Innovation and Technology Studies
Print ISSN: 3007-6919
Online ISSN: 3007-6927
DOI: https://doi.org/10.61784/its4029

COSTAR: EXPLAINABLE AND COST-SENSITIVE MACHINE LEARNING FOR TAX-AWARE CORPORATE FRAUD RISK DETECTION

Qiao Xu[1*], RuFeng Zhu[2]
[1]*The University of Delaware, Newark 19716, DE, USA.*
[2]*Brandeis University, Waltham 02453, MA, USA.*
[*]*Corresponding Author: Qiao Xu*

Abstract: Tax and securities authorities can examine only a small fraction of firms each year, so risk-based audit selection must cope with extremely rare and delayed fraud labels, firm-specific misclassification costs, and the obligation to justify every audit. This paper presents CoSTAR, a cost-sensitive, tax-aware and explainable framework for corporate fraud risk detection. CoSTAR enriches raw financial-statement items with twelve tax-aware indicators built from income-tax expense and taxes payable, scores firms with a multi-view balanced-bagging gradient-boosting ensemble whose views are selected on a case-disjoint calibration year, converts scores into audit decisions with an instance-dependent cost model, a budgeted expected-benefit rule and conformal risk control, and explains each flag with TreeSHAP reason codes, TreeSHAP-IQ Shapley interactions and the shape functions of a glass-box Kolmogorov–Arnold network. On the public AAER benchmark of U.S. listed firms (146,045 firm-years) with rolling out-of-time tests over 2003–2008, CoSTAR attains the highest mean AUC (0.707) among eight detectors, nearly triples NDCG@1% relative to the RUSBoost benchmark (0.0287 versus 0.0097), and delivers the largest net savings under a 1% audit budget (23.1% of the potential fraud loss). The same cost information yields far larger savings at decision time than inside the training loss (45.3% versus 9.0% at a 5% budget for LightGBM), and case-disjoint calibration markedly reduces the miss-rate excess of conformal guarantees under temporal shift.
Keywords: Corporate fraud detection; Tax risk; Cost-sensitive learning; Explainable artificial intelligence; Conformal risk control; Kolmogorov–Arnold networks

1 INTRODUCTION

Corporate fraud is costly and largely hidden. Dyck *et al.* estimate that only about one third of corporate frauds are detected in normal times and that fraud destroys about 1.6% of equity value every year [1]. Tax and securities authorities therefore rely on risk-based selection to allocate scarce examination resources, and data-driven tax risk detection has become an active research field [2]. The financial statements of listed firms are a natural common data source for this task because tax and financial reporting are tightly linked: book–tax differences predict the adjustments proposed by tax auditors [3], aggressive tax reporting co-moves with aggressive financial reporting [4], and tax aggressiveness is systematically related to the incidence of alleged accounting fraud [5-6].

Turning these signals into an audit list raises four coupled requirements. First, labels are extremely rare and delayed: enforcement releases appear years after the misconduct, and fewer than 1.3% of firm-years are labeled fraudulent in any year of the benchmark used here [7]. Second, misclassification costs are asymmetric and firm-specific: an undetected fraud in a large company destroys far more value than one in a small company, whereas examination effort grows less than proportionally with client size [8-9]. Third, an audit is an administrative act whose grounds must be explained to examiners, review bodies and the audited firm. Fourth, an authority needs a statistical statement about the frauds an audit list leaves out, not only a ranking score.

Machine-learning fraud detectors in accounting are typically trained and compared on cost-agnostic criteria such as AUC and NDCG@*k* [7, 10-12]. Cost-sensitive fraud classifiers embed costs into training but rarely provide explanations or guarantees [13-15], and comparative evidence indicates that the decision stage can matter more than the training objective [16]. Tax-fraud studies mostly analyze confidential declarations with unsupervised anomaly detection or supervised models on proprietary samples [17-21], which limits reproducibility and comparison.

This paper introduces **CoSTAR** (COst-Sensitive, Tax-Aware and explainable Risk detection), a framework that addresses the four requirements jointly on a public benchmark. The contributions are as follows.

1) A tax-aware representation that adds twelve indicators derived from income-tax expense and taxes payable (effective tax rate level, deviation, change and volatility; book–tax gap and its change; tax payable and tax expense dynamics; interest tax shield) to raw statement items and the Dechow ratios.

2) A multi-view balanced-bagging scorer that fuses a raw-statement view and a tax-aware view, together with a case-disjoint calibration protocol that extends the serial-fraud treatment of [7] to the calibration year.

3) A decision layer that separates risk estimation from cost: an instance-dependent cost model, a budgeted expected-benefit rule, and conformal risk control (CRC) that turns a target miss rate into an audit threshold [22].

© By the Author(s) 2024, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

4) An explanation layer that combines TreeSHAP reason codes [23], any-order Shapley interactions computed with TreeSHAP-IQ [24], and the univariate shape functions of a glass-box Kolmogorov–Arnold network (KAN) [25].
5) A rolling out-of-time evaluation on 146,045 U.S. firm-years against seven detectors, covering ranking quality, net savings, risk control and explanations.

2 RELATED WORK

2.1 Accounting Fraud and Tax Risk Detection

Early detectors relied on ratio-based scores such as the M-score and the F-score [10-11]. Bao *et al.* showed that RUSBoost trained on 28 raw accounting items outperforms ratio-based logistic models out of time and released the AAER-labeled data set used in this study [7, 26]. Bertomeu *et al.* applied gradient-boosted trees to accounting, market and audit variables and found that accounting variables become informative mainly through interactions [12], while Kim *et al.* cast misstatement detection as a multi-class cost-sensitive problem [13]. On the tax side, Mills linked book–tax differences to proposed audit adjustments [3], Frank *et al.* documented a positive relation between tax and financial reporting aggressiveness [4], and Lennox *et al.* related tax aggressiveness to alleged accounting fraud [5]; Hanlon and Heitzman review the measurement of tax avoidance [6]. Machine-learning tax-fraud studies include false-invoice detection [20], neural networks for personal income tax [21], spectral clustering for under-reporting [17], scalable anomaly detection for value-added tax and hybrid outlier detection for evasion risk [18-19]. A recent survey organizes fourteen families of tax risk detection methods and identifies unexplainable outputs and label scarcity as central bottlenecks [2].

2.2 Cost-Sensitive Learning

Cost-sensitive classification assigns different costs to different errors. Example-dependent cost-sensitive trees and instance-dependent logistic and boosting models minimize expected cost inside the training objective [14-15]. Vanderschueren *et al.* compared such predict-and-optimize strategies with predict-then-optimize pipelines on nine data sets and found the decision-making strategy to be generally more influential than the training loss [16]. Audit-fee research provides an empirical basis for cost structures: audit effort rises with client size with an elasticity well below one [8-9].

2.3 Explainability and Risk Control

SHAP unifies additive attribution methods [23], and TreeSHAP-IQ computes exact any-order Shapley interactions for tree ensembles in a single recursive traversal [24]. Glass-box alternatives include explainable boosting machines (EBM) and KANs [25, 27], whose learnable univariate edge functions can be inspected directly. Conformal risk control generalizes split conformal prediction to bound the expectation of any monotone loss with finite-sample validity under exchangeability [22], and its non-exchangeable extension weights calibration points by their relevance [28]. These components have not yet been combined for corporate fraud risk detection with tax-aware inputs.

3 PROBLEM FORMULATION

Let firm i in fiscal year t be described by a feature vector $\mathbf{x}_i$ and a label $y_i \in \{0, 1\}$, where $y_i = 1$ if the firm-year is later named in an SEC Accounting and Auditing Enforcement Release (AAER). A detector produces a risk score $s(\mathbf{x}_i)$, converted into a within-year percentile rank $r_i \in (0, 1]$. An audit decision $\delta_i \in \{0, 1\}$ incurs an investigation cost c_i whenever $\delta_i = 1$ and a loss L_i whenever a fraud is left unaudited. Following the evidence on economies of scale in audit effort and on fraud losses that scale with firm value [1, 8-9], both costs are expressed relative to the median firm of the year:

$$c_i = \left(A_i / \bar{A}_t\right)^{\beta}, \quad L_i = \rho \left(A_i / \bar{A}_t\right), \tag{1}$$

where A_i denotes total assets (winsorized at the yearly 99th percentile and floored at \$1 million), $\bar{A}_t$ is the yearly median, $\beta = 0.5$ is the size elasticity of audit effort, and ρ is the loss-to-audit-cost ratio of a median-sized firm ($\rho = 20$ by default and $\rho \in [2, 100]$ in the sensitivity analysis). The total cost of a decision vector is

$$C(\delta) = \sum_i \delta_i c_i + \sum_i (1 - \delta_i) y_i L_i, \tag{2}$$

and decisions are evaluated by the net savings relative to the no-audit policy,

$$\mathrm{NS}(\delta) = \left[\sum_i \delta_i y_i L_i - \sum_i \delta_i c_i\right] / \sum_i y_i L_i, \tag{3}$$

which is the fraction of the potential fraud loss recovered net of examination costs. Under a budget, the authority audits at most $k \cdot n_t$ of the n_t firms of a year. Given calibrated probabilities p_i, the unconstrained cost-minimizing rule is the Bayes rule $\delta_i = 1[p_i L_i > c_i]$, and the budgeted optimum audits the $k \cdot n_t$ firms with the largest expected benefit $b_i = p_i L_i - c_i$. In addition, the authority may require that the audit rule misses at most a fraction α of frauds, i.e., $\mathrm{E}[1\{\delta = 0\} \mid y = 1] \le \alpha$.

4 THE COSTAR FRAMEWORK

Figure 1 summarizes the four layers of CoSTAR. Each layer can be inspected separately, which keeps risk estimation, cost policy, statistical guarantees and justification apart—an organization that mirrors how audit units, policy makers and review bodies interact in practice.

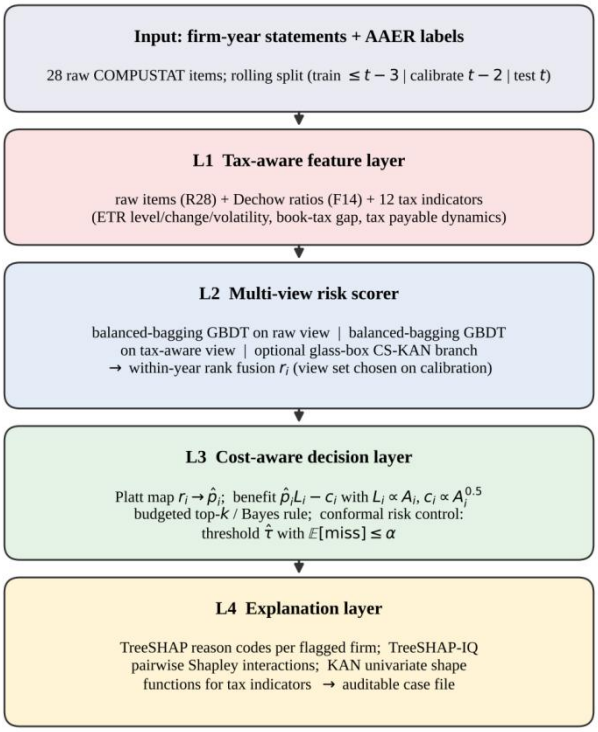

Figure 1 Architecture of CoSTAR

Note: Layer L1 builds tax-aware features, L2 produces a fused within-year risk rank, L3 converts ranks into cost-aware and risk-controlled audit decisions, and L4 attaches explanations to every flagged firm.

4.1 Tax-Aware Feature Layer

CoSTAR retains the 28 raw items (R28) of [7] and the 14 ratios (F14) of [11] and adds the 12 tax indicators (TAX) listed in Table 1, computed from income taxes (txt), taxes payable (txp), income before extraordinary items (ib), interest expense (xint) and total assets (at). Pre-tax income is approximated by PI = ib + txt, and the statutory federal rate τ_t equals 34% before 1993 and 35% afterwards. The book–tax gap grosses up tax expense by the statutory rate, following the logic used to relate book–tax differences to audit adjustments and tax avoidance [3, 6]. Changes are computed only when the previous fiscal year of the same firm is observed, the ETR is undefined for loss firms, and all continuous tax ratios are winsorized at the yearly 1st and 99th percentiles.

Table 1 Tax-Aware Indicators (TAX Block)

Indicator	Definition	Signal
etr	txt/PI if PI > 0, clipped to [0, 1]	tax burden on book income
etr_dev	etr minus yearly median	deviation from peers
d_etr	etr(t) − etr(t−1)	abrupt rate change
etr_vol	std. of etr over t−2…t	tax smoothing / instability
btd	(PI − txt/τ)/at	book–tax gap
d_btd	btd(t) − btd(t−1)	widening divergence
txp_at	txp/at	unpaid tax intensity
d_txp	Δtxp / average at	payable build-up
txt_at	txt/at	tax expense intensity
d_txt	Δtxt / average at	tax expense dynamics
txp_txt	txp/txt if txt > 0	share of tax unpaid
tax_shield	τ·xint/at	interest tax shield

4.2 Multi-View Balanced-Bagging Scorer

Extreme class imbalance is handled by balanced bagging. Each of B = 10 bags contains all positive training firm-years and an equally large random sample of negatives; a LightGBM learner with 300 trees, learning rate 0.03, 15 leaves and row/column subsampling of 0.8/0.6 is fitted to each bag [29], and bag probabilities are averaged. Two views are trained: a raw view on R28 and a tax-aware view on R28 ∪ F14 ∪ TAX. Because the views produce scores on different scales, they are fused through within-year percentile ranks,

$$r_i = (1/V) \sum_v \mathrm{rank}_v\big(s_v(\mathbf{x}_i)\big) \,/\, n_t, \tag{4}$$

which also removes year-specific score drift and requires no labels. A glass-box branch is available in the form of a cost-sensitive KAN (CS-KAN). With a single layer, the model is $f(\mathbf{x}) = b + \Sigma_j \varphi_j(x_j)$ with $\varphi_j(x) = w_j \cdot \mathrm{silu}(x) + \Sigma_m c_{jm} B_m(x)$, where B_m are cubic B-splines on a six-interval grid over quantile-normalized inputs [25]. CS-KAN is trained with class-balanced weighted cross-entropy, an L2 penalty and a second-difference roughness penalty (Adam, 20 epochs). The view set—raw and tax-aware views, optionally extended with CS-KAN, the logistic F-score model or a ratio-and-tax view—is chosen by pooled AUC on the calibration years only.

4.3 Case-Disjoint Temporal Calibration

For test year t, all models are trained on fiscal years 1991 to $t - 3$ and calibrated on year $t - 2$, so that calibrated quantities are estimated on data that lie outside the training window, as test data do. Bao *et al.* relabel training firm-years of AAER cases that continue into the test year as non-fraud [7], because a fraud case typically spans several consecutive years. CoSTAR applies the same rule to the calibration year: firm-years of cases that are active in year $t - 2$ are relabeled in the training data. Without this step, many calibration-year frauds are continuations of training cases, their scores are inflated, and every calibrated quantity—probabilities, thresholds and conformal guarantees—becomes optimistic (Section 6.4).

4.4 Cost-Aware Decision Layer

Ranks are mapped to probabilities by Platt scaling fitted on the calibration year, $p_i = \sigma(a \cdot \mathrm{logit}(r_i) + b)$. Three decision rules are supported: risk ranking (top-k by p_i), benefit ranking (top-k by $b_i = p_i L_i - c_i$), and the unconstrained Bayes rule $\delta_i = 1[b_i > 0]$. Costs enter only at decision time; the scorer itself is trained on balanced bags without cost weights. Section 6.2 examines this design against instance-dependent cost-weighted training.

4.5 Conformal Risk Control

For an audit rule $\delta_i = 1[z_i \geq \tau]$ with a fixed score z, the miss loss of a fraud is $\ell_i(\tau) = 1[z_i < \tau]$, which is monotone in τ and bounded by 1. Given the n calibration frauds, CRC selects [22]

$$\hat{\tau} = \sup\{\tau : (n/(n+1))\,\hat{R}_n(\tau) + 1/(n+1) \leq \alpha\}, \tag{5}$$

where $\hat{R}_n(\tau)$ is the empirical miss rate of the calibration frauds. Under exchangeability, a new fraud is missed with probability at most α. Two scores are used: the risk score $z_i = \mathrm{logit}(r_i)$ and the cost-weighted score $z_i = \mathrm{logit}(r_i) + \log(L_i/c_i)$, which orders firms by cost-weighted odds, as the Bayes rule does, and remains label-free so that the guarantee is not affected by the calibration fit.

4.6 Explanation Layer

Every flagged firm receives TreeSHAP reason codes computed exactly on each bag model and averaged across bags [23]; global importance aggregates mean |SHAP| by feature and by block (R28, F14 and TAX). Pairwise structure is quantified with the second-order k-SII index computed by TreeSHAP-IQ for the highest-risk firms [24]. The CS-KAN branch contributes glass-box shape functions φ_j that display how each tax indicator shifts the log-odds of fraud over its entire range.

5 EXPERIMENTAL SETUP

5.1 Data and Protocol

The AAER benchmark released by [7] contains 146,045 firm-years of U.S. listed firms (fiscal years 1990–2014) with 964 fraud firm-years belonging to 412 AAER cases; the yearly fraud prevalence ranges from 0.07% to 1.27%. As in [7], the evaluation covers the test years 2003–2008, for which detection lags are least likely to truncate the labels. Table 2 lists the six folds: training windows grow from 60,804 to 91,011 firm-years with 300–684 frauds, the calibration years contain 24–46 case-disjoint frauds, and the test years contain 35,166 firm-years with 261 frauds.

Table 2 Rolling Out-of-Time Folds

Test year	Training window	Train firm-years	Train frauds	Calib. frauds	Test firm-years	Test frauds
2003	1991-2000	60,804	300	45	5,981	69
2004	1991-2001	67,166	354	46	5,934	58
2005	1991-2002	73,233	427	43	5,863	45
2006	1991-2003	79,214	494	42	5,908	33
2007	1991-2004	85,148	558	34	5,868	30
2008	1991-2005	91,011	684	24	5,612	26

5.2 Detectors

Seven detectors are compared with CoSTAR: (1) a logistic model on the F14 ratios, i.e., the F-score specification of [11]; (2) RUSBoost on R28 with 300 trees, learning rate 0.1 and minimum leaf size 5, replicating [7, 26]; (3) LightGBM on the full feature set with 500 trees, learning rate 0.02 and 31 leaves [29]; (4) LightGBM-CW with class-balanced weights;

(5) LightGBM-IDCS with instance-dependent cost weights (L_i for frauds and c_i for non-frauds), the cost-sensitive training strategy of [14-15]; (6) EBM with five pairwise interactions [27]; and (7) the stand-alone CS-KAN. Except for the two replications, every detector uses the full tax-aware feature set, so differences reflect modeling rather than inputs.

5.3 Metrics and Implementation

Ranking quality is measured by AUC, average precision (AP), and the NDCG, sensitivity and precision within the top 1% of each test year, as in [7]. Yearly metrics are averaged over the six folds; the pooled AUC uses within-year ranks, and a paired bootstrap with 1,000 resamples gives 95% confidence intervals for AUC differences. Cost performance is measured by the pooled net savings (3), i.e., the ratio of the numerator and denominator summed over 2003–2008, and by the exposure recall $\Sigma\delta_i y_i L_i / \Sigma y_i L_i$. CRC is assessed by the pooled realized miss rate and the audit rate. All experiments run on a single CPU core with LightGBM 4.7, imbalanced-learn, InterpretML, JAX and shapiq; one CoSTAR fit (two views × ten bags) takes 10–17 s per fold.

6 RESULTS

6.1 Ranking Performance

Table 3 Out-of-Time Ranking Performance, 2003–2008

Detector	AUC (sd)	Pooled AUC	AP	NDCG @1%	Sens. @1% (%)	Prec. @1% (%)	ΔAUC
Logit F-score	0.686 (0.026)	0.681	0.0177	0.0342	4.45	2.54	+0.032*
RUSBoost	0.699 (0.031)	0.702	0.0161	0.0097	1.14	1.12	+0.012
LightGBM	0.694 (0.025)	0.697	0.0164	0.0181	1.65	1.41	+0.017*
LightGBM-CW	0.674 (0.033)	0.679	0.0201	0.0349	2.18	1.97	+0.034*
LightGBM-IDCS	0.591 (0.035)	0.582	0.0160	0.0286	1.99	1.70	+0.132*
EBM	0.675 (0.017)	0.678	0.0146	0.0191	2.04	1.96	+0.035*
CS-KAN	0.650 (0.026)	0.653	0.0147	0.0269	3.21	2.82	+0.060*
CoSTAR	**0.707 (0.036)**	**0.714**	**0.0185**	**0.0287**	**2.60**	**2.26**	**—**

Note: Yearly means over six test years (standard deviation of AUC in parentheses). ΔAUC: pooled AUC of CoSTAR minus that of the detector; * marks a paired-bootstrap 95% interval excluding zero.

Table 3 reports the ranking results. CoSTAR obtains the highest mean AUC (0.707) and the highest pooled AUC (0.714). The paired-bootstrap intervals of the pooled AUC difference exclude zero against the logistic F-score model (+0.032, 95% CI [0.004, 0.062]), LightGBM (+0.017, [0.001, 0.032]), LightGBM-CW (+0.034), EBM (+0.035), CS-KAN (+0.060) and LightGBM-IDCS (+0.132); against RUSBoost the difference is +0.012 (95% CI [−0.004, 0.028], one-sided $p = 0.074$). At the top of the list, CoSTAR nearly triples the NDCG@1% of RUSBoost (0.0287 versus 0.0097) and more than doubles its sensitivity (2.60% versus 1.14%) and precision (2.26% versus 1.12%) at the 1% cut-off.

No single detector dominates every ranking metric. LightGBM-CW reaches the highest AP (0.0201) and NDCG@1% (0.0349) but a markedly lower AUC (0.674), and the linear F-score model attains the highest top-1% sensitivity (4.45%) with an AUC of 0.686. Instance-dependent cost weights damage ranking severely (AUC 0.591): weights proportional to firm size make the learner concentrate on separating large frauds while the rest of the population is ranked almost at random. The two glass-box detectors trail CoSTAR by 0.03–0.06 in AUC, which positions them as explanation devices rather than as stand-alone scorers.

6.2 Cost-Aware Audit Selection

Table 4 Pooled Net Savings at ρ = 20

Detector	Rule	NS @1%	NS @5%	NS @10%	ER @5%	NS Bayes
Logit F-score	risk	-0.001	-0.012	0.053	0.014	—
Logit F-score	benefit	0.205	0.487	0.499	0.675	0.485
RUSBoost	risk	0.096	0.391	0.387	0.513	—
RUSBoost	benefit	0.214	0.386	0.427	0.583	0.429
LightGBM	risk	0.076	0.143	0.319	0.231	—
LightGBM	benefit	0.133	0.453	0.486	0.637	0.500
LightGBM-CW	risk	0.123	0.217	0.326	0.301	—
LightGBM-CW	benefit	0.153	0.466	0.458	0.644	0.458
LightGBM-IDCS	risk	0.094	0.090	0.143	0.131	—
LightGBM-IDCS	benefit	0.143	0.314	0.312	0.432	0.316
EBM	risk	0.144	0.230	0.276	0.329	—
EBM	benefit	0.176	0.419	0.453	0.602	0.450
CS-KAN	risk	0.180	0.228	0.408	0.336	—
CS-KAN	benefit	0.162	0.421	0.445	0.609	0.445
CoSTAR	risk	0.205	0.344	0.390	0.458	—
CoSTAR	**benefit**	**0.231**	**0.398**	**0.462**	**0.590**	**0.459**
Size-only	size	-0.001	0.358	0.421	0.564	—

Note: NS: net savings (3) pooled over 2003–2008; ER: exposure recall; risk = top-k by probability; benefit = top-k by $p_iL_i - c_i$; Bayes = unconstrained rule $p_iL_i > c_i$.

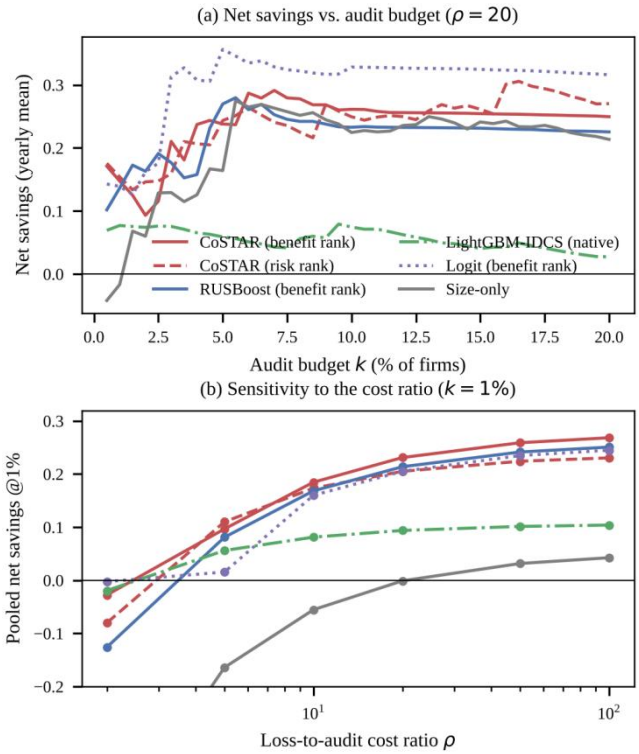

Figure 2 Cost-Aware Evaluation

Note: (a) Net savings versus audit budget at ρ = 20 (mean of the yearly values). (b) Pooled net savings at a 1% budget versus the loss-to-audit-cost ratio ρ.

Table 4 and Figure 2 report the cost results, from which three findings emerge. First, CoSTAR with benefit ranking recovers 23.1% of the potential fraud loss net of audit costs under a 1% budget, the highest value among all detector–rule combinations (RUSBoost 21.4%, logistic F-score 20.5%, EBM 17.6%, CS-KAN 16.2%, LightGBM 13.3%), whereas auditing the largest 1% of firms recovers nothing (−0.1%). Figure 2(b) shows that CoSTAR keeps the highest 1%-budget savings among the plotted policies for every cost ratio $\rho \geq 10$; at $\rho = 100$, for example, CoSTAR recovers 26.8% against 25.1% for RUSBoost and 24.5% for the logistic model.

Second, the decision rule matters more than the training loss. Benefit ranking raises the savings of seven of the eight detectors at both the 1% and the 5% budget. LightGBM trained with instance-dependent cost weights recovers only 9.0% at a 5% budget with its native ranking and 31.4% with benefit ranking, whereas the cost-insensitive LightGBM recovers 45.3% when the same cost information is applied at decision time; under the unconstrained Bayes rule the comparison is 31.6% versus 50.0%. This corroborates in an audit setting where costs scale with firm size [16].

Third, as the budget widens, exposure weighting becomes more valuable than ranking quality. At 5% and 10% budgets, the savings of all benefit-ranked detectors except LightGBM-IDCS lie between 38.6% and 49.9%; the calibrated logistic model is the most profitable (48.7% and 49.9%), and CoSTAR yields 39.8% and 46.2%, above the size-only policy (35.8% and 42.1%). Figure 2(a) displays the same pattern on yearly averages: CoSTAR leads at the 1% budget, and the logistic benefit curve dominates from 3% onward. The budget therefore determines which component carries the value: ranking quality under tight budgets and calibrated exposure weighting under wide ones.

6.3 Ablation Study

Table 5 View Selection and Ablation

Configuration	Calib. AUC	AUC	Pooled AUC	AP	NDCG @1%
† UB(raw) + UB(full)	**0.784**	**0.707**	**0.714**	**0.0185**	**0.0287**
UB(raw) + UB(full) + CS-KAN	0.780	0.694	0.699	0.0183	0.0318
UB(raw) + UB(full) + Logit	0.778	0.718	0.722	0.0170	0.0170
UB(raw) + UB(ratio+tax)	0.776	0.708	0.712	0.0173	0.0194
UB(raw) + UB(ratio+tax) + CS-KAN	0.774	0.692	0.697	0.0175	0.0230
UB(raw) + UB(raw+ratio)	—	0.703	0.708	0.0180	0.0275
UB(raw) + UB(ratio)	—	0.703	0.708	0.0182	0.0298
UB(raw) + UB(ratio) + CS-KAN	—	0.690	0.695	0.0178	0.0306
UB(raw) only	—	0.706	0.712	0.0163	0.0244
UB(full) only	—	0.691	0.697	0.0184	0.0338
UB(ratio+tax) only	—	0.660	0.662	0.0216	0.0331

Configuration	Calib. AUC	AUC	Pooled AUC	AP	NDCG @1%
no bagging: LightGBM(full)	—	0.694	0.697	0.0164	0.0181

Note: UB: balanced-bagging LightGBM; raw = R28, full = R28 ∪ F14 ∪ TAX, ratio = F14. † Configuration selected by calibration AUC (CoSTAR). The first five rows are the candidate set; calibration AUC is reported for candidates only.

Table 5 lists the five candidate fusions and the corresponding ablations. The raw-plus-tax-aware pair attains the highest calibration AUC (0.784) and is therefore the CoSTAR scorer. Removing the twelve tax indicators from the tax-aware view lowers the pooled AUC from 0.714 to 0.708; the paired bootstrap difference of 0.0055 (95% CI [0.0019, 0.0089]) is small but significant, and NDCG@1% decreases from 0.0287 to 0.0275. Fusion combines complementary strengths: the raw view alone reaches a similar AUC (0.706) but a lower AP (0.0163) and NDCG@1% (0.0244), whereas the tax-aware view alone has a higher NDCG@1% (0.0338) but a lower AUC (0.691). On the same features, balanced bagging nearly doubles the NDCG@1% of a single LightGBM (0.0338 versus 0.0181) at a comparable AUC. Adding the CS-KAN branch raises NDCG@1% to 0.0318 but lowers the AUC to 0.694, which supports its use as an explanation companion. The candidate that adds the logistic model reaches the highest test AUC (0.718) without the highest calibration AUC; its difference from CoSTAR is not significant (95% CI [−0.002, 0.017]). Three bagging seeds give mean AUCs between 0.705 and 0.708.

6.4 Conformal Risk Control

Table 6 Conformal Risk Control on Out-of-Time Test Years

Detector	Score	α	Miss rate	Audit rate (%)	Exposure recall	NS
CoSTAR	risk	0.2	0.333	39.0	0.893	0.450
CoSTAR	risk	0.3	0.421	29.2	0.842	0.462
CoSTAR	risk	0.4	0.556	22.1	0.769	0.453
CoSTAR	cost-wtd	0.2	0.307	39.6	0.969	0.436
CoSTAR	cost-wtd	0.3	0.398	31.6	0.947	0.458
CoSTAR	cost-wtd	0.4	0.559	20.6	0.903	0.490
RUSBoost	risk	0.2	0.253	43.3	0.907	0.444
RUSBoost	risk	0.3	0.337	34.0	0.882	0.460
RUSBoost	risk	0.4	0.502	23.3	0.777	0.447
RUSBoost	cost-wtd	0.2	0.299	41.4	0.966	0.421
RUSBoost	cost-wtd	0.3	0.402	32.2	0.955	0.460
RUSBoost	cost-wtd	0.4	0.506	24.5	0.911	0.482
LightGBM	risk	0.2	0.284	42.6	0.905	0.499
LightGBM	risk	0.3	0.410	30.4	0.767	0.436
LightGBM	risk	0.4	0.567	21.0	0.706	0.443
LightGBM	cost-wtd	0.2	0.276	42.6	0.970	0.437
LightGBM	cost-wtd	0.3	0.421	28.6	0.934	0.488
LightGBM	cost-wtd	0.4	0.598	21.4	0.842	0.487

Note: Miss rate and NS pooled over 2003–2008; audit rate and exposure recall are yearly means; ρ = 20.

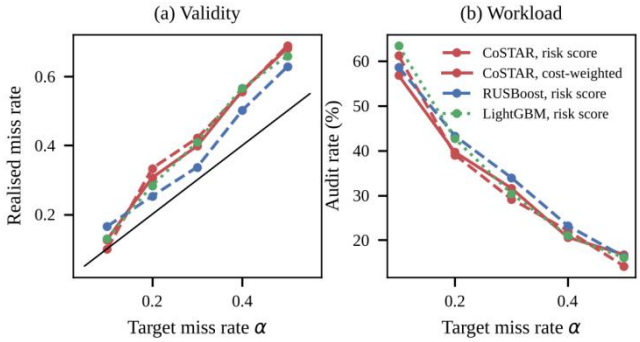

Figure 3 Conformal Risk Control

Note: (a) Pooled realized miss rate versus target α (diagonal: exact control). (b) Audit rate required by the calibrated threshold.

Table 6 and Figure 3 summarize CRC. Case-disjoint calibration is decisive: when calibration-year frauds include continuations of training cases, the realized miss rate of the same scorer averages 0.487 for $\alpha = 0.2$ and 0.644 for $\alpha = 0.3$, compared with 0.345 and 0.430 after case-disjoint relabeling, and the calibration-year AUC falls from 0.854 to 0.784, closer to the test-year level. At $\alpha = 0.1$ the guarantee is met (pooled miss rate 0.100 with the risk score) at an audit rate of 61%. For $0.2 \leq \alpha \leq 0.4$, the realized miss rates of all detectors exceed the targets by 0.04–0.20, which quantifies the temporal shift between calibration and test years that lie two fiscal years apart. The cost-weighted score changes which frauds are missed: at $\alpha = 0.3$ it audits 31.6% of firms and captures 94.7% of the loss exposure, against 29.2% and 84.2% for the risk score, and at $\alpha = 0.4$ it captures 90.3% versus 76.9% of the exposure with fewer audits (20.6% versus 22.1%). Recency-weighted calibration over several years, as enabled by non-exchangeable CRC [28], is the natural way to absorb the remaining shift.

6.5 Explanations

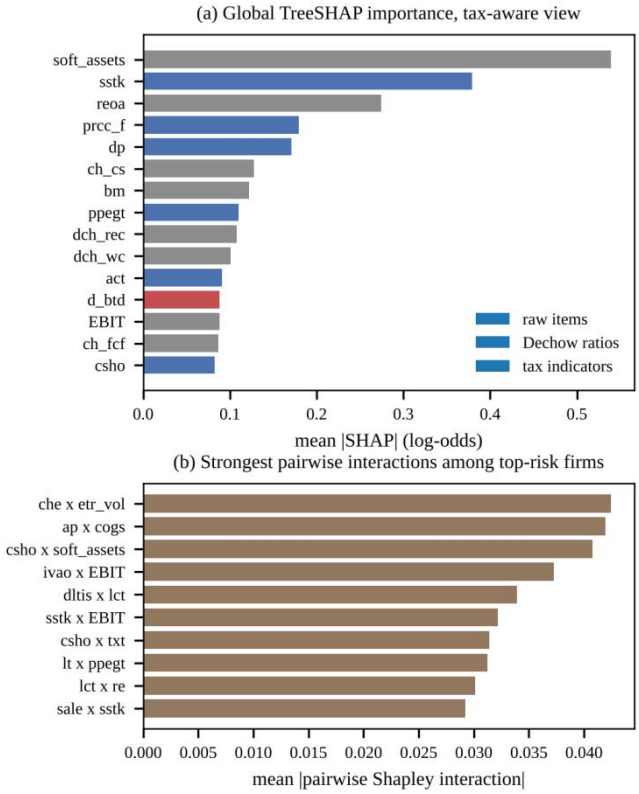

Figure 4 Explanation Layer

Note: (a) Mean |TreeSHAP| of the tax-aware view over all test years, colored by feature block. (b) Strongest second-order k-SII interactions computed with TreeSHAP-IQ for the 40 highest-risk firms of the 2003 fold.

Global TreeSHAP attribution of the tax-aware view (Figure 4(a)) is led by soft assets (soft_assets), common stock issuance (sstk), retained earnings over assets (reoa), the year-end share price (prcc_f), depreciation and amortization (dp) and the change in cash sales (ch_cs). Soft assets and equity issuance coincide with the misstatement determinants identified in [11], which supports the face validity of the learned risk model. The raw items account for 48.1% of the total attribution, the Dechow ratios for 39.8% and the tax indicators for 12.1%; the most influential tax signals are the change in the book–tax gap (d_btd), ETR volatility (etr_vol), the interest tax shield, the tax expense intensity and the book–tax gap itself.

Tax information acts strongly through interactions. Among the 40 highest-risk firms of the 2003 fold, the strongest k-SII pair is cash holdings × ETR volatility (che × etr_vol), and the pair formed by shares outstanding and total income tax (csho × txt) also ranks among the seven strongest pairs (Figure 4(b)). This pattern is consistent with the view that tax and financial reporting choices are made jointly [4-5], and it explains why the tax block adds discriminative power in the fused model although its marginal attribution share is moderate.

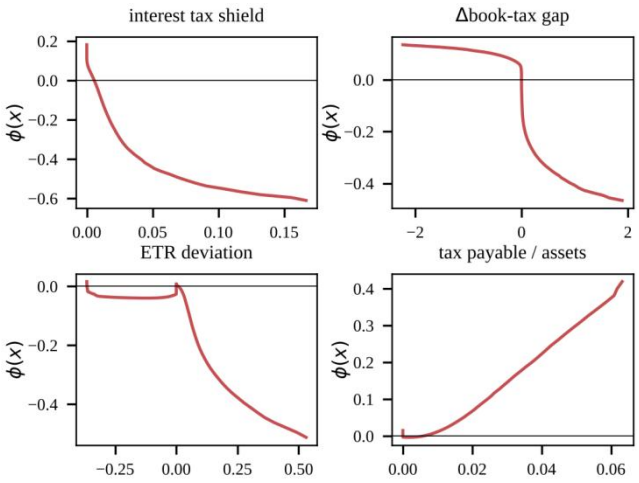

Figure 5 Glass-Box CS-KAN Shape Functions φ(x) of the Four Tax Indicators with the Largest Functional Range (2008 Fold), Plotted over the 3rd–97th Percentile of Each Indicator

Note: Each curve is the exact additive contribution of the indicator to the log-odds of fraud.

Figure 5 shows the CS-KAN shape functions of the tax indicators with the largest functional range. Because CS-KAN is additive, each curve is the exact contribution of the indicator to the log-odds of fraud rather than a post-hoc approximation, so an examiner can read off how a given value of each indicator shifts the risk of a firm. The fraud log-odds decline as the interest tax shield and the ETR deviation increase, drop sharply once the change in the book–tax gap turns positive (a widening gap), and rise almost linearly with tax payable intensity; the lower risk attached to a widening book–tax gap agrees with the finding of [5] that firms committing accounting fraud tend to be less tax-aggressive. Across all inputs, the largest functional ranges belong to soft assets, gross property, plant and equipment, share price and retained earnings, followed by shares outstanding, stock issuance, current assets and total assets, with the interest tax shield (tax_shield) as the highest-ranked tax indicator, an ordering that agrees closely with the TreeSHAP ranking of the black-box view.
For the highest-ranked detected fraud of the 2003 test year (the 99.88th risk percentile), the reason codes are soft assets of 0.88 (+0.62 log-odds), a book-to-market ratio of 0.11 (+0.32), a change in return on assets of -0.12 (+0.29), common stock issuance (+0.28), the depreciation index (+0.23) and retained earnings over assets (+0.21). Together they form an auditable narrative: an asset-light, richly valued firm whose profitability is deteriorating while it raises equity.

7 DISCUSSION

The results translate into four design rules for tax and securities authorities. First, the audit budget decides which component carries the value: with budgets near 1%, ranking quality drives savings and the multi-view scorer is preferable, whereas with wider budgets calibrated exposure weighting dominates and simpler, well-calibrated scorers become competitive. Second, cost information belongs in the decision layer; embedding firm-size costs into training weights degrades both ranking and savings. Third, any calibrated or guaranteed quantity in data with multi-year fraud cases requires case-disjoint calibration, otherwise thresholds are systematically too permissive. Fourth, the explanation layer yields case files that combine reason codes, interaction evidence and glass-box tax curves, which supports the justification of each audit.
The SEC AAER labels capture enforcement-detected accounting fraud, and the tax indicators are computed from public statements; tax administrations hold richer return-level data and audit outcomes that plug into the same four layers without modification. Natural extensions are recency-weighted non-exchangeable risk control [28], network features that link related taxpayers, and the joint modeling of tax adjustments and accounting enforcement.

8 CONCLUSION

This paper presented CoSTAR, a cost-sensitive, tax-aware and explainable framework for corporate fraud risk detection that integrates tax-aware features, a multi-view balanced-bagging scorer, an instance-dependent cost-aware decision layer with conformal risk control, and an explanation layer built on TreeSHAP, TreeSHAP-IQ and KAN shape functions. On 146,045 U.S. firm-years with rolling out-of-time tests, CoSTAR achieved the highest AUC among eight detectors, nearly tripled the NDCG@1% of the RUSBoost benchmark, and delivered the largest net savings under a 1% audit budget. The experiments further established that cost information yields far larger savings at decision time than inside the training loss, and that case-disjoint calibration is required for credible conformal guarantees. These findings provide a reproducible basis for risk-based audit selection in tax administrations and securities regulators.

COMPETING INTERESTS

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

REFERENCES

[1] Dyck A, Morse A, Zingales L. How pervasive is corporate fraud? Review of Accounting Studies, 2024, 29(1): 736-769.
[2] Zheng Q, Xu Y, Liu H, et al. A survey of tax risk detection using data mining techniques. Engineering, 2024, 34: 43-59.
[3] Mills L F. Book-tax differences and Internal Revenue Service adjustments. Journal of Accounting Research, 1998, 36(2): 343-356.
[4] 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.
[5] Lennox C, Lisowsky P, Pittman J. Tax aggressiveness and accounting fraud. Journal of Accounting Research, 2013, 51(4): 739-778.
[6] Hanlon M, Heitzman S. A review of tax research. Journal of Accounting and Economics, 2010, 50(2-3): 127-178.
[7] 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.
[8] Simunic D A. The pricing of audit services: Theory and evidence. Journal of Accounting Research, 1980, 18(1): 161-190.
[9] Hay D C, Knechel W R, Wong N. Audit fees: A meta-analysis of the effect of supply and demand attributes. Contemporary Accounting Research, 2006, 23(1): 141-191.
[10] Beneish M D. The detection of earnings manipulation. Financial Analysts Journal, 1999, 55(5): 24-36.

[11] Dechow P M, Ge W, Larson C R, et al. Predicting material accounting misstatements. Contemporary Accounting Research, 2011, 28(1): 17-82.
[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] Kim Y J, Baik B, Cho S. Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. Expert Systems with Applications, 2016, 62: 32-43.
[14] Correa Bahnsen A, Aouada D, Ottersten B. Example-dependent cost-sensitive decision trees. Expert Systems with Applications, 2015, 42(19): 6609-6619.
[15] Höppner S, Baesens B, Verbeke W, et al. Instance-dependent cost-sensitive learning for detecting transfer fraud. European Journal of Operational Research, 2022, 297(1): 291-300.
[16] Vanderschueren T, Verdonck T, Baesens B, et al. Predict-then-optimize or predict-and-optimize? An empirical evaluation of cost-sensitive learning strategies. Information Sciences, 2022, 594: 400-415.
[17] de Roux D, Pérez 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 and Data Mining, 2018: 215-222.
[18] Vanhoeyveld J, Martens D, Peeters B. Value-added tax fraud detection with scalable anomaly detection techniques. Applied Soft Computing, 2020, 86: Art. 105895.
[19] 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: Art. 116409.
[20] González P C, Velásquez J D. Characterization and detection of taxpayers with false invoices using data mining techniques. Expert Systems with Applications, 2013, 40(5): 1427-1436.
[21] Pérez López C, Delgado Rodríguez M J, de Lucas Santos S. Tax fraud detection through neural networks: An application using a sample of personal income taxpayers. Future Internet, 2019, 11(4): Art. 86.
[22] Angelopoulos A N, Bates S, Fisch A, et al. Conformal risk control. In: Proceedings of the International Conference on Learning Representations (ICLR), 2024.
[23] Lundberg S M, Lee S-I. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, 2017, 30: 4765-4774.
[24] Muschalik M, Fumagalli F, Hammer B, et al. Beyond TreeSHAP: Efficient computation of any-order Shapley interactions for tree ensembles. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2024, 38(13). DOI: 10.1609/aaai.v38i13.29352.
[25] Liu Z, Wang Y, Vaidya S, et al. KAN: Kolmogorov–Arnold networks. arXiv:2404.19756, 2024.
[26] 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.
[27] Nori H, Jenkins S, Koch P, et al. InterpretML: A unified framework for machine learning interpretability. arXiv:1909.09223, 2019.
[28] Farinhas A, Zerva C, Ulmer D, et al. Non-exchangeable conformal risk control. In: Proceedings of the International Conference on Learning Representations (ICLR), 2024.
[29] 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.