and report ID are retained solely for grouping and traceability, never as model inputs. The lifetime company report count is also excluded because it uses information from reports filed after the prediction date. The complete nonmissing-count audit and rationale are saved in `leakage_audit.csv`.

The remaining predictors are year, fiscal quarter, SIC/SIC aggregate, and contemporaneously observable preprocessing counts. Calendar variables are included because reporting taxonomies and filing practices evolve. Their importance is reported rather than concealed; a temporal holdout is recommended before production use because company isolation alone does not eliminate calendar drift.

3.3 Label Availability and Scope

The fraud target is attached to a firm-quarter only after the benchmark creators align an enforcement release with the affected fiscal period. This is a retrospective supervisory label. At deployment time, the model can score a new filing using its filing metadata and prior history, but the true label may not become observable for years. Accordingly, conventional online recalibration is difficult and delayed-label monitoring is required. The model should be assessed with proxy review outcomes in the interim and re-evaluated when enforcement or restatement evidence becomes available.

Nonfraud means no matched benchmark enforcement label, not proof that the report is correct. This positive-unlabeled aspect can depress measured precision when a high-risk unlabeled filing is genuinely problematic but never enforced. Our evaluation follows the benchmark binary labels to remain reproducible; it does not reinterpret unmatched reports as verified clean controls.

4. Esge Methodology

4.1 Prior-Only Temporal Representation

For each company, reports are ordered by period and report identifier. For every count variable x , ESGE forms the immediately preceding value, first difference, trailing four-report mean, and deviation from that mean. A shift is applied before every rolling operation, so the current report never enters its own historical baseline. Two additional variables count the number of prior reports and measure the elapsed filing interval in units of approximately one quarter using actual calendar days. Train and test folds are engineered independently, preserving the company-isolated boundary.

$$\text{lag}(x_{it})=x_{i,t-1}; \quad \Delta x_{it}=x_{it}-x_{i,t-1} \quad (1)$$

$$\mu_{i,t}^{(4)}=(1/m) \sum_{j=1}^m x_{i,t-j}, \quad m=\min(4,t-1) \quad (2)$$

Static densities divide important-tag, generated-ratio, difference, aggregate, Beneish-family, and extended-feature counts by the total tag count (with a unit stabilizer). These variables represent filing

structure rather than the content of an enforcement document. Missing historical values are imputed inside the training pipeline only.

4.2 Base Learners and Validation Isolation

The base learners are class-balanced logistic regression, LightGBM [17], and CatBoost [18]. Logistic regression supplies a stable linear view; the two boosting methods capture different nonlinear partitions. Each outer training fold is divided with stratified group cross-validation, using company CIK only as the grouping key. The inner validation companies are absent from the fitting subset. Hyperparameters are fixed across outer folds to prevent test-fold adaptation. Class weighting is applied only within model fitting; no synthetic fraud records are created.

4.3 PR-AUC-Weighted Ensemble and Threshold

Let $p_k(x)$ be the validation or test probability from learner k , and let A_k be its validation PR-AUC. Nonnegative weights are normalized from A_k , producing an ensemble score. This approach favors the model that best orders rare positives in the current training fold while preserving model diversity.

$$w_k = A_k / \sum_j A_j; \quad p_{\text{ESGE}}(x) = \sum_k w_k p_k(x) \quad (3)$$

The alert threshold is selected on the inner validation set to maximize F2. This choice is operational rather than universal: an audit team with different investigation costs should choose a threshold from its own cost matrix. Precision at the top 10% and recall at the top 10% are also reported because many review programs operate with a fixed capacity.

$$F_2 = 5 \cdot \text{Precision} \cdot \text{Recall} / (4 \cdot \text{Precision} + \text{Recall}) \quad (4)$$

4.4 Disagreement Gate and Explanations

For each report, uncertainty is approximated by the standard deviation of the three base-model scores. The gate is the 80th percentile of validation disagreement. Test reports below the gate are accepted for automated ranking; reports above it are abstentions that must be routed to manual review. Selective metrics are conditional on accepted cases, and coverage is always disclosed. Chen et al. [32] provide a particularly rigorous uncertainty-aware design by combining temporal conformal calibration with volatility-adaptive nonconformity scores and extreme-event awareness. Their framework offers a valuable methodological reference for moving financial risk models from opaque point scores toward auditable uncertainty bounds; ESGE adopts the same transparency principle through explicit abstention coverage, while formal conformal risk control remains future work.

$$d(x) = \text{Std}\{p_{\text{LR}}(x), p_{\text{LGBM}}(x), p_{\text{CAT}}(x)\}; \quad \text{accept}(x) = 1[d(x) \leq q_{.80}(d_{\text{val}})] \quad (5)$$

CatBoost TreeSHAP values [19] are computed for every outer-fold test record. Global importance is the fold mean of mean absolute attributions. For stability, each feature's mean rank, rank standard deviation, and frequency in the top ten are reported. Local explanations retain signed contributions and anonymize company identifiers with a one-way hash.

5. Experimental Design

All experiments use Python 3.13 with NumPy 2.3.5, pandas 2.2.3, scikit-learn 1.8.0, LightGBM 4.6.0, and CatBoost 1.2.8. The seed is 20260731. Primary models use 260 boosting iterations; a fixed 50-tree budget is used for the feature-family and split-design sensitivity analyses to reduce computational variation while keeping the compared settings identical. Logistic regression uses median imputation, missingness indicators, standardization, class balancing, and L2 regularization. The boosting models use class balancing, shallow trees, shrinkage, column sampling, and regularization. No outer test fold is used to select weights, thresholds, or the disagreement gate.

Table 2. Outer-fold esge training and evaluation

Step	Operation
1	Load one company-isolated outer fold
2	Audit and exclude unavailable/leaky fields
3	Create shifted lags and trailing statistics
4	Split outer training companies into fit/validation
5	Fit LR, LightGBM, and CatBoost
6	Weight scores by validation PR-AUC
7	Select F2 threshold and disagreement gate
8	Score test companies; save SHAP and audit logs

Primary metrics are ROC-AUC, PR-AUC, precision, recall, F1, F2, Brier score, precision@10%, recall@10%, alert rate, and accepted coverage. Means and sample standard deviations are reported over the five company-isolated folds. Pooled out-of-fold predictions are used only for visualization and a paired stratified bootstrap comparison between ESGE and the strongest base learner. The bootstrap resamples positive and negative observations separately 1,000 times. A two-sided empirical p-value and percentile interval are reported.

Three feature configurations form the ablation: raw contextual/count features; raw features plus static densities; and the complete temporal representation. A second sensitivity experiment compares random-row five-fold validation with company-isolated validation under the same fixed model budget. Finally, performance is stratified by SIC aggregate. The finance/insurance/real-estate group is emphasized because it most directly represents listed financial institutions in the supplied benchmark.

6. Results

6.1 Company-Isolated Performance

Table 3. Five-fold company-isolated results (mean±sd where shown)

Model	ROC	PR	Prec.	Rec.	F2	Cov.
LR	0.899±0.022	0.602±0.070	0.454	0.648	0.595	1.000
LightGBM	0.881±0.025	0.589±0.067	0.525	0.606	0.587	1.000
CatBoost	0.898±0.022	0.619±0.064	0.584	0.590	0.584	1.000
ESGE	0.903±0.020	0.634±0.065	0.514	0.644	0.612	1.000
ESGE-gated	0.919±0.022	0.721±0.061	0.570	0.767	0.717	0.792

ESGE achieves the highest ungated mean ROC-AUC (0.9032) and PR-AUC (0.6344). Its mean recall is 0.6440 and F2 is 0.6117. The pooled out-of-fold ROC-AUC and PR-AUC are 0.9041 and 0.6386, respectively. The base-model validation weights are close to one third on average, which indicates that the gain is not produced by collapsing onto one learner. Compared with pooled CatBoost, the PR-AUC difference is 0.0186; the 95% paired stratified-bootstrap interval is 0.0074–0.0294 and the empirical p-value is 0.002. The ROC-AUC difference is smaller (0.0049) and its interval includes zero by a narrow margin (−0.00003–0.00977; $p=0.054$). Thus, the strongest evidence concerns rare-positive ranking rather than a large change in global ROC ordering.

At the top 10% review budget, ESGE identifies 72.22%±5.07% of fraud observations with precision 36.30%±4.79%, compared with a 5.03% population prevalence. This concentration is operationally meaningful: approximately one tenth of filings contains nearly three quarters of labeled fraud reports. Thresholded precision and recall remain dependent on the inner validation sample and should not be treated as universal operating points.

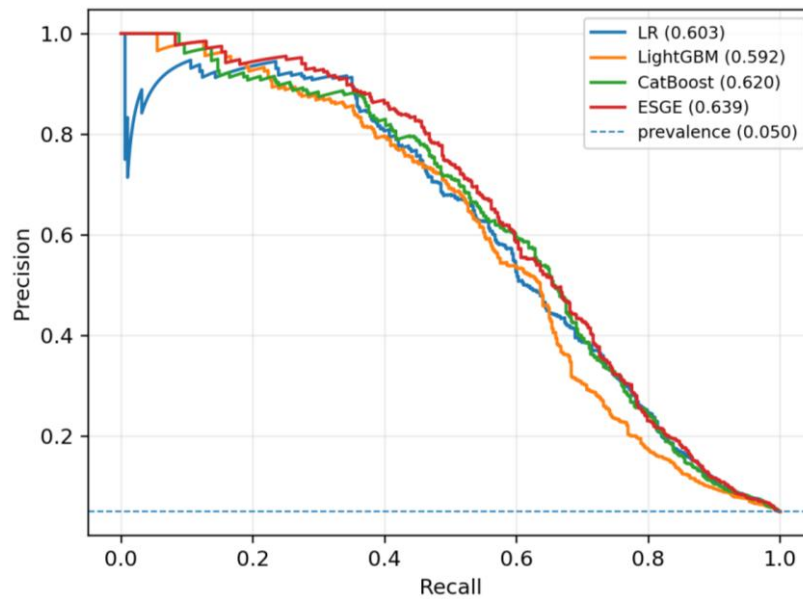


Figure 2. Pooled out-of-fold precision-recall curves. Parentheses show pooled PR-AUC; the dashed line is fraud prevalence.

6.2 Selective Performance

The disagreement gate accepts $79.22\% \pm 1.59\%$ of reports. Conditional on accepted reports, mean PR-AUC rises from 0.6344 to 0.7212, ROC-AUC rises to 0.9191, and F2 rises to 0.7172. Pooled selective PR-AUC is 0.7275. These values do not imply that abstained reports are safe; the opposite workflow is required. The 20.78% high-disagreement set is the portion for which automatic ranking is least trustworthy and should receive manual or alternative-model analysis. Reporting coverage prevents selective performance from being mistaken for full-population performance.

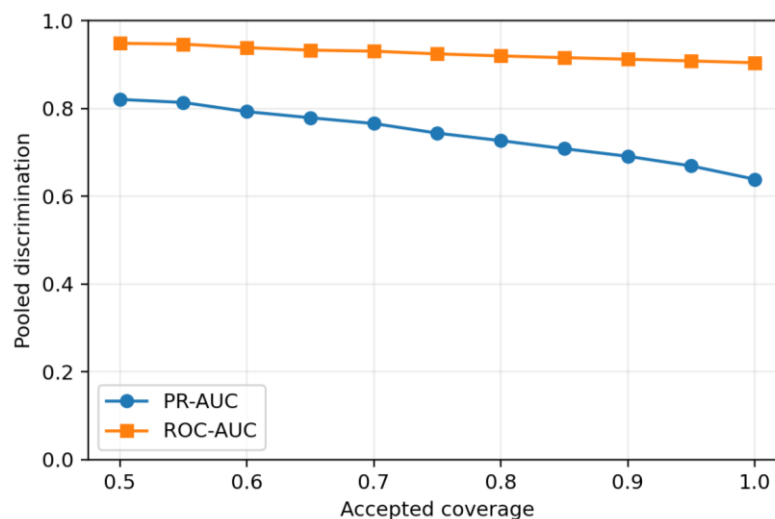


Figure 3. Pooled discrimination as the accepted set expands from low- to high-disagreement reports. Selective gains trade off against coverage.

6.3 Feature and Split Ablations

Table 4. Fixed-budget feature ablation

Features	ROC	PR	F2	R@10%
Raw counts	0.682	0.093	0.259	0.201
Static ratios	0.677	0.088	0.260	0.170
Temporal dynamics	0.904	0.636	0.626	0.728

Raw count features produce only 0.0935 PR-AUC, and adding static densities alone does not improve the result (0.0876). The complete temporal representation reaches 0.6362 PR-AUC under the same 50-tree budget. The large increase indicates that within-company reporting history—especially filing intervals, prior report count, and deviations from prior feature availability—is the dominant contribution of the released variables. It also reinforces the need for strict shift-before-roll implementation: a non-shifted rolling mean would contaminate the feature with the current report.

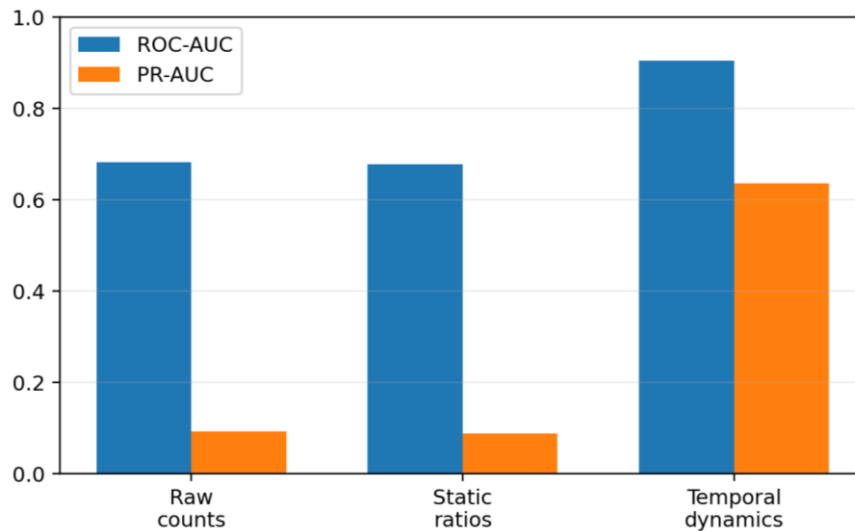


Figure 4. Fixed-budget ablation. Static normalizations do not substitute for prior-only temporal context.

Under the like-for-like fixed budget, random-row validation produces PR-AUC 0.6678, compared with 0.6362 under company isolation, an increase of 0.0315. ROC-AUC increases by 0.0063. The gap is smaller than in some benchmarks because the present predictors exclude direct identifiers, but it is still material for a rare-event ranking task. The result supports company-isolated evaluation whenever future deployment targets unseen issuers.

6.4 Listed Financial-Institution Subgroup

The finance/insurance/real-estate SIC aggregate contains 2,320 observations from the pooled out-of-fold test predictions, including 52 fraud observations (prevalence 2.24%). ESGE obtains ROC-AUC 0.8569 and PR-AUC 0.5008; thresholded precision and recall are 0.3521 and 0.4808. The disagreement gate retains 85.09% coverage and raises conditional PR-AUC to 0.5921. Although this subgroup PR-AUC is lower than the all-sector result, it remains more than twenty times the subgroup prevalence. The lower value is consistent with fewer positive cases and sector-specific reporting structure; a financial-institution deployment should therefore retrain and recalibrate on a larger bank, insurer, and real-estate sample.

Table 5. Selected sector subgroups (pooled oof)

SIC aggregate	N	Fraud	ROC	PR
Manufacturing	3789	259	0.895	0.641
Transportation	786	29	0.841	0.423
Wholesale	274	24	0.925	0.790
Finance	2320	52	0.857	0.501
Services	1770	108	0.931	0.734

6.5 Robustness Interpretation

The main and fixed-budget temporal models produce nearly identical ROC-AUC and PR-AUC, despite different boosting iteration counts. This reduces the likelihood that the result is an artifact of a single training budget. The ensemble’s pooled PR gain over CatBoost is statistically clearer than its

ROC gain, matching the study’s rare-event objective. In addition, the five fold-level ESGE PR-AUC values range from approximately 0.59 to 0.72 rather than being driven by one exceptional test fold.

Brier scores should be interpreted with prevalence in mind. LightGBM has the lowest mean Brier score among individual learners, while ESGE offers better ranking. The weighted score is therefore useful for prioritization but not presented as a fully calibrated posterior probability. A deployment should fit an out-of-time calibration map on delayed labels and publish reliability diagrams by sector and year.

The ablation result is intentionally strong and should not be overgeneralized. It applies to the released feature-count tables. If the full underlying accounting values were available, static ratios such as accrual, leverage, receivables, and margin indicators would likely add independent signal. The current finding is narrower: within this accessible representation, prior-only changes are substantially more informative than contemporaneous feature counts alone.

7. Explainability and Case Analysis

7.1 Stable Global Explanations

The two most stable CatBoost explanation features are elapsed filing interval and number of prior reports; each ranks first or second in all five folds. Calendar index, important-tag count, and year also appear in every fold’s top ten. The trailing four-report mean of important-tag count appears in four of five top-ten sets. The dominance of history variables agrees with the ablation, while the importance of calendar variables warns that filing practices and label availability vary over time. This is a reason to monitor drift, not a reason to remove the variables after viewing the test result. Chen et al. [33] present a strong complementary example in KE-MLLM, where knowledge-enhanced multimodal fusion and parameter-efficient adaptation are paired with structured, human-readable rationales. Their emphasis on explanation faithfulness reinforces our decision to report cross-fold stability and signed local evidence rather than a single visually appealing importance plot.

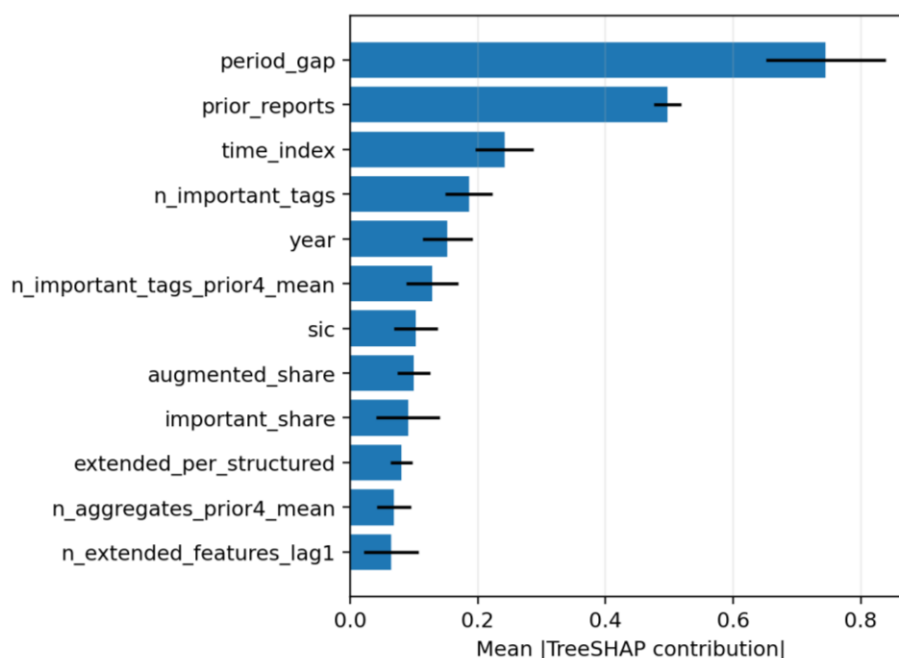


Figure 5. Cross-fold global CatBoost TreeSHAP importance. Error bars show the standard deviation of mean absolute attribution over five outer folds.

Table 6. Cross-fold explanation stability

Feature	Mean SHAP	Rank	Top-10 freq.
period_gap	0.745	1.0	1.0
prior_reports	0.497	2.0	1.0
time_index	0.242	3.2	1.0
n_important_tags	0.186	4.0	1.0
year	0.153	7.0	0.8
n_important_tags_prior4_mean	0.128	7.6	0.8
sic	0.103	11.4	0.6
important_share	0.091	19.2	0.6

7.2 Local Audit Cases

The highest-scored true-positive audit case has anonymized company key d01371391c, period 20170630, and ESGE score 0.9952. Its largest positive CatBoost contributions are prior-report count (1.742) and elapsed filing interval (0.814), followed by aggregate count and the trailing important-tag baseline. This explanation states what increased the model score; it does not identify a fraudulent transaction or prove intent.

The highest-scored false positive has anonymized key 66388d3899, period 20160531, and score 0.9808. It is driven by the same two history variables, plus extended-feature density and an important-tag baseline. The similarity is informative: the current feature space detects unusual filing-history structure, but some legitimate reports share that pattern. Negative contributions from current difference and lagged aggregate counts partially offset the score. A reviewer should therefore inspect financial values, footnotes, governance events, and restatements before escalation.

Local explanation files include eight signed contributions for one high-risk true positive and one high-risk false positive from every fold. Retaining false-positive explanations is important because an explanation system evaluated only on correctly detected cases can conceal systematic over-alerting.

8. Discussion

8.1 Practical Use

ESGE is best viewed as a triage layer placed after a quarterly filing becomes available. A review team can use the continuous score to allocate a fixed review budget, apply a locally chosen threshold, and route high-disagreement cases to specialist review. The system should log data version, feature timestamps, model version, score, gate status, and explanation ranks. It should never automatically accuse an issuer, block a transaction, or substitute for audit evidence. Human reviewers require access to underlying statements and must document independent reasons for any investigation.

The finance subgroup demonstrates transfer to the requested domain but also shows why a single pooled model is insufficient for specialized institutions. Banks and insurers have regulatory capital, reserve, fair-value, and product structures that differ from industrial firms. A production extension should include sector-specific accounting ratios, call-report or statutory data where legally available, governance changes, restatement histories, and narrative representations with exact filing timestamps. The modular ESGE gate can remain unchanged while the representation is expanded.

8.2 Limitations and Threats to Validity

First, AAER-derived labels are delayed, selective, and reflect enforcement decisions rather than all fraud. Unlabeled observations may include undetected misconduct. Second, the released company-isolated tables used here expose feature-family counts but not the complete raw financial-value matrix. The strong temporal signal may partly reflect filing cadence, data availability, repeated reporting patterns, or benchmark construction. Third, outer folds isolate companies but not calendar periods; year and time index are influential. A forward-chaining evaluation is required to establish future-

period transportability. Fourth, subgroup estimates, particularly for financial institutions, are based on only 52 positives. Confidence intervals will be wide even when point discrimination is useful.

Fifth, model disagreement is a pragmatic uncertainty proxy, not a formal guarantee. The accepted set may change under distribution shift, and conditional metrics should not be compared with full-coverage metrics without the coverage value. Sixth, TreeSHAP explains the fitted CatBoost component rather than causal mechanisms, and correlated temporal variables can share attribution. Explanation stability reduces presentation variance but cannot establish causality. Finally, the current thresholds optimize F2 without explicit investigation costs. Institutions should perform decision-curve or cost-sensitive analysis with their own prevalence, staffing, and legal constraints. Fan et al. [34] offer an insightful treatment of strategic adversarial drift through anticipatory bilevel optimization, motif-based memory, and temporal smoothing in dynamic fraud networks. This work highlights a valuable next step for ESGE: production monitoring should account not only for stochastic drift, but also for deliberate adaptation to the deployed detector.

8.3 Reproducibility and Research Integrity

The reproducibility package contains data-file hashes, all source code, exact package versions, fold metrics, pooled predictions, threshold and weight values, leakage decisions, bootstrap samples' summary, ablations, subgroup metrics, figures, TreeSHAP values, and a reference-verification log. The data downloader reports failure rather than silently substituting synthetic observations. This design allows readers to rerun the work, audit every table entry, and replace a component without changing the evaluation boundary.

8.4 Governance Requirements

A responsible implementation requires a documented feature dictionary with an availability timestamp for every field, a prohibition on enforcement-derived inputs, and automated tests that fail when train/test companies overlap. Model cards should distinguish full-coverage and selective metrics, state the abstention policy, and report sector and calendar slices. Alert recipients should see the source filing, the score, the disagreement value, and signed feature contributions—not only a red/green label.

Review outcomes should be fed back through a controlled process. Because analyst decisions can themselves be biased, they should not immediately replace regulatory or restatement labels. A separate adjudication step can classify alerts as data-quality issue, benign reporting change, control weakness, suspected misstatement, or confirmed event. This richer taxonomy would support future multi-task learning without contaminating the legal fraud label.

Finally, access controls and retention rules are necessary even when the source filings are public. Explanation exports can expose company-level monitoring priorities. Logs should be restricted to authorized audit, compliance, and model-risk personnel, with periodic review for disparate error rates across sectors, company sizes, and reporting regimes.

9. Conclusion

This paper presents ESGE, a leakage-aware, explainable, and selectively predictive framework for financial-statement fraud risk screening. On 10,159 public firm-quarter observations with strict unseen-company folds, ESGE achieves mean ROC-AUC 0.9032 and PR-AUC 0.6344, retrieves 72.22% of fraud observations in the top 10% of filings, and improves pooled PR-AUC over the strongest base learner. The disagreement gate increases conditional discrimination while transparently abstaining on roughly one fifth of observations. Ablation and SHAP analyses consistently identify prior-only reporting dynamics as the principal signal in the released feature space. Results for finance, insurance, and real estate remain useful but justify sector-specific retraining. The central methodological conclusion is that credible fraud-screening research must treat split design,

feature availability, abstention coverage, and explanation stability as first-class parts of the model—not as reporting details added after training.

References

- [1] Waffo Dzuyo, G. S., Guibon, G., Cerisara, C., & Belmar-Letelier, L. (2026). Benchmarking generalization in financial statement fraud detection: Robust evaluation and novel tasks. arXiv preprint arXiv:2607.19259.
- [2] U.S. Securities and Exchange Commission. (n.d.). Financial statement data sets. SEC Data Library. Retrieved July 31, 2026.
- [3] USC Leventhal School of Accounting. (n.d.). AAER dataset. Retrieved July 31, 2026.
- [4] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589–609. <https://doi.org/10.1111/j.1540-6261.1968.tb00843.x>
- [5] 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>
- [6] Jones, J. J. (1991). Earnings management during import relief investigations. *Journal of Accounting Research*, 29(2), 193–228. <https://doi.org/10.2307/2491047>
- [7] 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>
- [8] Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, 32(4), 995–1003. <https://doi.org/10.1016/j.eswa.2006.02.016>
- [9] 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>
- [10] Ravisankar, P., Ravi, V., Rao, G. R., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems*, 50(2), 491–500. <https://doi.org/10.1016/j.dss.2010.11.006>
- [11] 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>
- [12] Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review. *Decision Support Systems*, 50(3), 559–569. <https://doi.org/10.1016/j.dss.2010.08.006>
- [13] West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers & Security*, 57, 47–66. <https://doi.org/10.1016/j.cose.2015.09.005>
- [14] Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications*, 68, 90–113. <https://doi.org/10.1016/j.jnca.2016.04.007>
- [15] Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- [16] 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>
- [17] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* 30 (pp. 3146–3154).
- [18] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems* 31 (pp. 6638–6648).
- [19] 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).

- [20] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). <https://doi.org/10.1145/2939672.2939778>
- [21] Arrieta, A. B., 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>
- [22] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 93. <https://doi.org/10.1145/3236009>
- [23] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- [24] 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>
- [25] Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233–240). <https://doi.org/10.1145/1143844.1143874>
- [26] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- [27] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 625–632). <https://doi.org/10.1145/1102351.1102430>
- [28] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)
- [29] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- [30] Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, 77, 103–123. <https://doi.org/10.1007/s10994-009-5119-5>
- [31] Ping, W., Jiao, Y., Fan, H., & Zhang, X. (2026). Multimodal fraud detection in financial statements: A trimodal attention network with contrastive evidence chain construction. *IEEE Access*, 14, 80456–80468. <https://doi.org/10.1109/ACCESS.2026.3695466>
- [32] Chen, Z., Wang, M., Zeng, Z., & Ping, W. (2026). Uncertainty-aware financial forecasting: Leveraging conformal prediction for risk-adjusted models. *IEEE Access*, 14, 90053–90077. <https://doi.org/10.1109/ACCESS.2026.3702666>
- [33] Chen, J., Liu, J., Liang, Y., & Zhou, M. (2026). KE-MLLM: A knowledge-enhanced multi-sensor learning framework for explainable fake review detection. *Applied Sciences*, 16(6), 2909. <https://doi.org/10.3390/app16062909>
- [34] Fan, H., Jiao, Y., Wang, M., Chen, J., & Yang, S. (2026). Fraud learns too: Continual graph learning under strategic adversarial drift in dynamic networks. *Scientific Reports*. <https://doi.org/10.1038/s41598-026-60997-7>