Eurasia Journal of Science and Technology
Print ISSN: 2663-1024
Online ISSN: 2663-1016
DOI: https://doi.org/10.61784/ejst4162

EXACT CONCEPT-LEVEL EXPLANATIONS FOR EARLY DETECTION OF FINANCIAL MISSTATEMENTS

JiaYing Chen[1*], Xu Han[2], XuGuang Zhang[3], YongBin Yang[4]
[1]*Cornell University, Ithaca 14853, New York, USA.*
[2]*Renmin University of China, Beijing 100872, China.*
[3]*Independent Researcher.*
[4]*University of Southern California, Los Angeles 90089, California, USA.*
[*]*Corresponding Author: JiaYing Chen*

Abstract: Early warnings of material financial misstatements are useful only when auditors and regulators can understand why a firm is flagged. Machine learning models trained on financial statement data rank misstatement risk well, but their post-hoc explanations refer to individual line items, change when the model is retrained and often contradict established accounting evidence. This study proposes AGX, an accounting-grounded explainable model for the early detection of misstatement risk. AGX derives double-entry articulation features from consecutive financial statements, organises all predictors into ten accounting concepts, restricts every boosted tree to a single concept, imposes the monotone relations documented in the fraud-prediction literature and averages balanced ensembles on the log-odds scale. We prove that the resulting risk score decomposes exactly into concept contributions that coincide with group Shapley values and never contradict the imposed accounting priors. On the public AAER benchmark of 146,045 U.S. firm-years from 1990 to 2014, evaluated with a temporal design that uses only labels available at deployment, AGX matches the ranking accuracy of leading ensemble methods and detects the first year of misstatement episodes significantly better than LightGBM and the Explainable Boosting Machine (onset-year AUC of 0.722 versus 0.698 and 0.692). It is also the most robust model to measurement noise and missing prior-year data, and its explanations are exact, faithful, the most stable across retraining and free of theory violations.
Keywords: Financial misstatement; Accounting fraud detection; Explainable artificial intelligence; Shapley values; Monotone constraints; Early warning

1 INTRODUCTION

Material misstatements of financial statements impose large losses on investors, creditors and employees, and they erode trust in the information that capital markets rely on. The U.S. Securities and Exchange Commission (SEC) documents the most serious cases in its Accounting and Auditing Enforcement Releases (AAERs), but these releases typically arrive several years after the misstated statements were filed [1-2]. Investors, auditors and regulators therefore need tools that flag misstatement *risk* at the moment a set of statements becomes public, and ideally as soon as a misstatement episode begins.

Machine learning has substantially advanced this task. Building on accounting-based risk scores such as the Beneish M-score and the Dechow F-score [2-3], ensemble learners trained on raw financial statement items now achieve markedly higher out-of-sample detection rates [4-6]. The same flexibility that drives these gains, however, makes the resulting risk scores difficult to justify. Post-hoc explanation methods such as SHAP are routinely applied to such models [7-8], yet in the accounting domain they face four specific bottlenecks.

1. Vocabulary mismatch. Attributions are produced for individual line items (e.g., “common shares outstanding”), whereas auditors reason about accounting constructs such as accrual intensity, revenue recognition or financing pressure.
2. Incoherent aggregation. Accounting items are tightly linked by identities and are strongly dependent. Summing feature-level attributions within a construct does not, in general, yield the attribution of the construct itself, and the discrepancy depends on the attribution algorithm [8-9].
3. Theory-inconsistent and unstable explanations. With only a few hundred positive firm-years, unconstrained learners may fit response shapes that contradict well-established accounting evidence (e.g., lower risk for faster receivable growth), and retraining with a different random seed can reorder the explanation of the same firm-year.
4. Late rather than early detection. Pooled evaluations are dominated by later years of multi-year misstatement episodes, in which accumulated distortions are easy to see. The economically most valuable signal—the *onset* of an episode—is rarely measured, and training labels that would not yet be observable at deployment time are often used.

This paper addresses these bottlenecks with AGX (Accounting-Grounded eXplainable risk model), a misstatement risk model whose structure is dictated by accounting knowledge and whose explanations are exact by construction. AGX rests on three technical components. First, a set of *double-entry articulation features* measures how well a firm’s balance sheet changes reconcile with its income statement and financing flows; because every earnings overstatement must be booked against some balance sheet account, these residuals target where manipulation has to surface. Second, a *concept-partitioned, theory-monotone gradient boosting* learner restricts every tree to the features of a single accounting

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

concept and imposes the directional relations documented by Dechow et al. and Beneish [2-3]. Third, *balanced bagging* handles the extreme class imbalance while preserving the additive concept structure. We prove that the resulting risk score decomposes *exactly* into concept contributions that coincide with interventional group Shapley values, that these contributions are hierarchically coherent with feature-level Shapley values, and that they can never contradict the imposed theory.

The main contributions are as follows.

• We formalise early misstatement-risk detection under realistic label availability, including onset-year detection, episode-level detection delay and one-year-ahead onset warning, and we provide a temporal protocol that only uses labels that would have been observable at deployment.

• We derive a compact family of double-entry articulation features (clean-surplus, equity, debt and tax roll-forward residuals, abnormal receivable, inventory and payable growth) that encode accounting identities rather than generic ratios.

• We propose a concept-additive, theory-monotone boosting ensemble and establish three formal properties: exact concept attribution (Proposition 1), hierarchical coherence with feature-level and Owen-type attributions (Proposition 2), and closure under bagging together with monotone theory consistency (Proposition 3). Concept attributions cost a single pass over the trees instead of 2^m coalition evaluations.

• We conduct a comprehensive evaluation on the public AAER benchmark of Bao et al. (146,045 firm-years, 1990–2014) against ten baselines, with development-period tuning, a held-out test period, paired bootstrap tests, ablations, sensitivity and robustness analyses, and quantitative explanation-quality metrics (exactness, coherence, removal fidelity, seed and temporal stability, theory violations) [4].

The empirical results support these claims. Over the 2003–2008 test period AGX reaches a mean AUC of 0.713, statistically indistinguishable from XGBoost, LightGBM, EBM and the RUSBoost model of Bao et al. [4], with the smallest year-to-year dispersion among the non-linear models. It ranks misstatement onsets better than LightGBM and EBM (onset AUC 0.722 versus 0.698 and 0.692, $p < 0.05$), achieves the highest AUC in the later period 2009–2011 (0.732), loses no accuracy under 10% measurement noise, and is clearly more accurate for firms without prior-year statements (0.743 versus 0.675–0.695). Its concept explanations are exact to machine precision ($< 10^{-15}$), fully faithful under concept removal (removal error 0 versus 0.26–0.31 for the baselines), the most stable across random seeds and over time, and free of theory violations, whereas the baselines contradict accounting evidence for 9–42% of the perturbations we apply.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 formalises the detection tasks and the explanation requirements. Section 4 presents AGX and its theoretical properties. Section 5 describes the experimental design, Section 6 reports the results, and Section 7 discusses implications and concludes.

2 RELATED WORK

2.1 Detecting Financial Misstatements

Early detection models relied on a small number of theory-motivated accounting ratios. Beneish combined indices of receivables, margins, asset quality, sales growth, depreciation, leverage and accruals into the M-score [3], and later work showed that the score also predicts returns [10]. Dechow et al. assembled a comprehensive AAER sample and derived the F-score from accrual quality, financial performance, non-financial, off-balance-sheet and market-based measures [2]; their analysis established the directional evidence that we impose as theory priors. Subsequent studies introduced statistical and machine learning classifiers, including support vector machines with a financial kernel [11], systematic comparisons of classifiers and misclassification costs [12], meta-learning [13], multi-class cost-sensitive learning that separates intentional misstatements [14] and data-analytics preprocessing for rare-event prediction [6].

Bao et al. showed that an ensemble learner (RUSBoost) applied to 28 raw financial statement items outperforms ratio-based logistic models [4], and they published the benchmark data used in this paper together with a serial-fraud adjustment that prevents the same misstatement episode from appearing in both training and test samples. Bertomeu et al. used gradient-boosted trees with a broad set of accounting, market, governance and audit variables and interpreted the resulting models [5]. Other work adds textual information from annual reports [15-18], distributional properties of reported numbers [19], tax reporting [20] or dynamic ensemble selection [21]. Karpoff et al. document that conclusions depend on how misconduct is measured and when it is revealed [1], and Zavitsanos et al. show that evaluation choices such as temporal splitting and detection delay materially change reported performance [22]. Our work keeps the publicly available accounting data of Bao et al. fixed and focuses on how accounting structure can be built into the model so that its explanations are exact, coherent and consistent with theory, and on how early in an episode misstatements can be flagged [4].

2.2 Explainable Artificial Intelligence

Post-hoc local explanations approximate a trained model around an instance [23] or allocate the prediction among features with Shapley values [7]. TreeSHAP computes Shapley values for tree ensembles in polynomial time [8]. When features are dependent, the choice of value function matters: interventional (marginal) value functions evaluate the model on synthetic combinations of feature values, whereas conditional value functions respect the data distribution but

are harder to estimate [9]. Financial statement items are an extreme case of such dependence because many of them are linked by accounting identities.

An alternative line of research argues for models that are interpretable by design in high-stakes settings [24]. Generalised additive models with selected pairwise interactions (GA^{2}M) reach accuracy close to full-complexity models in many domains [25-26], and the Explainable Boosting Machine (EBM) implements this idea efficiently [27]. AGX shares the glass-box philosophy but differs in two respects: additivity is imposed over *accounting concepts* rather than individual features, so that rich non-linear interactions are retained *within* each concept, and directional accounting evidence is imposed through monotone constraints. Our propositions connect this structure to group-level Shapley values and to the Owen value for coalitional structures, which allows exact concept explanations without approximation.

3 PROBLEM FORMULATION

3.1 Firm-Year Panel, Episodes and Labels

Let i index firms and t fiscal years. For every firm-year we observe a feature vector $\mathbf{x}_{i,t} \in \mathbb{R}^p$ computed from the financial statements of year t and, where available, year $t-1$. The label $y_{i,t} \in \{0,1\}$ equals one if the statements of year t were subsequently identified as materially misstated in an AAER. Misstatements typically persist for several consecutive years; we call a maximal run of misstated years of the same firm that is covered by the same enforcement release an *episode* $e = (i, [s_e, c_e])$, with onset year s_e and last misstated year c_e. The onset indicator is $o_{i,t} = 1$ if $t = s_e$ for some episode of firm i, and the one-year-ahead onset indicator is $z_{i,t} = 1$ if $y_{i,t} = 0$ and $o_{i,t+1} = 1$.

3.2 Label Availability at Deployment Time

A model deployed at the end of fiscal year τ scores the statements of year τ, but AAERs about recent years have not yet been issued. Following the gap design of Bao et al. [4], the training sample for test year τ only contains firm-years $t \le \tau - 2$. In addition, an episode that is still running in year $\tau - 1$ or later cannot have been revealed at deployment time; its training firm-years are therefore relabelled as non-misstated:

$$\tilde{y}_{i,t} = y_{i,t}\,\mathbf{1}\left[c_{e(i,t)} < \tau - 1\right], \qquad t \le \tau - 2. \tag{1}$$

This rule is slightly stricter than the serial-fraud adjustment of Bao et al. [4], which relabels episodes that reach the test year τ; Section 6.6 shows that conclusions are unchanged under their rule.

3.3 Detection Tasks

We study three tasks that differ in how early a misstatement is flagged.

• Nowcasting (h=0). Rank all firm-years of test year τ by misstatement risk, $\hat{r}_{i,\tau} = F(\mathrm{x}_{i,\tau})$, and evaluate against $y_{i,\tau}$.

• Onset detection. Evaluate the same scores only on onset firm-years ($o_{i,\tau} = 1$) versus clean firm-years ($y_{i,\tau} = 0$), and track for every episode the first year in which the firm is flagged (episode detection delay).

• One-year-ahead onset warning (h=1). Using clean firm-years of year τ, rank firms by the risk that a misstatement episode starts in $\tau + 1$, evaluated against $z_{i,\tau}$. Training uses firm-years $t \le \tau - 3$, so that every training label refers to a year $t + 1 \le \tau - 2$.

3.4 Explanation Requirements

Let $\mathcal{I} = \{I_1, ..., I_m\}$ be a partition of the p features into m accounting concepts. For a scored firm-year we require concept attributions $\phi_g(\mathbf{x})$, $g = 1, ..., m$, that satisfy:

1. *Exactness and efficiency*: $\sum_g \phi_g(\mathbf{x}) = F(\mathbf{x}) - \mathbb{E}_D[F(\mathbf{X})]$ for a stated background distribution D, computed without sampling error.
2. *Faithfulness*: $\phi_g(\mathbf{x})$ equals the change in the risk score when concept g is replaced by background values.
3. *Coherence*: feature-level attributions within a concept add up to the concept attribution.
4. *Theory consistency*: a change of a feature in the direction that accounting evidence associates with higher risk never lowers the risk score or the corresponding concept attribution.
5. *Stability*: attributions of the same firm-year should not depend materially on random seeds or on the training window.

Post-hoc attributions of unconstrained ensembles generally violate (R2)–(R4); Section 4 constructs a model class that satisfies (R1)–(R4) by design, and Section 6 measures (R5).

4 THE AGX MODEL

Figure 1 summarises AGX. Financial statements of two consecutive years are mapped to 78 features organised into ten accounting concepts; a balanced ensemble of concept-partitioned, theory-monotone boosted trees produces a risk score whose concept decomposition is exact.

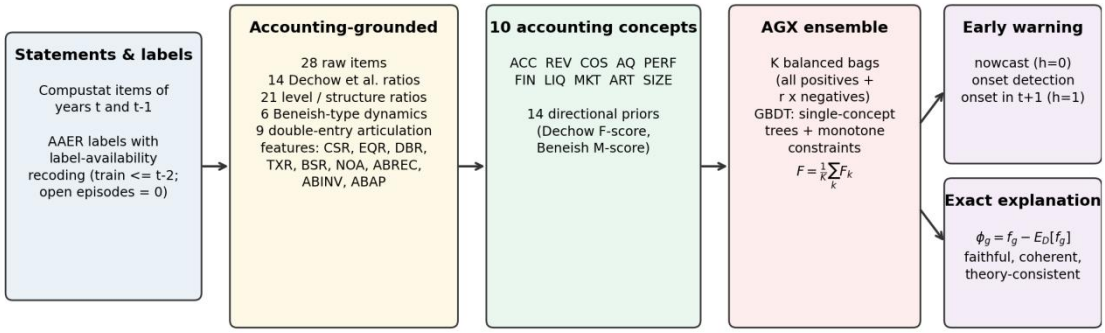


Figure 1 Overview of AGX

Note: Statements of years t and $t-1$ are mapped to raw items, Dechow-type ratios and accounting-grounded features (levels, Beneish-type dynamics and double-entry articulation residuals), partitioned into ten accounting concepts. Each of K balanced bags trains gradient-boosted trees that may only split on features of one concept and respect theory-based monotone constraints; the averaged log-odds score decomposes exactly into concept contributions.

4.1 Accounting-Grounded Representation

The input combines three blocks. (i) The 28 raw Compustat items and (ii) the 14 financial ratios of Dechow et al. supplied with the benchmark of Bao et al. [2, 4]. (iii) Thirty-six accounting-grounded (AG) features constructed from the statements of years t and $t-1$: 21 level and structure ratios (e.g., return on assets, gross margin, leverage, receivable and inventory turnover, financing raised), six Beneish-type dynamic indices [3] and nine double-entry articulation features introduced below. Let $\bar{A}_t = (AT_t + AT_{t-1})/2$ denote average total assets and Δ the year-over-year change.

Double-entry articulation residuals. Under double-entry bookkeeping, each balance sheet account rolls forward from its opening balance through flows that are reported elsewhere in the statements. When a roll-forward identity is evaluated with reported numbers, the residual captures unreported or irregular flows such as direct equity adjustments, restatements of opening balances or reclassifications [28]. We compute

$$\mathrm{CSR}_t = (\Delta RE_t - NI_t)/\bar{A}_t, \tag{2}$$

$$\mathrm{EQR}_t = (\Delta CEQ_t - NI_t - SSTK_t)/\bar{A}_t, \tag{3}$$

$$\mathrm{DBR}_t = (\Delta(DLTT_t + DLC_t) - DLTIS_t)/\bar{A}_t, \tag{4}$$

$$\mathrm{TXR}_t = (TXT_t - \Delta TXP_t)/\bar{A}_t, \tag{5}$$

$$\mathrm{BSR}_t = (AT_t - LT_t - CEQ_t - PSTK_t)/AT_t, \tag{6}$$

$$\mathrm{NOA}_t = [(AT_t - CHE_t) - (LT_t - DLC_t - DLTT_t)]/AT_{t-1}, \tag{7}$$

where CSR is the clean-surplus residual of retained earnings [29], EQR and DBR are the equity and debt roll-forward residuals, TXR relates tax expense to the change in taxes payable, BSR is the residual of the balance sheet identity and NOA measures net operating assets [30].

Abnormal working-capital growth. An overstatement of revenue must be booked against receivables (or cash), and an understatement of cost of goods sold against inventory. We therefore measure the growth of these accounts that is not explained by the firm's own prior turnover:

$$\mathrm{ABREC}_t = \frac{\Delta RECT_t}{\bar{A}_t} - \frac{RECT_{t-1}}{SALE_{t-1}} \cdot \frac{\Delta SALE_t}{\bar{A}_t}, \tag{8}$$

$$\mathrm{ABINV}_t = \frac{\Delta INVT_t}{\bar{A}_t} - \frac{INVT_{t-1}}{COGS_{t-1}} \cdot \frac{\Delta COGS_t}{\bar{A}_t}, \tag{9}$$

$$\mathrm{ABAP}_t = \frac{\Delta AP_t}{\bar{A}_t} - \frac{AP_{t-1}}{COGS_{t-1}} \cdot \frac{\Delta COGS_t}{\bar{A}_t}. \tag{10}$$

ABREC is positive when receivables grow faster than sales at the prior collection period, a classic signature of premature or fictitious revenue recognition; ABINV and ABAP play the same role for inventory and payables.

Beneish-type dynamics. We compute the days-sales-in-receivables index $\mathrm{DSRI}_t = (RECT_t/SALE_t)/(RECT_{t-1}/SALE_{t-1})$, the gross margin index $\mathrm{GMI}_t = GM_{t-1}/GM_t$ (positive margins only), the asset quality index $\mathrm{AQI}_t = SOFT_t/SOFT_{t-1}$ based on the soft-asset share, the sales growth index $\mathrm{SGI}_t = SALE_t/SALE_{t-1}$, the leverage

index LVGI_t and the change in the cash-to-assets ratio. Ratios with near-zero denominators are set to missing, which the tree learner handles natively.

Accounting concepts. Table 1 assigns every feature to exactly one of ten concepts that mirror how auditors and analysts organise fraud risk factors: accrual intensity, revenue and receivables, inventory, cost and margin, asset quality and capitalisation, profitability and performance, financing and leverage, liquidity and working capital, market valuation, double-entry articulation, and scale [2-3].

Table 1 Accounting Concepts and Their Features (Compustat Mnemonics for Raw Items; $p = 78$)

Concept	Features
ACC: Accrual intensity	dch_wc (WC accruals), ch_rsst (RSST accruals), NOA
REV: Revenue & receivables	dch_rec, ch_cs, DSRI, SGI, ABREC, rect/at, rect/sale, sale, rect
COS: Inventory, cost & margin	dch_inv, ch_cm, GMI, ABINV, invt/at, invt/cogs, gross margin, cogs, invt
AQ: Asset quality & capitalisation	soft_assets, AQI, dpi, ppegt/at, ivao/at, dp, ppegt, ivao
PERF: Profitability & performance	ch_roa, ch_fcf, ROA, EBIT, reoa, ib, ni, re, txt, txt/at
FIN: Financing & leverage	issue, (sstk+dltis)/at, sstk, dltis, leverage, LVGI, debt/at, xint/debt, xint, dltt, dlc, lt, pstk
LIQ: Liquidity & working capital	current ratio, che/at, Δ(che/at), ap/cogs, ABAP, act, lct, che, ivst, ap, txp
MKT: Market valuation	bm, E/P, S/P, log market value, prcc_f, csho
ART: Double-entry articulation	CSR, EQR, DBR, TXR, BSR
SIZE: Scale	log at, log sale, at, ceq

4.2 Concept-Partitioned, Theory-Monotone Boosting

Each ensemble member is a gradient-boosted tree model $F_k(\mathrm{x}) = b_k + \sum_{s=1}^{T} \eta\ h_{k,s}(\mathrm{x})$ trained on the logistic loss [31-32]. Two families of constraints encode accounting knowledge.

Concept interaction constraints. During tree growth, a feature may be used for a split only if it belongs to the same concept as every feature already used on the path from the root. Because every path contains the root split, each tree uses features of a single concept $g(s)$, and hence

$$F_k(\mathrm{x}) = b_k + \sum_{g=1}^{m} f_{k,g}\left(\mathrm{x}_{I_g}\right), \qquad f_{k,g} = \sum_{s:\, g(s)=g} \eta\ h_{k,s}. \tag{11}$$

Interactions *within* a concept (e.g., receivable growth conditional on sales growth) remain unrestricted, while interactions *across* concepts are excluded.

Theory-based monotone constraints. For the set $\mathcal{M}$ of fourteen features with established directional evidence we require F_k to be non-decreasing in $s_j x_j$, where $s_j \in \{-1, +1\}$ is the sign documented in the literature: WC and RSST accruals, change in receivables, change in inventory, soft assets, change in cash sales, actual issuance $(+)$ and change in ROA $(-)$ from the F-score model of Dechow et al. [2]; DSRI, GMI, AQI and SGI $(+)$ from the M-score of Beneish [3]; and ABREC and ABINV $(+)$, which formalise the same receivable and inventory mechanisms. Monotonicity is enforced split by split in the tree learner.

4.3 Balanced Bagging

Misstated firm-years represent less than 1% of the sample. Following the exploratory undersampling idea [33], AGX trains K members on all positive firm-years and an independent random sample of r times as many negative firm-years, and averages the members on the log-odds scale,

$$F(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} F_k(\mathbf{x}). \tag{12}$$

Averaging on the log-odds scale is essential: by Proposition 3 it preserves the concept-additive structure, whereas averaging probabilities would not. Rankings are unaffected by the intercept shift induced by undersampling.

4.4 Exact Concept Attribution

For a background distribution D (in our experiments, a random sample of training firm-years), the interventional concept game of firm-year $\mathbf{x}$ is

$$v_{\mathbf{x}}(S) = \mathbb{E}_{\mathbf{X}\sim D}\left[F\left(\mathbf{x}_{I_S}, \mathbf{X}_{\overline{I_S}}\right)\right], \qquad S \subseteq \{1, \ldots, m\}, \tag{13}$$

where $I_S = \cup_{g\in S} I_g$. The Shapley value of concept g in this game is the group Shapley value of the concept [34].

Proposition 1 (Exact concept attribution). *If* $F(\mathrm{x}) = b + \sum_g f_g\left(\mathrm{x}_{I_g}\right)$, *the Shapley value of concept* g *in the game (13) is* $\phi_g(\mathrm{x}) = f_g\left(\mathrm{x}_{I_g}\right) - \mathbb{E}_D\left[f_g\left(\mathrm{X}_{I_g}\right)\right]$, *(14) and* $\sum_g \phi_g(\mathrm{x}) = F(\mathrm{x}) - \mathbb{E}_D[F(\mathrm{X})]$.

Proof. Substituting the additive form into (13) gives $v_{\mathrm{x}}(S) = b + \sum_{g\in S} f_g\left(\mathrm{x}_{I_g}\right) + \sum_{g\notin S} \mathbb{E}_D f_g\left(\mathrm{X}_{I_g}\right)$. The marginal contribution of g to any coalition $S \not\ni g$ is therefore the constant $f_g\left(\mathrm{x}_{I_g}\right) - \mathbb{E}_D f_g\left(\mathrm{X}_{I_g}\right)$. Since the Shapley weights of all coalitions not containing g sum to one, ϕ_g equals this constant; efficiency follows by summation. □

The same argument shows that removing concept g (replacing x_{I_g} by background values) changes the expected score by exactly $\phi_g(\mathrm{x})$, so the attribution is faithful in the sense of (R2). The next result links concept attributions to feature-level explanations.

Proposition 2 (Hierarchical coherence). *Let $\psi_j(\mathrm{x})$ be the interventional Shapley value of feature j in the feature game $u_{\mathrm{x}}(T) = \mathbb{E}_D[F(\mathrm{x}_T, \mathrm{X}_{\bar{T}})]$. If F is concept-additive, then $\sum_{j\in I_g} \psi_j(\mathrm{x}) = \phi_g(\mathrm{x})$ for every concept g, and ψ coincides with the Owen value of the coalition structure $\mathcal{I}$.*

Proof. Concept additivity gives $u_{\mathrm{x}} = c + \sum_g u_{\mathrm{x}}^{(g)}$, where $u_{\mathbf{x}}^{(g)}(T) = \mathbb{E}_D f_g\left(\mathbf{x}_{T\cap I_g}, \mathbf{X}_{I_g\setminus T}\right)$ depends only on players in I_g. By linearity of the Shapley value and the null-player axiom, ψ_j for $j \in I_g$ equals the Shapley value of j in $u_{\mathbf{x}}^{(g)}$, and efficiency of that game yields $\sum_{j\in I_g} \psi_j = f_g\left(\mathbf{x}_{I_g}\right) - \mathbb{E}_D f_g = \phi_g$. The Owen value first splits the payoff among unions according to the quotient game, which here is the additive game (13), and then within each union according to the Shapley value of $u_{\mathbf{x}}^{(g)}$ [35]; both steps coincide with ψ. □

For a general model with cross-concept interactions, $\sum_{j\in I_g} \psi_j \neq \phi_g$ and the two explanations can rank concepts differently; Section 6.3 quantifies this gap for unconstrained boosting.

Proposition 3 (Closure under bagging and theory consistency). *(i) If $F_1, \ldots, F_K$ are concept-additive with respect to the same partition, then $F = \frac{1}{K}\sum_k F_k$ is concept-additive and $\phi_g^F = \frac{1}{K}\sum_k \phi_g^{F_k}$. (ii) If every F_k is non-decreasing in $s_j x_j$ for $j \in \mathcal{M} \cap I_g$, then F and ϕ_g are non-decreasing in $s_j x_j$, while ϕ_h for $h \neq g$ does not depend on x_j.*

Proof. (i) follows from linearity of (11) and of (14) in F. (ii) Monotonicity is preserved by averaging; ϕ_g differs from f_g only by a constant, and f_h with $h \neq g$ does not take x_j as an argument. □

Computation. Proposition 1 reduces explanation to evaluating each tree once: for every tree, the leaf value reached by x is credited to the tree's concept, and the background mean is subtracted. In practice we obtain the same quantity from TreeSHAP contributions [8], which for single-concept trees are confined to the tree's concept, and centre them on D. The cost is $O(KT\,\mathrm{depth})$ per firm-year, whereas the exact group Shapley value of a general model requires $2^m|D|$ model evaluations ($1{,}024|D|$ for $m = 10$).

Algorithm. For each test year τ: (1) build the 78 features for all firm-years; (2) form the training set $\{(\mathrm{x}_{i,t}, \tilde{y}_{i,t}): t \leq \tau - 2\}$ with relabelling (1); (3) for $k = 1, \ldots, K$, draw all positives and r times as many negatives, and fit boosted trees with concept interaction constraints and monotone constraints; (4) score year-τ firm-years with (12); (5) report concept attributions (14) and, within each concept, the centred feature contributions, which add up to the concept attribution by Proposition 2.

5 EXPERIMENTAL DESIGN

5.1 Data

We use the public benchmark released by Bao et al. [4], which links AAER-identified misstatements to Compustat fundamentals of U.S. listed firms. The file contains 146,045 firm-years from fiscal years 1990 to 2014 with the 28 raw items, the 14 ratios of Dechow et al. [2], the misstatement label and the AAER identifier that groups misstated firm-years. Fiscal year 1990 is used only to compute year-$(t-1)$ quantities. The sample contains 964 misstated firm-years associated with 412 AAER identifiers; grouping consecutive misstated years of the same firm and identifier yields 432 episode onsets, 289 of which are preceded by a clean firm-year of the same firm and therefore serve as positives for the one-year-ahead task. The misstatement rate is 0.66% overall and ranges from 0.07% to 1.27% across years. Table 2 reports training and test samples for each test year.

Table 2 Training and Test Samples per Test Year after the Label-Availability Recoding of Eq. (1)

Test year	Train firm-years	Train misstated	Test firm-years	Test misstated	Test onsets
2003	67,166	354	5,981	69	19
2004	73,233	427	5,934	58	17
2005	79,214	494	5,863	45	7
2006	85,148	557	5,908	33	12
2007	91,011	684	5,868	30	11
2008	96,919	734	5,612	26	13

Note: Onsets are first years of misstatement episodes.

5.2 Temporal Protocol

For each test year τ, every model is trained on firm-years up to $\tau - 2$ with the relabelling of Eq. (1) and evaluated on all firm-years of τ; the training window expands over time. All design choices and hyper-parameters were selected on a separate development period ($\tau \in \{1999, ..., 2002\}$) by the mean AUC across development years, with average precision as tie-breaker. The final evaluation uses the test period 2003–2008 of Bao et al. [4], and the years 2009–2011 serve as an additional out-of-period check in which labels are increasingly incomplete because of the AAER lag.

5.3 Baselines

We compare AGX with nine baselines that cover the accounting, machine-learning and interpretable-modelling literatures: (i) Logit-D14, a re-estimated logistic model on the 14 ratios of Dechow et al. [2]; (ii) Logit-Beneish on the indices of Beneish [3] (DSRI, GMI, AQI, SGI, depreciation index, LVGI and working-capital accruals); (iii) an L_2-regularised logit on all 78 features; (iv) RUSBoost-R28, the model of Bao et al. [4] on the 28 raw items with 300 trees, learning rate 0.1, minimum leaf size five and 1:1 random undersampling [36]; (v) a random forest; (vi) XGBoost [31]; (vii) LightGBM [32]; (viii) a multilayer perceptron (two hidden layers of 64 and 32 units, bagged over five undersampled sets); and (ix) the Explainable Boosting Machine (EBM), a GA ^{2}M with ten pairwise interactions [25, 27]. Except for Logit-D14, Logit-Beneish and RUSBoost-R28, all models receive the same 78 features as AGX, so that differences reflect model structure rather than information.

Hyper-parameters were tuned on the development period over the following grids: LightGBM leaves $\in \{7,15,31\}$; XGBoost depth $\in \{3,4,6\}$; random forest minimum leaf size $\in \{5,25\}$; MLP weight decay $\in \{10^{-4}, 10^{-3}, 10^{-2}\}$; logit inverse regularisation $C \in \{0.01,0.1,1\}$; and for AGX the tree size and number of rounds $\in \{(7,400), (7,600), (3,600)\}$, the negative-to-positive ratio $r \in \{5,10,20\}$ and an optional double weight on onset firm-years. The selected AGX uses $K = 10$ bags, $r = 5$, trees with seven leaves, 400 rounds and learning rate 0.03. On the development period it reached a mean AUC of 0.751, compared with 0.748 for the best XGBoost (depth four) and 0.735 for the best LightGBM (seven leaves). Weighting onset firm-years did not improve the development AUC and was therefore not used.

5.4 Evaluation Metrics

Ranking accuracy. Following Bao et al. [4], we report the area under the ROC curve (AUC) and, for the top $k = 1\%$ of test firm-years, NDCG@k, sensitivity and precision; we additionally report average precision (AP), which is informative under extreme imbalance. Metrics are computed per test year and averaged. Early detection. The onset AUC compares onset firm-years with clean firm-years of the same test year; onset recall@k is the share of onset firm-years ranked in the top $k\%$ of their year. For the 79 episodes that start between 2003 and 2008, we record whether the firm is flagged (top 5%) in its onset year and in any misstated year up to 2008, and the delay in years. For the one-year-ahead task we report AUC and recall@10%. Statistical inference. Differences in mean AUC and mean onset AUC between AGX and each baseline are assessed with a paired bootstrap that resamples firm-years within each test year (500 replications), reporting 95% confidence intervals and two-sided p-values. Explanation quality. For black-box models, concept attributions are obtained in the usual way by summing TreeSHAP contributions within each concept and centring them on the background sample [8]; for EBM, term contributions are summed within concepts, with pairwise terms split equally between their features. We measure (a) *exactness* as the maximum absolute difference between AGX attributions and brute-force group Shapley values computed over all 2^{10} coalitions; (b) *coherence* as the relative L_1 distance between aggregated feature attributions and exact group Shapley values for unconstrained models; (c) *removal fidelity* as the normalised L_1 error between $\phi_g(\mathbf{x})$ and the removal effect $\Delta_g(\mathbf{x}) = F(\mathbf{x}) - \mathbb{E}_D\left[F\left(\mathbf{x}_{-I_g}, \mathbf{X}_{I_g}\right)\right]$, together with their Spearman correlation and top-1 agreement; (d) *seed stability* as the mean pairwise Spearman correlation and top-1 agreement of attribution vectors across independently seeded retrainings (five for AGX, LightGBM and XGBoost; three for EBM) on a common set of the 300 highest-risk firm-years; (e) *temporal stability* as the mean pairwise Kendall τ between the global concept-importance rankings of the six annual models;

and (f) the *theory violation rate*, the share of firm-years for which moving a prior-constrained feature by one quarter of its inter-quartile range in the risk-increasing direction lowers the risk score. Exactness and coherence use the ten (AGX) or thirty (LightGBM, XGBoost) highest-ranked firm-years of 2005 and a background of 32 training firm-years; fidelity uses each model's 200 highest-ranked firm-years.

5.5 Implementation

All experiments were run in Python with LightGBM 4.7 (interaction and monotone constraints, intermediate monotone method), XGBoost 3.4, scikit-learn 1.8 [37], imbalanced-learn 0.14, InterpretML 0.7.8 and SHAP 0.54 on a single CPU core. Training one AGX model (ten bags) takes about 15 seconds per test year. Code, configuration files and all intermediate results are released with the paper.

6 RESULTS

6.1 Detection Accuracy

Table 3 reports ranking accuracy over the six test years and Table 4 the paired bootstrap comparisons. AGX attains a mean AUC of 0.713, which is statistically indistinguishable from XGBoost (difference $+0.000$, 95% CI $[-0.012,0.012]$), LightGBM (-0.003, $[-0.018,0.012]$), EBM ($+0.005$, $[-0.010,0.022]$) and RUSBoost-R28 (-0.005, $[-0.029,0.017]$). AGX significantly outperforms the random forest ($+0.020$, $p = 0.028$), the logit on all features ($+0.018$, $p = 0.044$) and the re-estimated Beneish model ($+0.144$, $p = 0.002$), and exceeds the Dechow logit by 0.030 ($p = 0.08$). Figure 2 shows that AGX is the most stable non-linear model across years: the standard deviation of its annual AUC is 0.019, compared with 0.022–0.035 for the other ensembles. XGBoost and LightGBM place slightly more misstatements in the top 1% (NDCG@1% of 0.030 and 0.028 versus 0.021); this gap closes as the inspection budget widens, and at a 5% cut-off AGX attains the highest sensitivity among all models (14.9%; Table 10).

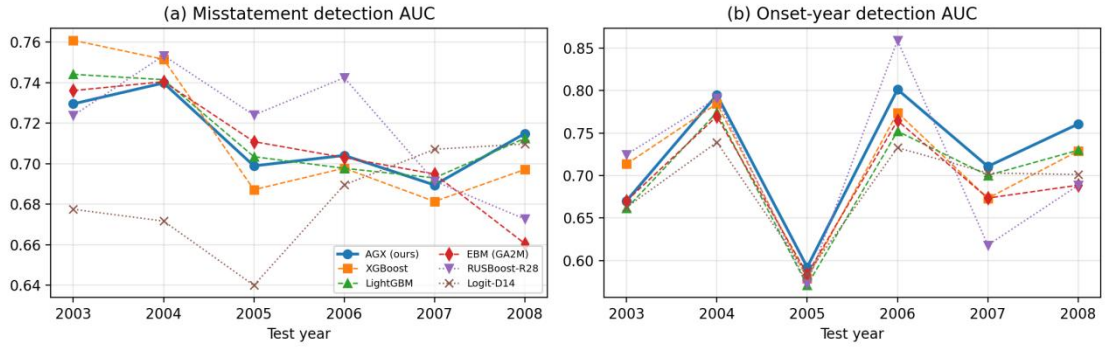


Figure 2 Annual AUC (a) and Onset-Year AUC (b) on the Test Years 2003–2008 for AGX and the Strongest Baselines

Table 3 Detection Performance on the Test Years 2003–2008 (Means over Years; Standard Deviation of the Annual AUC in Parentheses)

Model	AUC (s.d.)	AP	NDCG@1%	Sens@1%	Prec@1%	Onset AUC	Onset Rec@5%
AGX (ours)	0.713 (0.019)	0.019	0.021	2.4	2.0	0.722	10.1
XGBoost	0.713 (0.034)	**0.020**	**0.030**	**3.5**	**2.8**	0.709	12.7
LightGBM	0.715 (0.022)	0.019	0.028	3.3	2.3	0.698	**13.9**
EBM (GA^{2}M)	0.708 (0.029)	0.018	0.018	2.4	1.7	0.692	8.9
Random forest	0.693 (0.023)	0.016	0.009	1.0	1.1	**0.726**	12.7
MLP	0.702 (0.021)	0.017	0.017	1.8	1.4	0.684	6.3
RUSBoost-R28	**0.718 (0.031)**	0.017	0.012	1.4	1.4	0.709	11.4
Logit (all features)	0.694 (0.015)	0.016	0.021	2.2	2.2	0.685	7.6
Logit-D14	0.683 (0.026)	0.017	0.026	3.2	1.7	0.687	**13.9**
Logit-Beneish	0.569 (0.077)	0.009	0.005	0.8	0.6	0.630	11.4

Note: Sens@1%, Prec@1% and onset recall are percentages. Best values in bold.

Table 4 Paired Bootstrap Comparison of AGX with Each Baseline (500 Replications, Resampling Firm-Years within Test Years)

Comparison	ΔAUC [95% CI]	p	ΔOnset AUC [95% CI]	p
AGX vs. XGBoost	+0.000 [-0.012, +0.012]	0.992	+0.013 [-0.010, +0.034]	0.316
AGX vs. LightGBM	-0.003 [-0.018, +0.012]	0.752	+0.024 [+0.002, +0.046]	0.024
AGX vs. EBM (GA^{2}M)	+0.005 [-0.010, +0.022]	0.588	+0.030 [+0.002, +0.056]	0.032

Comparison	ΔAUC [95% CI]	p	ΔOnset AUC [95% CI]	p
AGX vs. Random forest	+0.020 [+0.003, +0.037]	0.028	-0.005 [-0.039, +0.028]	0.672
AGX vs. MLP	+0.011 [-0.009, +0.035]	0.288	+0.038 [-0.005, +0.078]	0.068
AGX vs. RUSBoost-R28	-0.005 [-0.029, +0.017]	0.696	+0.013 [-0.039, +0.057]	0.640
AGX vs. Logit (all features)	+0.018 [+0.001, +0.036]	0.044	+0.037 [+0.001, +0.070]	0.048
AGX vs. Logit-D14	+0.030 [-0.003, +0.060]	0.080	+0.035 [-0.034, +0.092]	0.276
AGX vs. Logit-Beneish	+0.144 [+0.097, +0.185]	0.002	+0.092 [+0.011, +0.166]	0.036

Note: Δ is the difference in mean AUC or mean onset AUC (AGX minus baseline).

6.2 Early Detection

Table 5 focuses on the onset of misstatement episodes. AGX achieves the highest onset AUC among the gradient-boosting, glass-box and logistic models (0.722), significantly above LightGBM ($+0.024$, $p = 0.024$), EBM ($+0.030$, $p = 0.032$) and the logit on all features ($+0.037$, $p = 0.048$), and above XGBoost and RUSBoost-R28 by 0.013. The random forest attains a similar onset AUC (0.726) but has the lowest overall AUC among the ensembles (0.693), so it trades general ranking accuracy for onset ranking, whereas AGX improves the latter without giving up the former. At a fixed 5% inspection budget, the share of the 79 episodes that are flagged in their onset year is similar across models (9–14%); AGX flags 19.0% of the episodes at some point before the end of the test window, the highest value jointly with LightGBM. The one-year-ahead task is markedly harder: with 54 future onsets, AGX reaches an AUC of 0.664 compared with 0.646 for XGBoost and 0.628 for LightGBM, and the parsimonious Dechow logit performs well (0.684), indicating that theory-based ratios carry most of the pre-onset signal; none of these differences is statistically significant.

Table 5 Early Detection on the Test Years 2003–2008

Model	Onset AUC	Onset Rec@1%	Onset Rec@5%	Episodes flagged	Delay (yrs)	$h = 1$ AUC	$h = 1$ Rec@10%
AGX (ours)	0.722	3.8	10.1	19.0	0.73	0.664	22.2
XGBoost	0.709	3.8	12.7	17.7	0.50	0.646	24.1
LightGBM	0.698	1.3	13.9	19.0	0.47	0.628	22.2
EBM (GA^{2}M)	0.692	2.5	8.9	15.2	0.58	–	–
RUSBoost-R28	0.709	0.0	11.4	17.7	0.50	0.625	18.5
Logit-D14	0.687	5.1	13.9	15.2	0.08	0.684	18.5

Note: Onset AUC and recall concern onset firm-years; onset Rec@5% equals the share of the 79 episodes starting in 2003–2008 that are flagged (top 5%) in their onset year; "Episodes flagged" is the share flagged in any misstated year up to 2008 and "Delay" the mean delay of detected episodes; $h = 1$ columns refer to one-year-ahead onset warning (54 future onsets). Percentages except AUC and delay.

6.3 Explanation Quality

Exactness and coherence. For the ten highest-ranked firm-years of 2005, the closed-form AGX attributions differ from brute-force group Shapley values over all 2^{10} coalitions by at most 8.9×10^{-16}, and they satisfy efficiency up to 6.7×10^{-16}, confirming Proposition 1. Summing interventional TreeSHAP values of the individual AGX features within concepts reproduces the concept attributions up to 1.4×10^{-8}, as predicted by Proposition 2. For unconstrained models this coherence fails: aggregating feature-level attributions within concepts deviates from the exact group Shapley values by 2.1% (LightGBM) and 2.3% (XGBoost) in relative L_1 distance with interventional TreeSHAP, and by 8.0% and 10.1% with the path-dependent TreeSHAP that is the default in practice; for XGBoost, the most important concept changes for 16.7% of the firm-years.

Fidelity and stability. Table 6 summarises the remaining criteria. Removing a concept from a firm-year changes the AGX score by exactly its attribution (normalised error 3.8×10^{-16}, rank correlation 1.000), whereas aggregated attributions of LightGBM, XGBoost and EBM deviate from the removal effects by 26–31% and identify the most influential concept correctly for only 64–91% of their highest-risk firm-years. Across independently seeded retrainings, AGX produces the most stable concept explanations (Spearman correlation 0.963 and top-1 agreement 97.4%, compared with 0.885–0.938 and 76–88% for the baselines) and the most stable feature-level explanations (0.955). Over the six annual models, the global ranking of concepts is also most stable for AGX (Kendall $\tau = 0.887$ versus 0.819–0.834).

Table 6 Explanation Quality

Model	Rem. error	Rem. ρ	Rem. top-1	Seed ρ_c	Seed top-1	Seed ρ_f	Temp. τ	Viol. (%)
AGX (ours)	0.000	1.000	100.0	0.963	97.4	0.955	0.887	0.0

Model	Rem. error	Rem. ρ	Rem. top-1	Seed ρ_c	Seed top-1	Seed ρ_f	Temp. τ	Viol. (%)
LightGBM	0.307	0.902	80.5	0.885	78.1	0.816	0.819	8.9
XGBoost	0.306	0.920	90.5	0.938	87.8	0.892	0.834	14.4
EBM (GA^{2}M)	0.263	0.827	64.0	0.924	76.1	0.935	0.822	30.2

Note: Rem. error, ρ and top-1 compare concept attributions with the effect of removing each concept (200 highest-risk firm-years of 2005); seed statistics (ρ_c concept-level, ρ_f feature-level Spearman correlation) use independently seeded retrainings; temporal τ compares global concept rankings across the six annual models; theory violations (Viol.) are averaged over the 14 prior-constrained features and the test years.

Theory consistency. Table 7 reports how often moving a prior-constrained feature in the risk-increasing direction lowers the score. AGX never does so, as guaranteed by Proposition 3. All baselines violate the accounting evidence for a substantial share of firm-years: on average 8.9% for LightGBM, 9.2% for the random forest, 14.4% for XGBoost, 30.2% for EBM and 31.5% for the MLP, with at least nine of the fourteen features affected by more than 5% violations in each of these models. Logistic models violate the priors whenever an estimated coefficient has the wrong sign, which happens for two of the eight Dechow and two of the five Beneish variables available to them.

Table 7 Violations of Accounting Theory Priors (Shares of Firm-Years for Which a Risk-Increasing Move of a Prior-Constrained Feature Lowers the Risk Score; Test Years 2003–2008)

Model	Features used	Features >5%	Mean rate (%)	Worst (%)
AGX (ours)	14	0	0.0	0.0
XGBoost	14	13	14.4	29.8
LightGBM	14	9	8.9	27.5
EBM (GA^{2}M)	14	12	30.2	69.4
Random forest	14	14	9.2	14.7
MLP	14	12	31.5	76.6
RUSBoost-R28	0	0	–	–
Logit (all features)	14	8	42.0	98.8
Logit-D14	8	2	20.0	96.6
Logit-Beneish	5	2	39.2	99.1

What the model relies on. Figure 3 shows the share of mean absolute concept attribution by year. Asset quality and capitalisation (28.5% on average) and financing and leverage (20.2%) dominate AGX's assessments, followed by profitability (12.0%), liquidity (10.4%), market valuation (8.5%) and revenue and receivables (8.0%); the ranking is consistent with the importance of soft assets and external financing in the F-score of Dechow et al. [2]. Figure 4 contrasts the partial dependence of six prior-constrained features. AGX's response curves are monotone in the documented direction, while LightGBM and XGBoost fit non-monotone shapes, for instance a decreasing risk over parts of the range of receivable growth and soft assets.

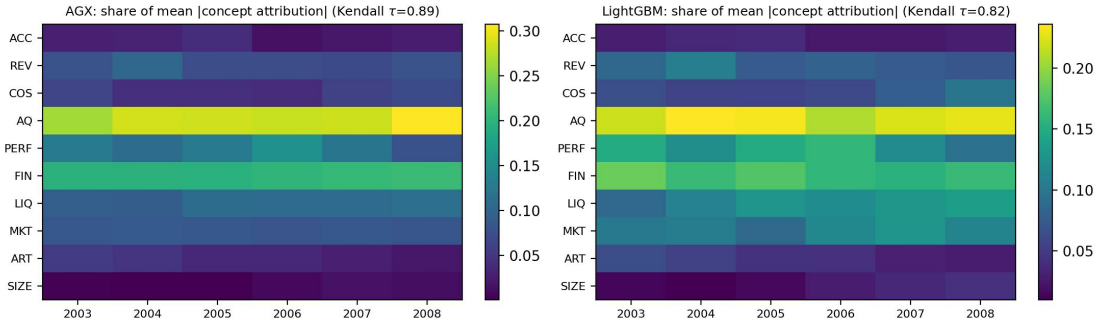


Figure 3 Share of Mean Absolute Concept Attribution per Test Year for AGX (Left) and LightGBM (Right), with the Mean Pairwise Kendall τ of the Yearly Concept Rankings

6.4 Ablation Study

Table 8 isolates the components of AGX. Removing the 36 accounting-grounded features causes the largest loss (AUC -0.029 to 0.684, onset AUC -0.012), confirming that the dynamic and structural information computed from two consecutive statements is essential. Removing the concept partition, or both constraint families, leaves the overall AUC essentially unchanged but lowers the onset AUC by 0.013–0.014, whereas removing only the monotone priors leaves both metrics unchanged and only forfeits the theory guarantees. The double-entry articulation features do not change

accuracy on their own (0.713 and 0.724), and their value lies in the articulation concept that they add to the explanations. Training a single model on all data instead of balanced bagging gives a comparable AUC but a lower NDCG@1% (0.018 versus 0.021) and onset AUC (0.718). Finally, two random partitions of the same 78 features into ten groups of the original sizes reach similar accuracy (AUC 0.716 and 0.720, onset AUC 0.731), showing that the onset gain stems from the additive structure itself; the accounting partition delivers the same accuracy while making every additive component an interpretable accounting argument (Figure 4).

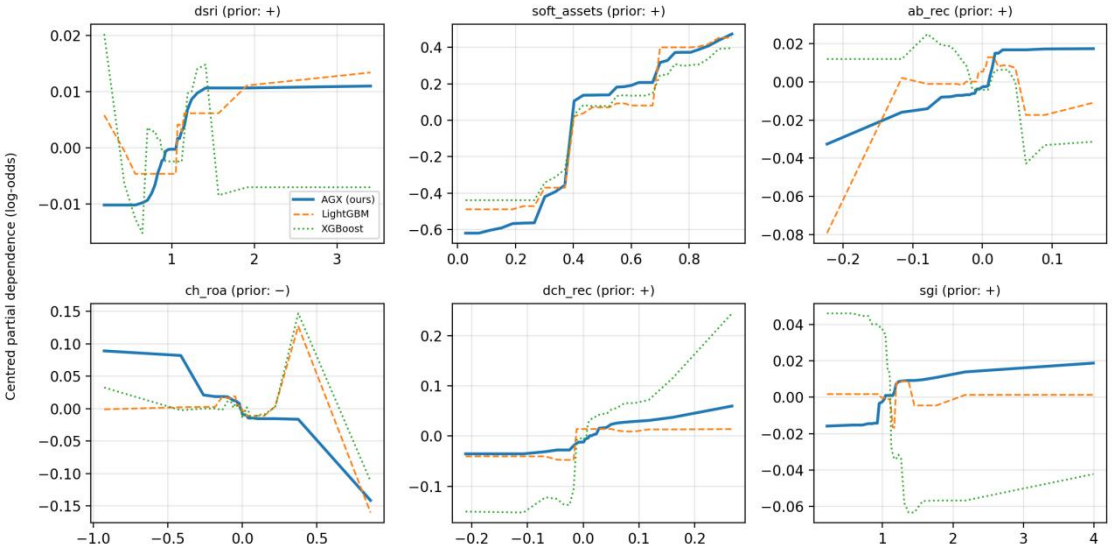


Figure 4 Centred Partial Dependence (Log-Odds) of Six Prior-Constrained Features for AGX, LightGBM and XGBoost (Test Year 2005 Models); the Expected Direction Is Given in Each Panel Title

Table 8 Ablation Study on the Test Years 2003–2008 (Means over Years)

Variant	AUC	AP	NDCG@1%	Onset AUC
AGX (full)	0.713	0.019	0.021	0.722
w/o accounting-grounded features	0.684	0.016	0.023	0.710
w/o articulation features	0.712	0.019	0.022	0.724
w/o monotone priors	0.714	0.019	0.020	0.722
w/o concept partition	0.714	0.020	0.021	0.708
w/o both constraints	0.713	0.021	0.022	0.709
w/o balanced bagging	0.716	0.018	0.018	0.718
random concept partition #1	0.716	0.019	0.019	0.731
random concept partition #2	0.720	0.019	0.021	0.731

6.5 Sensitivity Analysis

Table 9 and Figure 5 vary one hyper-parameter at a time. A single bag reduces both AUC (0.709) and onset AUC (0.709); five bags already recover most of the benefit and performance saturates at ten (0.713 and 0.722), with twenty bags giving the same values within 0.002. Doubling the negative-to-positive ratio from five to ten changes AUC by $+0.001$ and onset AUC by $+0.003$. Very small trees (three leaves) underfit (0.706), whereas larger trees (fifteen leaves) slightly improve both AUC (0.715) and onset AUC (0.729), indicating that richer interactions within concepts are beneficial and that the development-period choice is conservative. Table 10 shows that the relative position of AGX improves with the inspection budget.

Table 9 Sensitivity of AGX to the Number of Bags, the Negative-to-Positive Ratio and the Tree Size (Test Years 2003–2008)

Setting	AUC	AP	NDCG@1%	Onset AUC
n_bags = 1	0.709	0.019	0.023	0.709
n_bags = 5	0.713	0.019	0.020	0.720
n_bags = 10	0.713	0.019	0.021	0.722
n_bags = 20	0.712	0.019	0.023	0.720
neg_ratio = 5	0.713	0.019	0.021	0.722
neg_ratio = 10	0.714	0.019	0.020	0.725
num_leaves = 3	0.706	0.018	0.021	0.707
num_leaves = 7	0.713	0.019	0.021	0.722
num_leaves = 15	0.715	0.019	0.020	0.729

Note: The selected configuration is $K = 10$, $r = 5$ and seven leaves.

Table 10 NDCG@k / Sensitivity@k (%) for Alternative Inspection Budgets k (Test Years 2003–2008)

Model	k=1%	k=2%	k=3%	k=5%
AGX (ours)	0.021/2.4	0.039/5.3	0.058/9.0	0.084/14.9
XGBoost	0.030/3.5	0.050/7.0	0.071/11.2	0.087/14.5
LightGBM	0.028/3.3	0.046/6.5	0.066/10.6	0.086/14.5
EBM (GA^{2}M)	0.018/2.4	0.039/6.1	0.049/8.0	0.069/12.2
RUSBoost-R28	0.012/1.4	0.024/3.4	0.040/6.8	0.059/10.8

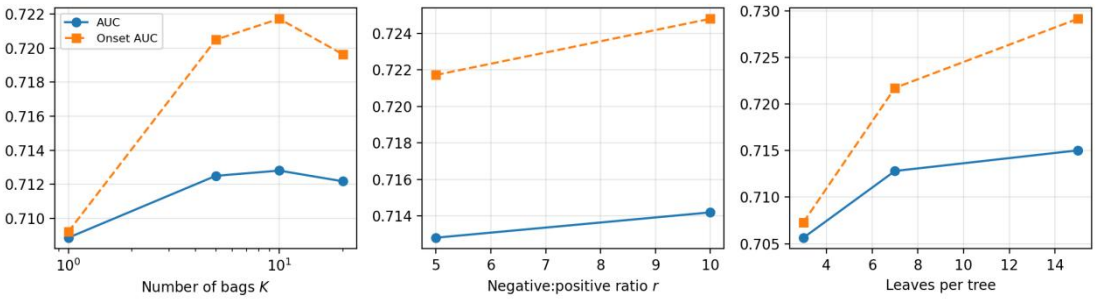


Figure 5 Sensitivity of AUC and Onset AUC to the Number of Bags K, the Negative-to-Positive Ratio r and the Number of Leaves per Tree

6.6 Robustness

Later period, measurement noise and labelling rule. Table 11 shows that AGX attains the highest AUC in the out-of-period years 2009–2011 (0.732 versus 0.726 for XGBoost and 0.719 for LightGBM), although XGBoost ranks the few onsets of that period better. When all reported amounts and ratios are perturbed by multiplicative Gaussian noise with standard deviation of 1%, 5% and 10% at deployment time, the AUC of AGX stays within 0.002 of its noise-free value, whereas XGBoost and EBM lose up to 0.007 (Figure 6). Under the serial-fraud rule of Bao et al. instead of Eq. (1), AGX obtains an AUC of 0.723 and an onset AUC of 0.727, compared with 0.715 and 0.700 for XGBoost (Table 12) [4].

Random seeds and missing history. Repeating the full AGX evaluation with three random seeds yields an AUC of 0.714 ± 0.001 and an onset AUC of 0.723 ± 0.001. For the 12.2% of test firm-years without prior-year statements, whose dynamic and articulation features are missing, AGX reaches an AUC of 0.743, compared with 0.676 for XGBoost, 0.675 for LightGBM and 0.695 for EBM, while all models perform similarly on firm-years with complete histories (0.707–0.718).

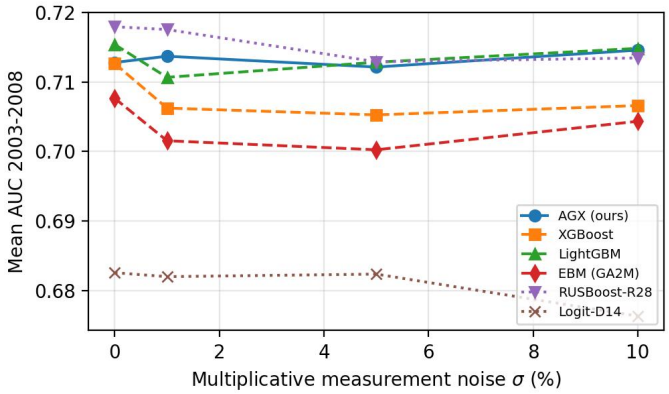


Figure 6 Mean AUC (2003–2008) under Multiplicative Measurement Noise Applied at Deployment Time

Table 11 Robustness: Out-of-Period Accuracy (2009–2011) and Accuracy under Multiplicative Measurement Noise of Standard Deviation σ Applied to All Reported Amounts at Deployment Time (Mean AUC 2003–2008)

Model	AUC 2009–11	Onset AUC 2009–11	AUC (σ=0)	σ=1%	σ=5%	σ=10%
AGX (ours)	0.732	0.701	0.713	0.714	0.712	0.715
XGBoost	0.725	0.716	0.713	0.706	0.705	0.707
LightGBM	0.719	0.691	0.715	0.711	0.713	0.715
RUSBoost-R28	0.702	0.661	0.718	0.718	0.713	0.713
Logit-D14	0.672	0.647	0.683	0.682	0.682	0.676

Table 12 Additional Robustness Checks on the Test Years 2003–2008

Check	AGX	Comparison
Serial-fraud rule of Bao et al. [4]: AUC / onset AUC	0.723 / 0.727	XGBoost: 0.715 / 0.700
Three random seeds: AUC (mean $\pm$ s.d.)	0.714 ± 0.001	–
Three random seeds: onset AUC (mean $\pm$ s.d.)	0.723 ± 0.001	–
Firm-years without prior-year data (12.2%): AUC	0.743	XGBoost 0.676; LightGBM 0.675; EBM 0.695
Firm-years with prior-year data: AUC	0.710	XGBoost 0.714; LightGBM 0.718; EBM 0.707

6.7 Case Studies

Figure 7 decomposes the AGX scores of the two highest-ranked onset firm-years of 2005 (firm identities are withheld). Firm A starts a three-year misstatement episode and is ranked at the 93.2th percentile; its score exceeds the background mean by 1.97 log-odds, of which asset quality contributes 0.98 (soft assets equal to 79% of total assets add 0.52 and a low share of gross property, plant and equipment adds 0.36), financing contributes 0.30 (mainly a USD 17.8 million equity issuance), and market valuation, liquidity and revenue add a further 0.73, while profitability lowers the risk by 0.11. Firm B, a single-year misstatement ranked at the 78.6th percentile, is flagged by asset quality (0.52), financing (0.37, including equity issuance and a leverage index of 2.84) and revenue and receivables (0.29, driven by receivables equal to 65% of sales), partly offset by profitability (-0.43). Because the decomposition is exact, these figures add up to the reported scores, and the within-concept feature contributions add up to each concept's contribution.

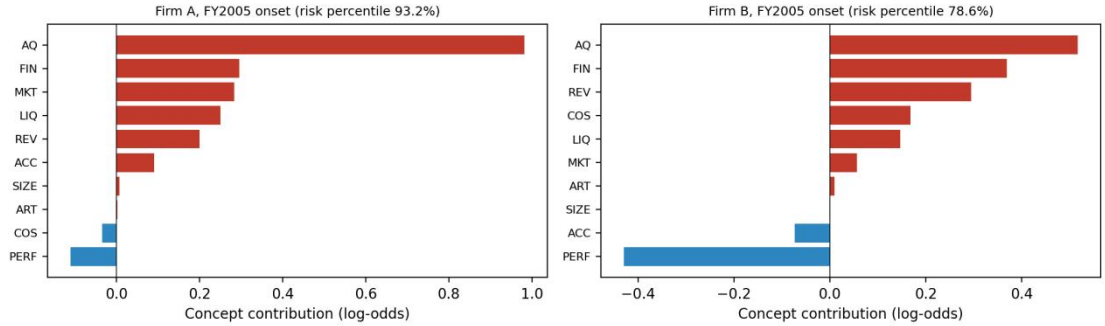


Figure 7 Exact Concept Decomposition of the AGX Risk Score for the Two Highest-Ranked Onset Firm-Years of 2005 (Log-Odds Relative to the Background Mean)

7 DISCUSSION AND CONCLUSION

Why concept additivity helps. Restricting trees to single concepts removes cross-concept interactions, which reduces the variance of a learner that sees only a few hundred positive firm-years per training window. The ablation shows that this restriction, rather than the particular grouping, is responsible for the improvement in onset detection, and the sensitivity analysis shows that the gain is obtained without sacrificing overall ranking accuracy. The same structure explains the robustness results: when the prior-year statements are missing, every lag-based feature is missing within its own concept, and the remaining concepts contribute exactly as they would for any other firm. Unconstrained ensembles, in contrast, route missing values through interactions learned mostly on firms with complete histories, which explains their markedly lower accuracy on the 12% of firm-years without prior-year data.

Implications for auditors, investors and regulators. AGX produces a risk score that can be read as a sum of accounting arguments. An auditor can see that a firm is flagged mainly because of its asset composition and its reliance on equity financing, drill down to the features that drive each concept, and rely on the fact that the drill-down adds up exactly. Because the monotone constraints encode published evidence, explanations never contradict that evidence, a property that unconstrained models violate for 9–32% of perturbations. Regulators can therefore use the model as a screening device whose reasoning can be documented and challenged in the vocabulary of financial statement analysis.

Scope and future research. The concept partition is a modelling choice that encodes accounting knowledge; it can be refined with additional constructs such as governance, audit or textual disclosure signals [5, 17], each of which would enter as a further concept with exactly the same guarantees. Conditional value functions and richer within-concept interactions are natural extensions [9], and the framework applies directly to other jurisdictions and enforcement datasets.

Conclusion. We introduced AGX, an accounting-grounded explainable model for early detection of financial misstatement risk. By combining double-entry articulation features, concept-partitioned and theory-monotone boosting and log-odds bagging, AGX attains the ranking accuracy of state-of-the-art ensembles on the AAER benchmark while improving onset-year detection, robustness and the stability of its explanations. Its concept attributions are exact interventional group Shapley values, coherent with feature-level attributions and consistent with accounting theory by

construction. These properties make machine-learning-based misstatement screening easier to audit, explain and act upon.

COMPETING INTERESTS

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

DATA AVAILABILITY

The data analysed in this study are publicly available at https://github.com/JarFraud/FraudDetection [4].

ACKNOWLEDGEMENTS

The authors thank Bao, Ke, Li, Yu and Zhang for making the AAER benchmark publicly available.

REFERENCES

[1] Karpoff J M, Koester A, Lee D S, et al. Proxies and databases in financial misconduct research. The Accounting Review, 2017, 92: 129-163. DOI: 10.2308/accr-51766.
[2] Dechow P M, Ge W, Larson C R, et al. Predicting material accounting misstatements. Contemporary Accounting Research, 2011, 28: 17-82.
[3] Beneish M D. The detection of earnings manipulation. Financial Analysts Journal, 1999, 55: 24-36.
[4] 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: 199-235. DOI: 10.1111/1475-679X.12292.
[5] Bertomeu J, Cheynel E, Floyd E, et al. Using machine learning to detect misstatements. Review of Accounting Studies, 2021, 26: 468-519. DOI: 10.1007/s11142-020-09563-8.
[6] Perols J L, Bowen R M, Zimmermann C, et al. Finding needles in a haystack: Using data analytics to improve fraud prediction. The Accounting Review, 2017, 92: 221-245. DOI: 10.2308/accr-51562.
[7] Lundberg S M, Lee S-I. A unified approach to interpreting model predictions. Advances in neural information processing systems 30, 2017: 4765-4774.
[8] 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: 56-67. DOI: 10.1038/s42256-019-0138-9.
[9] Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artificial Intelligence, 2021, 298: 103502. DOI: 10.1016/j.artint.2021.103502.
[10] Beneish M D, Lee C M C, Nichols D C. Earnings manipulation and expected returns. Financial Analysts Journal, 2013, 69: 57-82.
[11] Cecchini M, Aytug H, Koehler G J, et al. Detecting management fraud in public companies. Management Science, 2010, 56: 1146-1160. DOI: 10.1287/mnsc.1100.1174.
[12] Perols J. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. Auditing: A Journal of Practice & Theory, 2011, 30: 19-50.
[13] Abbasi A, Albrecht C, Vance A, et al. MetaFraud: A meta-learning framework for detecting financial fraud. MIS Quarterly, 2012, 36: 1293-1327.
[14] 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.
[15] Purda L, Skillicorn D. Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. Contemporary Accounting Research, 2015, 32: 1193-1223. DOI: 10.1111/1911-3846.12089.
[16] Hajek P, Henriques R. Mining corporate annual reports for intelligent detection of financial statement fraud – a comparative study of machine learning methods. Knowledge-Based Systems, 2017, 128: 139-152. DOI: 10.1016/j.knosys.2017.05.001.
[17] Brown N C, Crowley R M, Elliott W B. What are you saying? Using topic to detect financial misreporting. Journal of Accounting Research, 2020, 58: 237-291.
[18] Craja P, Kim A, Lessmann S. Deep learning for detecting financial statement fraud. Decision Support Systems, 2020, 139: 113421. DOI: 10.1016/j.dss.2020.113421.
[19] Amiram D, Bozanic Z, Rouen E. Financial statement errors: Evidence from the distributional properties of financial statement numbers. Review of Accounting Studies, 2015, 20: 1540-1593. DOI: 10.1007/s11142-015-9333-z.
[20] Lennox C, Lisowsky P, Pittman J. Tax aggressiveness and accounting fraud. Journal of Accounting Research, 2013, 51: 739-778. DOI: 10.1111/joar.12002.
[21] Achakzai M A K, Peng J. Detecting financial statement fraud using dynamic ensemble machine learning. International Review of Financial Analysis, 2023, 89: 102827. DOI: 10.1016/j.irfa.2023.102827.
[22] Zavitsanos E, Mavroeidis D, Bougiatiotis K, et al. Financial misstatement detection: A realistic evaluation. Proceedings of the 2nd ACM international conference on AI in finance (ICAIF'21), 2021: 34:1-34:9.

[23] Ribeiro M T, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD'16), 2016: 1135-1144.
[24] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 2019, 1: 206-215. DOI: 10.1038/s42256-019-0048-x.
[25] Lou Y, Caruana R, Gehrke J, et al. Accurate intelligible models with pairwise interactions. Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'13), 2013.
[26] Caruana R, Lou Y, Gehrke J, et al. Intelligible models for HealthCare: Predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining (KDD'15), 2015.
[27] Nori H, Jenkins S, Koch P, et al. InterpretML: A unified framework for machine learning interpretability. arXiv preprint arXiv:190909223, 2019.
[28] Hribar P, Collins D W. Errors in estimating accruals: Implications for empirical research. Journal of Accounting Research, 2002, 40: 105-134. DOI: 10.1111/1475-679X.00041.
[29] Ohlson J A. Earnings, book values, and dividends in equity valuation. Contemporary Accounting Research, 1995, 11: 661-687.
[30] Hirshleifer D, Hou K, Teoh S H, et al. Do investors overvalue firms with bloated balance sheets? Journal of Accounting and Economics, 2004, 38: 297-331. DOI: 10.1016/j.jacceco.2004.10.002.
[31] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD'16), 2016: 785-794.
[32] Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30, 2017: 3146-3154.
[33] Liu X-Y, Wu J, Zhou Z-H. Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2009, 39: 539-550.
[34] Shapley L S. A value for n-person games. In: Kuhn H W, Tucker A W, eds. Contributions to the theory of games II. Princeton University Press, 1953: 307-317.
[35] Owen G. Values of games with a priori unions. In: Henn R, Moeschlin O, eds. Mathematical economics and game theory: Essays in honor of oskar morgenstern. Springer, 1977: 76-88.
[36] 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: 185-197.
[37] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 2011, 12: 2825-2830.