World Journal of Economics and Business Research
Print ISSN: 2960-0073
Online ISSN: 2960-0081
DOI: https://doi.org/10.61784/wjebr4127

EXPLAINABLE MACHINE LEARNING FOR TAX REPORTING ANOMALY DETECTION AND COMPLIANCE RISK ASSESSMENT

Qiao Xu
The University of Delaware, Newark 19716, DE, USA.

Abstract: Tax administrations must rank large populations of filings for audit with scarce, delayed and selectively observed labels, and every selection has to be justified to the taxpayer. This paper presents TRACE, an explainable framework for tax reporting anomaly detection and compliance risk assessment. TRACE combines domain-informed tax risk features (tax burden, inter-period consistency and digit conformity), peer-relative normalisation within each fiscal-year cross-section and a dual-channel detector that fuses balanced-bagging gradient boosting with an ECOD-style tail-evidence channel. The fused risk score decomposes exactly into feature attributions and six auditable reason codes, and conformal p-values turn it into audit selections with a controlled false-alarm rate or false discovery rate. In an out-of-time evaluation on 146,045 real U.S. firm-years with SEC enforcement labels (test years 2003–2008), TRACE attains the highest AUPRC among ten detectors and an AUC of 0.704, within 0.007 of the RUSBoost benchmark and 0.034 and 0.065 above XGBoost and LightGBM. On a semi-synthetic benchmark that injects six tax misreporting schemes into real financial statements, TRACE gives the best AUC with known (0.861) and unseen (0.761) schemes, and its top reason code identifies the manipulated risk group in 72.5% of known-scheme cases, compared with 39.0% for LIME applied to the same score. On real public-sector audit data, its exact attributions agree more closely with the Shapley values of the auditor's own risk formula than KernelSHAP and LIME. Conformal calibration keeps the false-alarm rate on compliant filers within 0.006 of the nominal level across six test years, and Benjamini–Hochberg selection attains a realised FDR of 0.039 at $\alpha = 0.05$ with seven times the power of a LightGBM selector.
Keywords: Tax compliance; Audit selection; Anomaly detection; Explainable AI; Shapley values; Conformal prediction; False discovery rate

1 INTRODUCTION

Risk-based audit selection is the central enforcement instrument of modern tax administrations. Audit capacity covers only a small fraction of filings, so the selection model determines both how much non-compliance is found and how many compliant taxpayers are burdened with an unnecessary examination. Corporate filings make the task harder: tax and financial reporting decisions are taken jointly, and firms that report aggressively in one domain tend to do so in the other [1]. Abnormal tax burdens, inconsistencies with a filer's own history and digit-level irregularities are classic red flags of examiners, but turning them into a reliable and explainable ranking requires learning from data [2-3].

Five properties of the problem shape the design of a learning-based selector. First, labels are scarce, extremely imbalanced and delayed: confirmed cases emerge from enforcement actions years after filing and cover only schemes that were already detected [4-5]. Second, evaders adapt, so a selector must also surface schemes without a labelled precedent. Third, the filing population drifts through inflation, firm growth and regulatory change. Fourth, every selection must be justified to auditors and ultimately to the taxpayer, and post-hoc approximations of a black-box score can be unfaithful to the score that actually triggered the audit [6]. Fifth, false alarms impose real costs on compliant filers, so the error rate of the selection rule should be controlled rather than merely estimated.

Existing work addresses these requirements in isolation. Accounting-fraud models rank firms from financial-statement data; unsupervised detectors flag unusual tax declarations without labels; Shapley-value methods explain tree ensembles; and conformal inference supplies finite-sample guarantees for outlier tests [4-5, 7-12]. No existing framework combines these elements so that the explanation is exact for the deployed score and the audit decision inherits a statistical guarantee.

This paper proposes TRACE (Tax Reporting Anomaly detection with Calibrated Explanations). Its contributions are fourfold. (i) A dual-channel detector that combines domain-informed tax risk features and peer-relative normalisation with a balanced-bagging gradient-boosting channel for known schemes and a parameter-free tail-evidence channel for unseen schemes. (ii) Exact additive reason codes: the fused risk score decomposes exactly into feature attributions and six auditable risk groups, so the reason reported for an audit is the reason the score is high. (iii) Conformal compliance-risk calibration that converts scores into p-values and selects audits with a controlled false-alarm rate on compliant filers or a controlled false discovery rate. (iv) A three-part evaluation: out-of-time detection on 146,045 real U.S. firm-years with enforcement labels, a semi-synthetic benchmark that injects six tax misreporting schemes into real financial statements with field-level ground truth, and real audit data whose label-generating formula is known, which provides exact Shapley ground truth for validating explanations.

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

On 146,045 real U.S. firm-years with SEC enforcement labels, TRACE attains the highest AUPRC of ten detectors and an AUC within 0.007 of the strongest published benchmark. On the semi-synthetic benchmark it gives the best detection of known and unseen schemes, and its reason codes identify the manipulated risk group for 72.5% of the cases, against 39.0% for LIME on the same score. Its conformal audit selections keep the false-alarm and false-discovery rates at their nominal levels.

2 RELATED WORK

2.1 Detecting Tax Evasion and Financial Misreporting

Statistical screening of tax and financial reports has a long tradition. Nigrini used Benford's law to quantify unplanned evasion in individual income tax returns, and Amiram et al. showed that the first-digit divergence of a firm's financial statements predicts misstatements [2-3]. Dechow et al. derived the F-score from accrual, performance and financing ratios of firms subject to SEC Accounting and Auditing Enforcement Releases (AAERs), and Bao et al. showed that an ensemble learner on 28 raw accounting items outperforms ratio-based logistic models, introducing NDCG@k as a ranking metric for fraud prediction [4-5]. Frank et al. documented a strong positive association between aggressive tax reporting and aggressive financial reporting, which motivates financial-statement evidence for tax compliance risk [1]. Within tax administrations, de Roux et al. detected under-reporting declarations with unsupervised learning, Vanhoeyveld et al. scaled anomaly detection to value-added tax fraud, and Savić et al. combined unsupervised outlier detectors for tax evasion risk management, while Hooda et al. classified suspicious firms from external audit data [7-9, 13]. These studies optimise detection accuracy; the faithfulness of the explanations and the error rate of the resulting audit decisions are not addressed.

2.2 Anomaly Detection on Tabular Data

Isolation forest isolates anomalies with random axis-parallel partitions, and deep isolation forest extends it with randomly initialised neural representations [14-15]. COPOD and ECOD estimate tail probabilities from empirical distribution functions; they are parameter-free and interpretable per dimension [16-17]. XGBOD augments a supervised booster with unsupervised outlier scores, RUSBoost couples random undersampling with boosting for imbalanced data, and gradient-boosting systems such as XGBoost and LightGBM remain the strongest learners on tabular data [18-21]. TRACE follows the semi-supervised direction but keeps the unsupervised evidence additive and grouped, so that it can be explained exactly.

2.3 Explainability and Calibrated Decisions

SHAP attributes predictions with Shapley values; TreeSHAP computes them exactly in polynomial time for tree ensembles, whereas LIME and KernelSHAP approximate a model locally by sampling [10-11, 22]. Rudin argued that high-stakes decisions require explanations that are faithful to the deployed model [6]. Conformal prediction provides distribution-free validity under exchangeability, and Bates et al. proved that conformal p-values for outlier detection are positively dependent, so the Benjamini–Hochberg procedure controls the false discovery rate [12, 23-25]. TRACE brings these results to tax audit selection and couples them with explanations that are exact for the score being calibrated.

3 METHODOLOGY

3.1 Problem Formulation

Let x_{it} denote the reported line items of filer i in fiscal year t and $x_{i,t-1}$ its previous filing. A label $y_{it} \in \{0, 1\}$ marks a non-compliant (misreported) filing; labels exist only for earlier years, are extremely imbalanced and cover only detected schemes. For the filings of a target year, TRACE returns a risk score $S(x)$, reason codes $\rho_g(x)$ for a set $\mathcal{G}$ of auditable risk groups, and an audit set $\mathcal{A}_t$ whose error is controlled either as a false-alarm rate (FAR) on compliant filers, $P(i \in \mathcal{A}_t \mid y_{it} = 0) \le \alpha$, or as a false discovery rate (FDR), the expected share of compliant filings among the selected audits. Figure 1 summarises the pipeline.

3.2 Domain-Informed Tax Risk Features

TRACE starts from the 28 raw items and 14 financial ratios used by Dechow et al. and Bao et al. and adds three families of tax-specific evidence [4-5]. *Tax burden:* with pre-tax income π = ib + txt, the effective tax rate ETR = txt/π (defined for $\pi > 0$), a pre-tax loss flag, tax expense to sales, income taxes payable to assets, payable to expense, and the year-on-year changes of ETR and payable intensity capture understated tax expense and concealed tax liabilities. *Inter-period consistency:* sales and asset growth, gross margin and its change, return on assets, leverage and its change, receivables to sales and its change, and the implied interest rate xint/(dlc + dltt) compare a filing with the filer's own history, as an examiner compares consecutive returns; an abnormal implied rate indicates interest deductions that are

not backed by debt. *Digit conformity:* for the set of non-zero reported amounts of a filing, the first-digit mean absolute deviation from Benford's law is

$$\mathrm{MAD}_{it} = (1/9)\,\Sigma_{d=1..9}\,|\,\hat{p}_{it}(d) - \log_{10}(1 + 1/d)\,| \tag{1}$$

where $\hat{p}_{it}(d)$ is the share of amounts with leading digit d [2-3]. The round-figure share is the fraction of material amounts that carry at most two significant digits, a signature of invented numbers. The resulting $m = 61$ features are partitioned into six risk groups that correspond to audit programmes: Tax burden (TAX), Revenue & cost (REV), Accruals & earnings (ACC), Financing & interest (FIN), Digit conformity (DIG) and Scale & market (SCL).

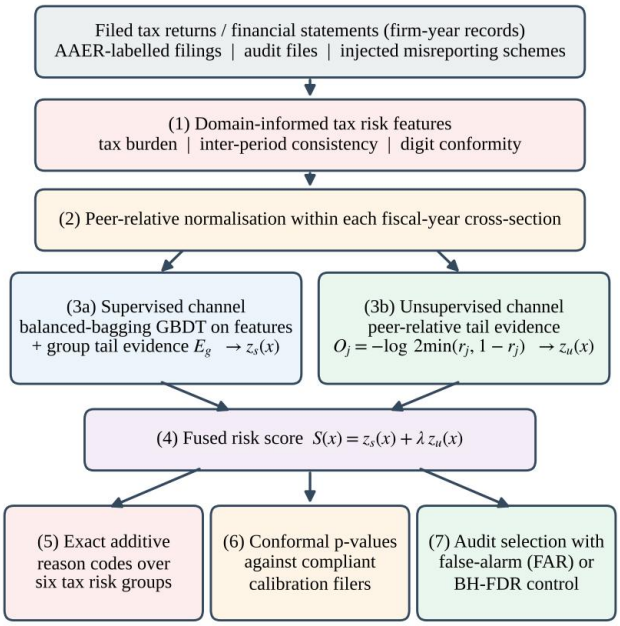


Figure 1 Overview of TRACE

Domain-informed features are normalised within each fiscal-year cross-section and scored by a supervised and an unsupervised channel; the fused score is decomposed exactly into reason codes and converted into conformal p-values for audit selection with error control.

3.3 Peer-Relative Normalisation and Tail Evidence

Monetary amounts are heavy-tailed and their scale drifts with inflation and firm growth. TRACE therefore maps every feature to its mid-rank percentile within the cross-section of filings of the same fiscal year,

$$r_j(x_{it}) = [\#\{k: x_{ktj} < x_{itj}\} + \tfrac{1}{2}\,\#\{k: x_{ktj} = x_{itj}\}] \,/\, n_t \tag{2}$$

which is invariant to monotone rescaling and identically distributed across years; missing values remain missing. Two-sided tail evidence follows the empirical-distribution principle of ECOD [17]:

$$O_j(x) = -\log(2\,\min\{r_j(x), 1 - r_j(x)\}) \tag{3}$$

For a typical filer r_j is uniform, hence O_j follows a unit exponential distribution with mean one; missing entries receive this null value. The group evidence $E_g(x) = |g|^{-1}\,\Sigma_{j\in g}\,O_j(x)$ measures how extreme a filing is within risk group g.

3.4 Dual-Channel Detection

Supervised channel. Because positives are rare, TRACE trains B gradient-boosted tree ensembles $f_1, \ldots, f_B$, each on all positive filings and a fresh random sample of $r{\cdot}n_+$ negatives, with the normalised features and the six group-evidence features as inputs [21]. The channel output is the average log-odds $s_s(x) = B^{-1}\,\Sigma_b\,f_b(x)$. Balanced bagging stabilises the ranking under extreme imbalance, and the group evidence lets the trees exploit joint extremeness within a risk group. *Unsupervised channel.* The mean tail evidence $s_u(x) = m^{-1}\,\Sigma_j\,O_j(x)$ rises for any filing that is extreme on many dimensions, including schemes without a labelled precedent. *Fusion.* With z-scores z_s and z_u computed from the training means and standard deviations σ_s and σ_u, the risk score is

$$S(x) = z_s(x) + \lambda\, z_u(x) \tag{4}$$

where λ is selected on validation years and fixed to 0.2 in all experiments.

3.5 Exact Additive Reason Codes

TreeSHAP yields exact Shapley attributions of each ensemble, $f_b(x) = \varphi_0^{(b)} + \Sigma_j\,\varphi_j^{(b)}(x)$ [11]. TRACE defines the attribution of feature j to the fused score as

$$\varphi_j(x) = \Sigma_b\,\varphi_j^{(b)}(x) \,/\, (B\sigma_s) + \lambda(O_j(x) - 1) \,/\, (m\sigma_u) \tag{5}$$

and the attribution φ_{Eg} of each group-evidence feature by the first term alone. Proposition 1 (completeness). For every filing, $S(x) = \varphi_0 + \Sigma_j \varphi_j(x) + \Sigma_g \varphi_{Eg}(x)$, where φ_0 does not depend on x. *Proof:* averaging the per-ensemble identities gives $s_s(x) = B^{-1}\Sigma_b \varphi_0^{(b)} + B^{-1}\Sigma_b\Sigma_j \varphi_j^{(b)}(x)$, and $s_u(x) = 1 + m^{-1}\Sigma_j(O_j(x) - 1)$; substituting both into (4) and collecting the terms that do not depend on x into φ_0 proves the claim. The reason code of group g is

$$\rho_g(x) = \Sigma_{j \in g} \varphi_j(x) + \varphi_{Eg}(x) \quad (6)$$

The codes sum exactly to $S(x) - \varphi_0$, are signed, so that exculpatory evidence is shown as well, and require no sampling. Unlike LIME or KernelSHAP applied to the same score, they are not approximations: the reason reported to the auditor is the reason the score is high. The top-ranked group is the primary reason, and the leading features inside it form the evidence trail.

3.6 Conformal Compliance-Risk Calibration

Let $\mathcal{D}_{cal} = \{x_1, \ldots, x_n\}$ be compliant filings that were not used for training, typically from the most recent labelled year. The conformal p-value of a new filing is [12, 23]

$$p(x) = [1 + \#\{i \in \mathcal{D}_{cal}: S(x_i) \geq S(x)\}] / (n + 1) \quad (7)$$

If calibration and test compliant filings are exchangeable, $P(p(x) \leq \alpha \mid y = 0) \leq \alpha$. TRACE offers two decision rules. *FAR control* audits every filing with $p(x) \leq \alpha$, which bounds the probability that a compliant taxpayer is audited by α irrespective of the strength of the detector. *FDR control* applies the Benjamini–Hochberg procedure to the p-values of a batch of filings; since conformal p-values are positively dependent, the expected share of compliant filings among the selected audits is at most α [12, 25]. Peer-relative normalisation supports these guarantees over time: every feature is uniform within each year, which removes level drift from the score distribution of compliant filers and reduces systematic departures from the exchangeability that conformal validity requires (Section 5.6).

4 EXPERIMENTAL SETUP

4.1 Datasets

Table 1 summarises the three evaluation settings, which together cover real enforcement labels, controlled misreporting schemes with ground truth and a real audit process with a known scoring rule.

Table 1 Evaluation Data

Data	Records	Positives	Feat.	Protocol
Real misstatements [5]	146,045 firm-years, 1990–2014	964 (AAER)	61	Out-of-time, test 2003–2008
Semi-synthetic tax schemes	45,911 train / 34,696 test	3% injected, S1–S6	47	Known / novel schemes
UCI Audit Risk [13]	763 firms	304 high-risk	8	5-fold CV, exact Shapley truth

Real misstatement data. We use the public dataset of Bao et al.: 146,045 firm-years of U.S. listed companies from 1990 to 2014 with 28 raw Compustat items, the 14 ratios of Dechow et al. and labels from SEC AAERs (964 misstated firm-years) [4-5]. Following Bao et al., test year $T \in \{2003, \ldots, 2008\}$ uses training years 1991 to $T - 2$, a 24-month gap that mimics the delay of enforcement outcomes, and serial misstatements that span the training and test periods are recoded as compliant in training [5]. Each test year contains 5,612–5,981 firm-years with 26–69 misstated filings. All hyperparameters were selected on the validation years 2000–2002 and fixed before testing.

Semi-synthetic tax misreporting benchmark. To evaluate unseen schemes and explanation fidelity against ground truth, six schemes are injected into real statements of firm-years that never appear in an AAER: S1 revenue under-reporting (sales and receivables reduced by 15–40%), S2 cost or deduction inflation (cost of goods sold raised by 10–30%, half of it booked as payables), S3 tax-expense understatement (40–80% of the income tax expense removed), S4 tax-payable concealment (60–100% of taxes payable removed with the matching current liabilities), S5 interest-deduction stuffing (interest expense raised by 50–150% without additional debt) and S6 fabricated round figures (8–14 amounts rounded to two significant digits). Schemes that shift taxable income adjust income and tax expense consistently with a 35% statutory rate. Each scheme is injected into 0.5% of the filings of every year. Training uses 1995–2001 and testing 2003–2008; detection is evaluated on the 24,143 test filings outside a 30% calibration split (1,038 injected). The primary risk group is REV for S1–S2, TAX for S3–S4, FIN for S5 and DIG for S6. In the known-scheme setting all schemes appear in training (1,380 positives); in the novel-scheme setting S5 and S6 are absent (920 positives). The 14 ratios are excluded here because they cannot be recomputed for injected statements, which leaves 47 features.

Real public-sector audit data. The UCI Audit Risk data of Hooda et al. contain the risk evidence of firms examined by an external government audit office in India (776 records, 763 unique) [13]. We use the eight raw evidence fields: sector score, location, the expenditure discrepancies PARA_A and PARA_B, the historical discrepancy score, money at risk, district loss and loss history. The label is a deterministic function of the auditor's component scores, Risk = $1\{AR > 1\}$ with AR = $0.5 \cdot IR \cdot CR$, where the inherent risk IR aggregates six evidence components and the control risk CR the district-loss and history components; sector score and location do not enter the formula. We compute the exact interventional Shapley values of the reconstructed function $0.5 \cdot (\Sigma_k c_k)(c_E + c_F)$ of the six published components (all 2^6 coalitions, marginal background) and use them as ground truth for local explanations. Evaluation uses stratified five-fold cross-validation.

4.2 Baselines and Explainers

Detection baselines are the logistic F-score model on the 14 ratios; RUSBoost on the 28 raw items, the benchmark of Bao et al. (1,000 trees, minimum leaf size 5, learning rate 0.1); XGBoost and LightGBM on the 42 raw items and ratios with class reweighting; XGBOD with isolation-forest, ECOD, COPOD and histogram-based outlier scores as additional inputs; and the unsupervised detectors IForest, ECOD, COPOD and DIF on the peer-normalised features, where DIF uses 20 randomly initialised two-layer representations, each isolated by an iForest of 50 trees [4-5, 14-21]. The exact TRACE codes are compared with LIME and KernelSHAP applied to the same fused score, with TreeSHAP of a stand-alone LightGBM, random forest and isolation forest, with linear SHAP of logistic regression and with ECOD dimension scores [10-11, 22].

4.3 Metrics

Detection is measured by AUC, AUPRC and the top-1% metrics of Bao et al.: NDCG@1%, sensitivity and precision among the 1% highest-scored filings of a test year [5]. With 26–69 misstated filings per year, the top 1% (about 59 firm-years) contains between zero and four of them, so these metrics vary strongly across years, whereas AUC and AUPRC use the full ranking. The semi-synthetic benchmark adds recall at a 5% audit budget. Reason-code fidelity is the share of injected cases whose top group (Hit@1) or top two groups (Hit@2) contain the true group; agreement with exact Shapley values is measured by the per-firm Spearman correlation and the top-1 agreement. Audit selection is evaluated by the empirical FAR, the realised FDR and the power.

4.4 Implementation

The supervised channel uses $B = 10$ LightGBM ensembles (300 trees, 15 leaves, learning rate 0.05, feature fraction 0.6), each trained on all positives and $r = 10$ negatives per positive, and $\lambda = 0.2$. All experiments run on a single CPU core; training TRACE on 75,000 firm-years and scoring a test year takes about 20 s.

5 RESULTS

5.1 Detection on Real Misstatement Data

Table 2 and Figure 2 report the out-of-time results. TRACE attains the highest AUPRC (0.0186) and an AUC of 0.704, second to the RUSBoost benchmark of Bao et al. (0.711); the gap of 0.007 is small relative to the year-to-year standard deviation of the AUC (0.03–0.04 for all supervised detectors). TRACE improves the AUC by 0.034 over XGBoost, 0.065 over LightGBM, 0.019 over XGBOD and 0.013 over the logistic F-score model trained on the same history, and it has the best AUC of all detectors in 2003 (0.741) and 2004 (0.766). All unsupervised detectors remain below 0.64: misstatements that lead to enforcement are not simply statistical outliers, and labelled history must drive the ranking. The top-1% metrics favour different detectors in different years (Figure 2, bottom) because only zero to four misstated filings fall into the top 1% of a test year; XGBOD and the logistic model reach the highest averages, and TRACE lies in the middle of the supervised group.

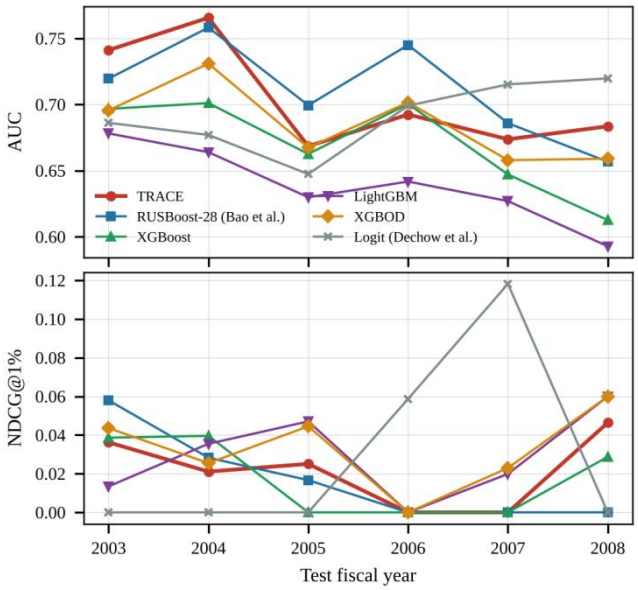


Figure 2 Out-of-Time AUC (Top) and NDCG@1% (Bottom) per Test Year on the Real Misstatement Data

Table 2 Out-of-Time Detection on Real Misstatement Data (Mean over 2003–2008)

Method	AUC	AUPRC	NDCG@1%	Sens@1%	Prec@1%
Logit (F-score)	0.691	0.0179	0.0295	**3.7%**	2.0%
RUSBoost-28	**0.711**	0.0172	0.0171	1.9%	2.0%
XGBoost	0.670	0.0158	0.0178	1.9%	1.7%
LightGBM	0.639	0.0172	0.0293	2.8%	2.0%
XGBOD	0.685	0.0180	**0.0327**	3.6%	**2.8%**
IForest	0.562	0.0140	0.0265	2.4%	1.7%
ECOD	0.532	0.0084	0.0020	0.2%	0.3%
COPOD	0.632	0.0116	0.0020	0.2%	0.3%
DIF	0.572	0.0130	0.0292	3.0%	2.3%
TRACE (ours)	0.704	**0.0186**	0.0215	2.0%	1.7%

5.2 Reason Codes on Real Filings

Figure 3(a) aggregates the absolute reason codes over all test firm-years. Accruals & earnings (28.4%), scale & market (22.0%) and financing & interest (21.5%) carry most of the attribution, followed by revenue & cost (14.5%), tax burden (10.2%) and digit conformity (3.3%). The leading inputs in Figure 3(b) are soft assets, stock issuance, share price, retained earnings to assets and depreciation, in line with the accrual and financing determinants of misstatement reported by Dechow et al.; the pre-tax loss flag is the most influential tax-burden input [4]. Figure 3(c) shows the reason codes of a misstated FY2008 filing ranked in the top 0.07% of its year: scale & market (+1.19), revenue & cost (+0.94), accruals & earnings (+0.80) and financing & interest (+0.55) dominate, the strongest single inputs are soft assets and stock issuance, and digit conformity is slightly exculpatory (−0.03). Such a profile directs the examination towards revenue recognition, working-capital accruals and the financing transactions of the year.
The deletion test in Figure 4(a) neutralises the k most attributed inputs of each top-1% filing by setting them to the median rank and records the drop of S. Ordering by the exact TRACE attributions produces the steepest drop: removing five inputs lowers S by 1.99 on average, against 1.37 for a global importance ordering, 0.87 for tail evidence alone and 0.36 for a random ordering, and the areas under the deletion curves are 18.5, 12.3, 8.5 and 3.5. Supervised-only TreeSHAP comes close (18.3) because the supervised channel dominates the fused score at $\lambda = 0.2$; the exact fused attributions remain the best ordering because they also account for the tail-evidence term.

5.3 Ablation and Fusion Weight

Table 3 isolates the components on the real data. Balanced bagging is the largest single contributor (AUC 0.640 without it), followed by the inter-period consistency features (−0.012) and the tax-burden features (−0.004); restricting the learner to the 28 raw items lowers the AUC to 0.698. The digit features hardly change the AUC (−0.001), as AAER cases rarely involve fabricated round figures, whereas they dominate the detection of scheme S6 in Section 5.4. Removing the tail-evidence augmentation, the peer-relative normalisation or the unsupervised fusion raises the AUC by 0.003–0.004: on labels that reflect previously detected schemes the unsupervised components cost a small amount of ranking accuracy, and in return they provide the sensitivity to unseen schemes and the calibration stability reported below. Figure 4(b) shows that λ governs this trade-off smoothly; the AUC decreases from 0.708 at $\lambda = 0$ to 0.704 at $\lambda = 0.2$ and 0.657 at $\lambda = 1$, while the detection of the unseen scheme S6 improves with λ (Table 4).

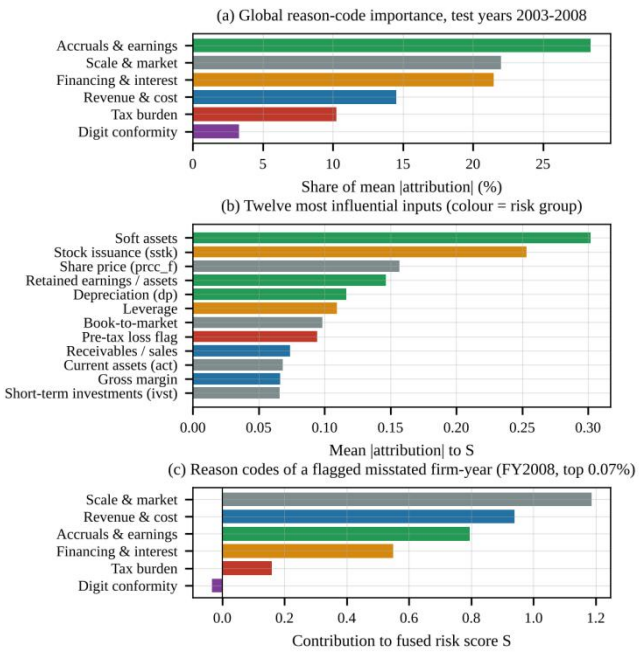


Figure 3 Reason Codes on Real Filings: (a) Share of the Mean Absolute Attribution per Risk Group; (b) The Twelve Most Influential Inputs; (c) Reason Codes of a Misstated FY2008 Filing Ranked in the Top 0.07% of Its Year

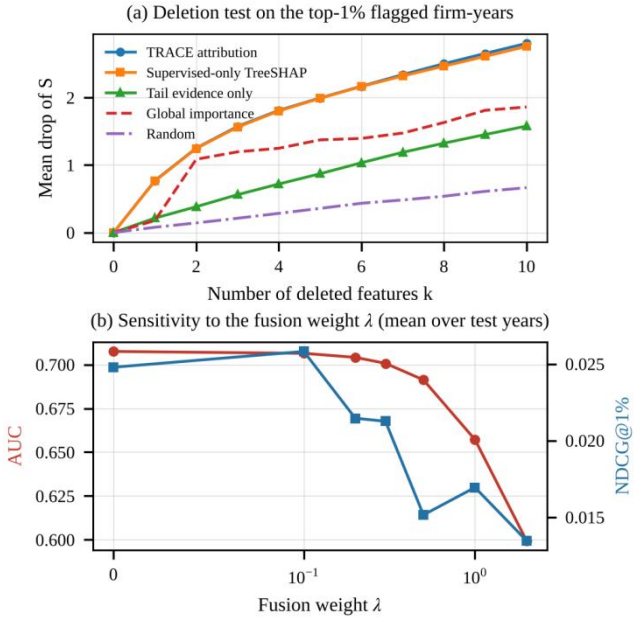


Figure 4 (a) Deletion Test: Mean Drop of the Fused Score after Neutralising the k Most Attributed Inputs of Each Top-1% Filing; (b) AUC and NDCG@1% as Functions of the Fusion Weight λ (Mean over Test Years)

Table 3 Ablation of TRACE on Real Misstatement Data

Variant	AUC	AUPRC	NDCG@1%	Prec@1%
TRACE (full)	0.7041	0.0186	0.0215	1.7%
w/o tax-burden features	0.6997	0.0199	0.0306	2.3%
w/o inter-period features	0.6921	0.0176	0.0230	1.7%
w/o digit features	0.7034	0.0189	0.0216	1.4%
w/o peer-relative normalisation	0.7079	0.0193	0.0292	2.5%
w/o tail-evidence augmentation	0.7082	0.0194	0.0215	1.4%
w/o unsupervised fusion ($\lambda = 0$)	0.7075	0.0189	0.0248	2.0%
w/o balanced bagging	0.6397	0.0148	0.0175	1.7%
28 raw items only	0.6979	0.0196	0.0313	2.5%

5.4 Known and Novel Tax Misreporting Schemes

Table 4 and Figure 5(a) summarise the semi-synthetic benchmark. With all schemes labelled, TRACE achieves the best AUC (0.861), AUPRC (0.477) and recall at a 5% audit budget (0.482), and the digit features make S6 almost perfectly detectable (AUC 0.996). In the novel-scheme setting, where S5 and S6 never appear in training, TRACE keeps the highest overall AUC (0.761) and AUPRC (0.314), while XGBoost has a slightly higher overall recall@5% (0.351 against 0.331). The supervised baselines rank the unseen round-figure scheme S6 below compliant filings (AUC 0.352 for XGBoost and 0.401 for LightGBM) because nothing in their training history links digit anomalies to risk; the tail-evidence channel lifts TRACE to 0.518 at $\lambda = 0.2$ and 0.564 at $\lambda = 1$, and its recall@5% on S6 rises from 0.006 for the baselines to 0.035 and 0.087, respectively. Purely unsupervised detectors reach 0.59–0.61 on S6 but stay below 0.60 overall, so neither channel alone covers both regimes. On the unseen interest-stuffing scheme S5 all supervised detectors are close to 0.58, with TRACE marginally ahead (0.584). The tax-specific features pay off on the tax schemes: for tax-payable concealment (S4) TRACE reaches an AUC of 0.887 in the novel setting, against 0.874 for LightGBM and 0.864 for XGBoost.

Table 4 Detection on the Semi-Synthetic Benchmark (K = Known, N = Novel Schemes)

Method	AUC K	R@5% K	AUC N	AUPRC N	R@5% N	AUC S5 N	AUC S6 N
IForest	0.593	0.104	0.596	0.065	0.107	0.568	0.590
ECOD	0.537	0.083	0.537	0.051	0.083	0.535	0.603
DIF	0.585	0.101	0.582	0.059	0.094	0.576	**0.605**
XGBoost	0.856	0.479	0.734	0.302	**0.351**	0.578	0.352
LightGBM	0.853	0.472	0.738	0.297	0.316	0.576	0.401
PR-tail channel	0.534	0.080	0.534	0.050	0.080	0.526	0.596
TRACE ($\lambda = 0.2$)	**0.861**	**0.482**	**0.761**	**0.314**	0.331	**0.584**	0.518
TRACE ($\lambda = 1$)	0.822	0.396	0.739	0.184	0.278	0.582	0.564

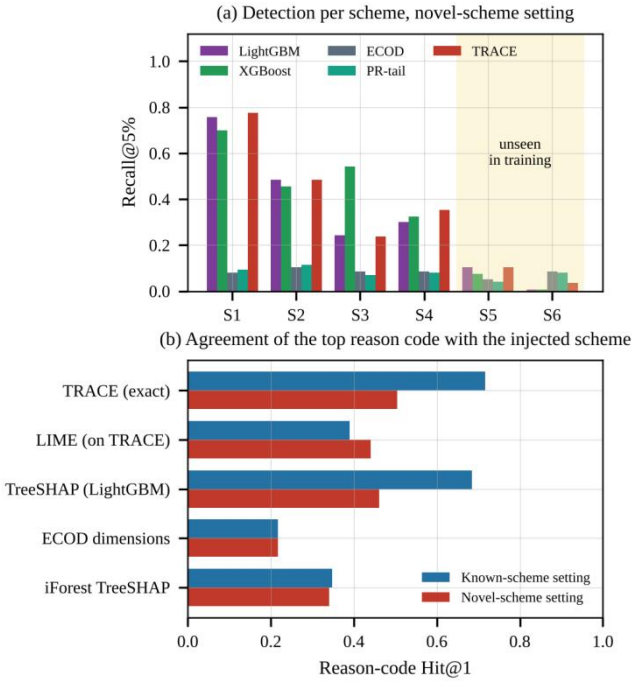


Figure 5 Semi-Synthetic Benchmark: (a) Recall at a 5% Audit Budget per Scheme in the Novel-Scheme Setting (S5 and S6 Unseen in Training); (b) Hit@1 of the Top Reason Code in the Known- and Novel-Scheme Settings

5.5 Fidelity of Reason Codes

Table 5 and Figure 5(b) measure whether the top reason code names the manipulated risk group. The exact TRACE codes are the most faithful explainer in both settings (Hit@1 of 0.716 and 0.505). On the identical 200-case subset, LIME applied to the same fused score reaches 0.390 in the known setting against 0.725 for the exact codes, so the difference stems from approximation error rather than from the model. Per scheme, the exact codes name the correct group for 99% of the S6 cases and 77% of the S4 cases, and the average reason-code profile of every scheme peaks at its true group: revenue & cost for S1 and S2, tax burden for S3 and S4, financing & interest for S5 and digit conformity for S6. Unseen schemes are harder to explain: for S5 and S6 in the novel setting the supervised term dominates the fused codes and Hit@1 falls to 0.05–0.07, while the ECOD dimension scores still point to the digit group for 66% of the S6 cases. Reporting the tail-evidence term of each reason code next to the fused code therefore shows the examiner when a filing is flagged by extremeness without a labelled precedent.

Table 5 Reason-Code Fidelity on Injected Cases (Matched: Common 200-Case Subset)

Explainer	Hit@1 K	Hit@2 K	Hit@1 N	Hit@2 N	Match. K	Match. N
TRACE (exact)	**0.716**	0.833	**0.505**	**0.615**	**0.725**	**0.445**
TreeSHAP (LightGBM)	0.685	**0.835**	0.461	0.591	0.670	0.390
LIME (on TRACE)	0.390	0.615	0.440	0.555	0.390	0.440
iForest TreeSHAP	0.348	0.545	0.340	0.588	0.315	0.320
ECOD dimensions	0.218	0.400	0.218	0.400	0.215	0.215

5.6 Conformal Audit Selection

Figure 6(a) evaluates FAR control on the real data: for test year T the model is trained on 1991 to $T-3$ and calibrated on the compliant filings of year $T-2$. Averaged over the six test years, the FAR of TRACE exceeds the nominal level by at most 0.006 (0.1045 at $\alpha = 0.10$), whereas the variant without peer-relative normalisation and LightGBM on raw amounts exceed it by up to 0.020 and 0.015. The year-to-year deviation is similar for all three (mean absolute deviation 0.012–0.014), which reflects the size of a single calibration year. At $\alpha = 0.10$, TRACE audits 10.6% of the population and recovers 23.3% of the misstated filings, against 21.8% for LightGBM.

For batch selection with FDR control (Table 6 and Figure 6(b)), the compliant filings of the semi-synthetic test period are split at random into calibration (30%) and test (70%) parts, 20 times. The realised FDR is at or below α in eleven of the twelve configurations; the remaining one (TRACE, novel setting, $\alpha = 0.10$) equals 0.110, within one standard error of the nominal level. At $\alpha = 0.05$ in the known setting, TRACE recovers 13.5% of the injected cases with a realised FDR of 0.039, seven times the power of LightGBM (1.9%), because its fused score separates the extreme tail of the score distribution more sharply. The unsupervised channel alone makes no discoveries at these levels: FDR control requires a strong ranking, not only an anomaly score.

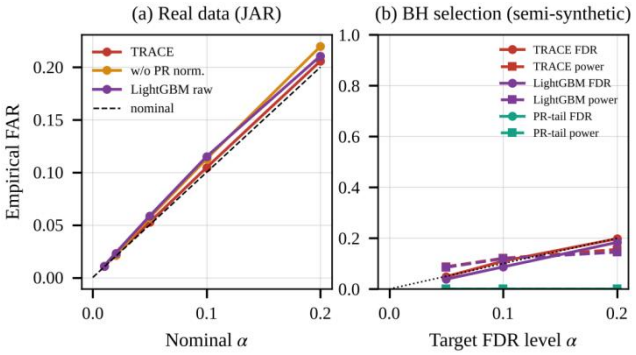


Figure 6 Conformal Audit Selection: (a) Empirical False-Alarm Rate versus Nominal Level on the Real Data (Mean over 2003–2008); (b) Realised FDR and Power of BH Selection in the Novel-Scheme Setting (Mean over 20 Calibration Splits)

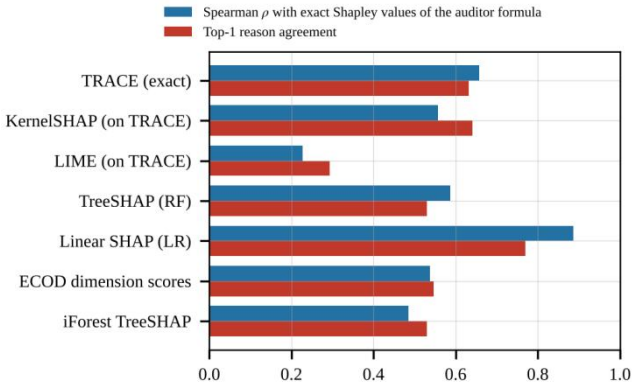


Figure 7 Agreement of Local Explanations with the Exact Shapley Values of the Auditor's Risk Formula for the 304 High-Risk Firms of the Audit Data

Table 6 BH Audit Selection with Conformal p-Values (Mean over 20 Splits)

Setting, α	TRACE FDR	TRACE power	LightGBM FDR	LightGBM power
Known, 0.05	0.039	**0.135**	0.024	0.019
Known, 0.10	0.097	**0.216**	0.077	0.122
Known, 0.20	0.194	**0.289**	0.189	0.277
Novel, 0.05	0.049	**0.088**	0.039	0.085
Novel, 0.10	0.110	**0.121**	0.087	0.120
Novel, 0.20	0.198	**0.156**	0.183	0.145

5.7 Validation Against an Auditor's Risk Formula

On the audit data all supervised detectors are almost perfect (AUC 0.999 ± 0.001 for TRACE, 0.999 for random forest, 0.998 for XGBoost and 0.983 for logistic regression), whereas isolation forest (0.858) and ECOD (0.822) are clearly weaker. Figure 7 compares the explanations of the 304 high-risk firms with the exact Shapley values of the auditor formula. The exact TRACE attributions reach a mean Spearman correlation of 0.658, above KernelSHAP (0.557) and LIME (0.227) applied to the same score and above TreeSHAP of a random forest (0.587), the ECOD dimension scores (0.537) and isolation-forest TreeSHAP (0.485); in top-1 agreement TRACE (63.2%) and KernelSHAP (64.1%) are on par. TRACE places 6.3% of its attribution mass on the two fields that do not enter the formula, against 22.8% for ECOD and 18.1% for isolation-forest TreeSHAP. The linear SHAP values of logistic regression agree most closely (0.886) because every component of the formula is piecewise linear in its evidence field; this agreement comes with the lowest supervised AUC (0.983). Among the high-accuracy detectors, the exact codes of TRACE are the most faithful to the auditor's own reasoning.

6 DISCUSSION

The results position TRACE as an audit-selection instrument rather than a pure ranking model. On historical enforcement labels its accuracy is within 0.007 AUC of the strongest published benchmark and above standard gradient boosting, and every score comes with an exact decomposition into risk groups that maps onto audit programmes: a dominant tax-burden code points to the tax provision and the reconciliation of taxes payable, a revenue & cost code to revenue recognition and deductions, a financing code to interest deductions and thin-capitalisation rules, and a digit code to the documentation of individual amounts. The fusion weight λ acts as a policy parameter: small values

maximise accuracy on previously detected schemes, larger values allocate part of the audit budget to filings whose extremeness has no labelled precedent. Conformal calibration turns audit capacity into an explicit error budget for compliant taxpayers, and peer-relative normalisation keeps this budget stable as the population drifts. The same construction applies to individual income tax returns, value-added tax declarations and payroll filings, with risk groups defined over the corresponding line items.

7 CONCLUSION

This paper presented TRACE, an explainable framework for tax reporting anomaly detection and compliance risk assessment that integrates domain-informed tax risk features, peer-relative normalisation, a dual-channel detector, exact additive reason codes and conformal audit selection. Experiments on real enforcement data, a semi-synthetic tax misreporting benchmark and real audit data show competitive detection with the highest AUPRC on enforcement labels, the best detection of known and unseen injected schemes, reason codes that are markedly more faithful than sampling-based explanations, and audit selections whose false-alarm and false-discovery rates stay at their nominal levels. Future work will extend the framework to transaction-level and network data of tax administrations and will learn the fusion weight from the audit capacity and the expected cost of missed schemes.

COMPETING INTERESTS

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

REFERENCES

[1] Frank M M, Lynch L J, Rego S O. Tax reporting aggressiveness and its relation to aggressive financial reporting. The Accounting Review, 2009, 84(2): 467-496.
[2] Nigrini M J. A taxpayer compliance application of Benford's law. Journal of the American Taxation Association, 1996, 18(1): 72-91.
[3] 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.
[4] Dechow P M, Ge W, Larson C R, et al. Predicting material accounting misstatements. Contemporary Accounting Research, 2011, 28(1): 17-82.
[5] 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.
[6] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 2019, 1(5): 206-215.
[7] de Roux D, Perez B, Moreno A, et al. Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018: 215-222.
[8] Vanhoeyveld J, Martens D, Peeters B. Value-added tax fraud detection with scalable anomaly detection techniques. Applied Soft Computing, 2020, 86: 105895.
[9] Savić M, Atanasijević J, Jakovetić D, et al. Tax evasion risk management using a hybrid unsupervised outlier detection method. Expert Systems with Applications, 2022, 193: 116409.
[10] Lundberg S M, Lee S I. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NIPS), 2017: 4765-4774.
[11] 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.
[12] Bates S, Candès E, Lei L, et al. Testing for outliers with conformal p-values. The Annals of Statistics, 2023, 51(1): 149-178.
[13] Hooda N, Bawa S, Rana P S. Fraudulent firm classification: A case study of an external audit. Applied Artificial Intelligence, 2018, 32(1): 48-64.
[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] Xu H, Pang G, Wang Y, et al. Deep isolation forest for anomaly detection. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(12): 12591-12604.
[16] Li Z, Zhao Y, Botta N, et al. COPOD: Copula-based outlier detection. In: Proceedings of the IEEE International Conference on Data Mining (ICDM), 2020: 1118-1123.
[17] 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.
[18] Zhao Y, Hryniewicki M K. XGBOD: Improving supervised outlier detection with unsupervised representation learning. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN), 2018: 1-8.
[19] 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.
[20] 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.

[21] Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems 30 (NIPS), 2017: 3146-3154.
[22] Ribeiro M T, Singh S, Guestrin C. '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, 2016: 1135-1144.
[23] Vovk V, Gammerman A, Shafer G. Algorithmic Learning in a Random World. New York, NY, USA: Springer, 2005.
[24] Angelopoulos A N, Bates S. Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning, 2023, 16(4): 494-591.
[25] Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 1995, 57(1): 289-300.