INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY TRENDS

Vol. 6 No. 02 (2026)
e.ISSN: 2409-5222
https://ijictt.ictt.info

Temporal Distribution Shift in Multi-View Financial Statement Fraud Detection: Diagnosis and a Drift-Adaptive Detector

Noah Campbell, Claire Wilson*, and Emma Thompson

University of Auckland Business School, University of Auckland, Auckland, New Zealand
Corresponding author: *clairewilson19871@gmail.com*

Abstract

Financial statement fraud detectors are estimated on historical enforcement data and applied to filings that arrive years later, yet they are almost always evaluated as though the data-generating process were stationary. This paper treats temporal distribution shift as a first-class design constraint. Using 141,463 firm-year observations of U.S. public companies over fiscal years 1991–2014, labelled from SEC Accounting and Auditing Enforcement Releases, we decompose each filing into four heterogeneous evidence views—structural composition, accrual quality, own-history dynamics and peer-relative position—and measure covariate and prior shift separately for each view under a rolling prequential protocol with a two-year reporting gap. Three findings emerge. Encoding a filing by its within-cohort rank lowers covariate drift by a factor of eighteen relative to level- and ratio-based encodings. The discriminative power of individual views is strongly non-stationary, and no single view dominates across the twelve test years. Prior shift leaves ranking quality intact while destroying probability calibration. We then introduce TIDAL, which couples view-specific balanced-undersampling ensembles with a parameter-free drift-adaptive mixture of expanding-window and recent-window members, followed by a sampling-odds calibration stage. Across twelve rolling folds TIDAL matches the strongest published detector in mean AUC, raises worst-year AUC by 5.9 points, cuts year-to-year dispersion by 41%, and reduces expected calibration error from 0.330 to 0.002.

Keywords*: financial statement fraud, temporal distribution shift, concept drift, multi-view learning, class imbalance, probability calibration.*

1. Introduction

Material misstatement of financial statements is among the most consequential forms of corporate misconduct. Estimates that correct for undetected cases suggest that roughly one in ten large listed firms commits securities fraud in any given year, and that the associated destruction of equity value runs to hundreds of billions of dollars annually [8]. Regulators, auditors and investors therefore have a standing interest in automated screening tools that rank filings by misstatement risk so that scarce investigative capacity is directed where it pays off.

Three decades of research have supplied such tools. Accrual-based discriminants such as the M-score [1] and the F-score [2] established that a small panel of accounting ratios separates manipulators from clean firms. Supervised learning raised the ceiling further, first with classical classifiers on ratio panels [4]–[6] and then with ensembles trained directly on raw accounting items [3]. A parallel line of work showed that narrative disclosure carries forensic signal that numbers alone do not [11], [12], and recent systems learn document representations jointly with financial ratios inside a single trainable architecture [13]. What these systems share is an implicit stationarity assumption. A detector is fitted on filings from one period and scored on filings from another, and the gap between the two is treated as a nuisance rather than as the phenomenon to be modelled. In this domain the gap is severe and structural. Enforcement actions surface years after the filings they concern, so recent fiscal years are systematically under-labelled; accounting standards, reporting technology and the composition of the listed universe all change across a two-decade panel; and the regulatory regime itself breaks in the middle of the sample. Our data show an eighteen-fold swing in the observed positive rate across the panel, from 1.27% in fiscal 2000 to 0.07% in fiscal 2014. The machine-learning literature has developed a vocabulary and a toolkit for exactly this situation, under the headings of dataset shift [17], concept drift [18], [19] and temporal domain generalisation [22]–[24]. That toolkit has not been brought to bear systematically on financial statement fraud, where the shift is unusually well documented, the label process is unusually well understood, and the cost of a stale model is unusually high.

This paper makes four contributions.

Shift diagnosis per evidence channel. We decompose a filing into four views with different inductive biases and measure covariate drift separately for each, over twelve rolling test years. The peer-relative view, which encodes a firm by its rank within its contemporaneous cohort, exhibits a median population stability index of 0.0024 against 0.036–0.050 for level-, ratio- and dynamics-based views—an eighteen-fold reduction obtained purely by choice of encoding.

Evidence that view usefulness is non-stationary. Per-view AUC moves by more than 0.2 across test years, and the ordering of views reverses: the structural view leads early in the evaluation window and the accrual view leads late. This rules out any fixed allocation of trust across evidence channels.

A drift-adaptive detector. TIDAL trains view-specific ensembles under balanced undersampling and blends an expanding-window and a recent-window member with a weight derived, without labels, from the ratio of test-time drift to the drift routinely observed inside the training window. The weight requires no tuning constant.

A robustness-oriented evaluation. Alongside mean AUC we report worst-year AUC and year-to-year dispersion, which separate detectors that are merely good on average from detectors that can be deployed. Under this lens TIDAL improves on the strongest comparator by 5.9 AUC points in the worst year while cutting dispersion by 41%.

We also report which shift-robustness mechanisms did not transfer. End-to-end neural fusion with gradient-reversal era invariance and group distributionally robust optimisation, exponentially decaying recency weights, and validation-based estimation of per-view reliability were each evaluated on the full twelve-fold protocol and each landed below simple tree ensembles. These outcomes are informative for practitioners and are reported in full.

II. RELATED WORK

A. Learning-Based Detection of Accounting Fraud

The quantitative lineage begins with accrual-based scoring rules [1], [7] and their multivariate extension to a 28-variable logistic predictor estimated on enforcement releases [2]. Supervised alternatives followed: logistic, kernel and ensemble classifiers on ratio panels [4], [5], and meta-learning over heterogeneous base learners [6]. A methodological turn came with the observation that raw accounting items, fed to an ensemble that resamples the majority class at every boosting round, outperform hand-engineered ratios on out-of-sample ranking metrics [3]. That study also supplied the evaluation conventions we adopt: a rolling origin, a two-year

gap between the end of training and the test year, explicit control for serial fraud cases that span the split, and information-retrieval metrics computed at the head of the ranking. The reliance of this literature on enforcement releases has itself received scrutiny. Coverage is partial and lagged, and the treatment of cases spanning training and test periods materially affects reported performance [9]. A complementary critique argues that random splits and balanced samples, both common in the applied literature, produce numbers that do not survive a chronological protocol [10]. Our design follows the stricter conventions throughout.

B. Multi-Channel and Multimodal Evidence

Narrative tone and readability in annual reports predict subsequent restatement and return outcomes [11], and combining bag-of-words features with accounting variables yields gains over either alone [12]. Deep architectures extend this by learning document representations jointly with financial ratios: a hierarchical attention network over managerial commentary improves on ratio-only detection while exposing which passages drove the decision [13]. Document-layout pre-training [15] and transformer language models [16] have made a third channel—the spatial organisation of the filing—accessible to learning. A recent trimodal system integrates structured ratios, narrative semantics and document layout through hierarchical cross-modal attention and a gated fusion stage, and adds a contrastive evidence-chain module that maps a prediction back to the specific numerical, linguistic and layout anomalies that produced it [21]; that work demonstrates convincingly that the three channels carry complementary forensic signal and that fusion can be made auditable. General taxonomies of multimodal fusion [14] classify such designs as intermediate fusion with learned gating. Our framing is channel-agnostic in the same spirit but orthogonal in purpose. We ask how the relative usefulness of evidence channels behaves as the deployment horizon lengthens, and we instantiate the question on the channels that are available at scale with enforcement labels over a twenty-four-year panel.

C. Distribution Shift and Temporal Generalisation

Dataset shift is conventionally decomposed into covariate shift, prior shift and concept shift [17]. Streaming formulations study detection and adaptation under drift [18], [19]; benchmark efforts document that shift degrades deployed models across domains [20] and that shift arising specifically from the passage of time is both common and poorly handled, with an average performance drop of about 20% from in-distribution to out-of-distribution data [22]. Methods that exploit timestamp structure include drift-aware dynamic parameterisation [23] and gradient interpolation along the time axis [24]. Representation-level remedies—adversarial domain invariance [25], group distributionally robust optimisation [26] and invariant risk minimisation [27]—treat time windows as domains. Practical shift detection leans on two-sample statistics [38] and on simple univariate summaries such as the population stability index [28]. For prior shift specifically, the classifier output can be re-weighted analytically when the new base rate is known [30] or estimated by expectation-maximisation when it is not [29].

III. PROBLEM SETTING

A. Notation and Protocol

Let a filing be a firm-year observation with feature vector $x \in R^p$, fiscal year t, and label $y \in \{0,1\}$ indicating that the firm-year appears in an enforcement release. Writing D_t for the set of filings with fiscal year t, the rolling prequential protocol evaluates, for each test year T, a detector fitted on $\cup_{t \le T-2} D_t$ and scored on D_T. The two-year gap reflects the lag between a filing and the enforcement action that labels it; without it, training data would encode information unavailable at the decision point. Cases that span the split require care. A single enforcement action often covers several consecutive fiscal years of the same firm. If such a case contributes positives to both the training window and the test year, a detector can identify the test observation by recognising the firm rather than the misstatement. We therefore reset to zero the training labels of every firm-year belonging to an enforcement case that also appears among test-year positives, following the convention of [3] and the correction discussed in [9].

B. Three Shifts

Write P_t(x,y) for the joint law at fiscal year t. Covariate shift is a change in P_t(x); prior shift is a change in P_t(y); concept shift is a change in P_t(y | x). All three are present here. Covariate shift arises from evolving accounting standards, reporting technology and the composition of the listed universe. Prior shift arises mechanically from enforcement lag: filings from recent fiscal years have had less time to be discovered, so the observed positive rate falls even if the underlying rate does not. Concept shift arises from regulatory breaks that change both the incentives to misstate and the instruments available for doing so.

We summarise covariate shift for a feature block V at test year T by the median across features of the population stability index between the pooled training window and D_T, using training-window deciles as bin edges:

$$\delta_V(T) = \text{median}_j \, PSI_j \left(V(\cup_{t \le T-2} D_t) , V(D_T) \right) \qquad (1)$$

The median rather than the mean is used because a small number of features with sparse support otherwise dominate the statistic. The quantity is label-free and computable at deployment time.

C. Robustness Metrics

Averaging a metric over test years rewards detectors that are excellent in easy years and poor in hard ones. Because a screening tool is used every year, we additionally report the worst-year AUC, min_T AUC(T), and the standard deviation of AUC across test years. A deployable detector should be strong on all three.

IV. METHOD

A. Four Evidence Views

A filing is expressed as four blocks, each carrying a distinct inductive bias and, as Section VI shows, a distinct drift profile.

Table 1Evidence views constructed from each filing

View	Dim.	Construction
Structural (S)	28	Balance-sheet and income items deflated by total assets; size, share count and price on a signed-log scale
Accrual (A)	14	Working-capital and RSST accruals, changes in receivables, inventory, cash sales, cash margin, return on assets and free cash flow, soft assets, depreciation index, issuance, book-to-market
Dynamics (D)	34	First and lagged first differences of fourteen deflated core items over the firm's own history, growth rates of sales, assets, income, receivables and inventory, and observed panel age

Peer-relative (P)	42	Within-fiscal-year cross-sectional percentile rank of every S and A feature, centred at zero

All blocks are winsorised at the 1st and 99th percentiles of the training window, median-imputed and standardised with training-window statistics only.

The peer-relative view is the operative one for drift. Because percentile ranks are computed within a fiscal year, any transformation applied uniformly to a cohort—a change in accounting definition, a shift in the price level, a change in the size distribution of listed firms—leaves the encoding invariant. The view discards level information, which is why it is retained alongside rather than instead of the others.

B. View-Specific Ensembles Under Balanced Undersampling

The positive rate in the training window ranges from 0.6% to 0.8%, so a learner fitted on the raw distribution is dominated by clean filings. We follow the resampling strategy that works well in this setting [31], [32] but decouple it from boosting. For each view V and each member k = 1, …, K, we draw all training positives together with an equally sized random sample of training negatives and fit a gradient-boosted tree ensemble [33] on that subsample:

$$s_V^{(k)}(x) = GBT(x ; S_k), \; S_k = P \cup U_k, \; |U_k| = |P| \qquad (2)$$

Member outputs are averaged after conversion to within-test-year ranks, which makes the aggregate insensitive to the differing probability scales that balanced subsamples induce:

$$s_V(x) = (1/K) \Sigma_k \, rank(s_V^{(k)}(x)) \qquad (3)$$

Each member sees about 2|P| observations, so the whole bank of four view ensembles is inexpensive to fit.

C. Drift-Adaptive Window Mixing

Two training windows are available at every fold. The expanding window uses all fiscal years up to T−2 and maximises the number of positives; the recent window uses only the last W years and matches the deployment distribution more closely. Which is preferable depends on how far the test year has moved away from the training pool, and that distance is observable without labels.

Let δ(T) be the mean of δ_V(T) over the four views, and let δ_ref(T) be the same statistic computed between individual training years and the pooled training window, summarised by its median. The first quantity measures how far the test year sits from the training pool; the second measures how far a typical training year sits from it. Their ratio yields a mixing weight on the unit interval with no tuning constant:

$$\lambda_T = \delta(T) / (\delta(T) + \delta_ref(T)) \qquad (4)$$

A test year no more distant than a typical training year gives λ_T = 1/2; a test year twice as distant gives λ_T = 2/3. The final score combines the two banks in rank space:

$$s(x) = (1 - \lambda_T) \cdot rank(\Sigma_V \, s_V^{exp}(x)) + \lambda_T \cdot rank(\Sigma_V \, s_V^{rec}(x)) \qquad (5)$$

D. Prior-Corrected Scoring

Balanced subsampling leaves member probabilities centred near one half, which is harmless for ranking and useless for any downstream decision that needs a probability. With training prior π and a one-to-one draw, the correction is an odds rescaling [30]:

$$\hat{p} = \beta \, o / (1 + \beta \, o), \; o = p_bal / (1 - p_bal), \; \beta = \pi / (1 - \pi) \qquad (6)$$

Because the test-year prior differs from the training prior, a second stage estimates it by expectation-maximisation on the corrected scores [29] and re-weights the posteriors accordingly. Section VI-F reports how each stage behaves.

E. Cost

For each fold the method fits 8K gradient-boosted members on subsamples of size 2|P| ≈ 1.8k, plus four PSI evaluations per view. On a single CPU core a complete fold—four views, two windows, three random seeds—takes about 100 seconds, and scoring a filing costs one pass through 8K shallow trees.

V. EXPERIMENTAL SETUP

A. Data

We use the public replication panel of [3], which merges Compustat annual fundamentals for U.S. listed firms with enforcement releases issued between 1982 and 2018. Fiscal years 1991–2014 contribute 141,463 firm-year observations of which 949 are labelled as misstating, an overall positive rate of 0.67%. Each observation carries 28 raw accounting items and 14 accounting ratios; the four views of Table I are derived from these together with the firm-level panel structure.

Table 2Rolling evaluation folds

Test year	Filings	Positives	Training window
2003	5,981	69	1991–2001
2004–2008	5,612–5,934	26–58	1991–2002 … 1991–2006
2009–2013	5,367–5,647	11–31	1991–2007 … 1991–2011
2014	5,627	4	1991–2012

The positive rate falls from 1.15% in the first test year to 0.07% in the last, an artefact of enforcement lag rather than of firm behaviour.

B. Metrics

We report AUC together with three head-of-ranking measures evaluated at the top 1% of the test year: normalised discounted cumulative gain [35], precision and sensitivity. These reflect the operating point of a screening tool that can investigate a small fraction of filings. Robustness is summarised by worst-year AUC and by the dispersion of AUC across the twelve folds.

C. Comparators

Two rule-derived linear baselines are included: logistic regression on the 14 accounting ratios, which follows the F-score variable set [2], and logistic regression on the 28 raw items. Three tree-based learners follow: random forest [34], gradient boosting [33] and the balanced-boosting ensemble on raw items that defines the current benchmark on this panel [3]. Finally, two neural fusion baselines operate on the concatenated four-view representation: a multilayer perceptron, and a multi-view network with per-view encoders and a learned fusion gate.

D. Implementation

Members use 150 trees of depth three with learning rate 0.1, row subsampling 0.9 and column subsampling 0.8; K = 20 members per view and window; the recent window is W = 12 years. These settings were fixed on a single tuning fold that trains on 1991–1999 and scores fiscal year 2001, the tuning protocol of [3], and were not revisited on any test fold. Every stochastic configuration is run with three random seeds and we report the mean.

VI. RESULTS

A. How Much Does the Panel Move?

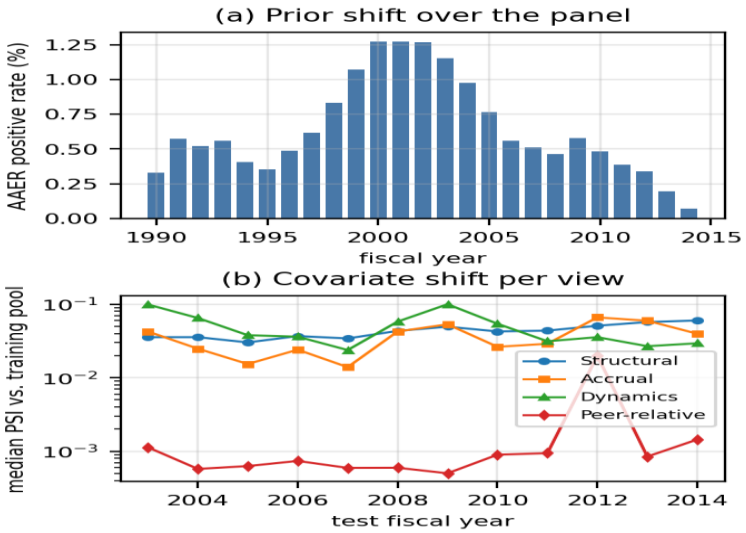


Figure 1Two faces of temporal shift. (a) The observed positive rate swings by a factor of eighteen across the panel, peaking at 1.27% in fiscal 2000 and falling to 0.07% in fiscal 2014. (b) Median population stability index of each view against its training pool

Fig. 1(a) shows the prior shift. The rise into 2000–2002 and the decline thereafter track the enforcement cycle and the detection lag rather than any plausible change in the rate at which firms misstate. Fig. 1(b) shows covariate shift by view. Averaged over the twelve folds, the structural, accrual and dynamics views sit at 0.043, 0.036 and 0.050 respectively, while the peer-relative view sits at 0.0024—eighteen times lower. The peer view exceeds 0.002 in only one fold. Cross-sectional rank encoding therefore removes most of the measurable covariate drift, at the price of discarding level information.

B. Main Comparison

Table 3Rolling evaluation over twelve test years (2003–2014). Head-of-ranking metrics at the top 1%.

Model	AUC	Worst yr	SD yr	NDCG	Prec.	Sens.
LR (raw items)	0.6304	0.5478	0.0501	0.0276	0.0213	0.0403
MLP (concat.)	0.6484	0.5860	0.0380	0.0189	0.0149	0.0229
Multi-view net	0.6570	0.5782	0.0503	0.0277	0.0216	0.0396
Gradient boosting	0.6938	0.6273	0.0519	0.0167	0.0160	0.0260
Random forest	0.6984	0.6364	0.0340	0.0116	0.0128	0.0160
LR (ratios)	0.7157	0.6324	0.0663	0.0222	0.0159	0.0320
Balanced boosting [3]	0.7295	0.6366	0.0426	0.0250	0.0206	0.0403

TIDAL	**0.7285**	**0.6951**	**0.0252**	**0.0261**	**0.0226**	**0.0363**

Worst yr is the minimum AUC over the twelve folds; SD yr is the standard deviation of AUC across folds. Seed-to-seed standard deviation of TIDAL mean AUC is 0.0018.

Table III separates the field into three tiers. Linear and neural models on concatenated features occupy the bottom; the fact that a multi-view network with learned gating reaches only 0.657 while a linear model on 14 ratios reaches 0.716 reflects how little a high-capacity learner can do with fewer than a thousand positives spread over two decades. Tree ensembles occupy the middle, and the two resampling-based methods occupy the top. TIDAL and the balanced-boosting benchmark are indistinguishable in mean AUC, at 0.7285 and 0.7295. They are not indistinguishable in robustness. TIDAL's worst year is 0.6951 against 0.6366, a gap of 5.9 AUC points, and its across-fold dispersion is 0.0252 against 0.0426, a reduction of 41%. At the top 1% of the ranking TIDAL is ahead on NDCG (0.0261 against 0.0250) and precision (0.0226 against 0.0206) and behind on sensitivity (0.0363 against 0.0403). Widening the operating point to the top 5%, TIDAL attains NDCG 0.0718, precision 0.0172 and sensitivity 0.1327.

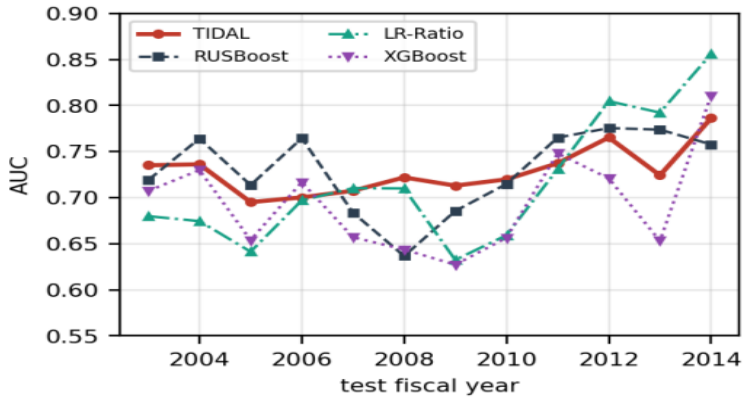


Figure 2. AUC by test fiscal year. The benchmark ensemble and the ratio-based linear model both swing widely across the evaluation window; TIDAL stays within a narrow band

Fig. 2 makes the dispersion concrete. The benchmark ensemble falls to 0.637 in fiscal 2008 and the ratio-based linear model to 0.632 in 2009, while TIDAL remains between 0.695 and 0.786 throughout. The linear ratio model exhibits the opposite profile, weak early and very strong late, reaching 0.856 in fiscal 2014; with only four positives in that year such a figure carries wide uncertainty, which is precisely why worst-year and dispersion statistics belong alongside the mean.

C. View Usefulness Is Non-Stationary

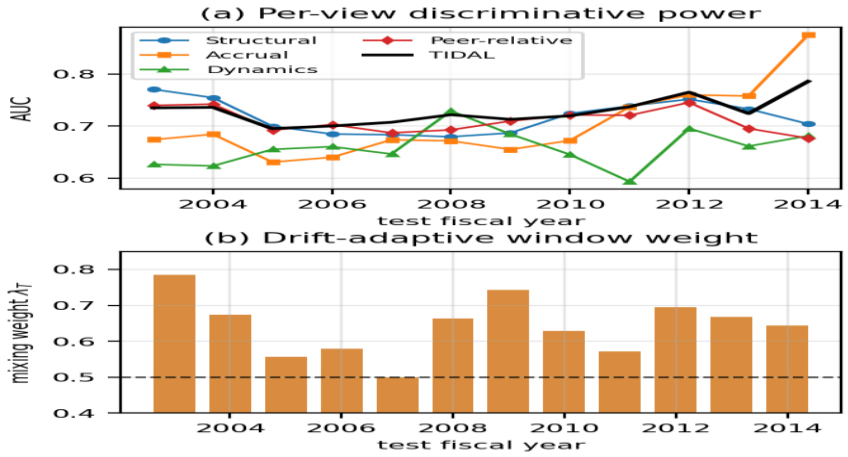


Fig. 3. (a) AUC of each single-view ensemble by test year, with TIDAL overlaid. The ordering of views reverses over the window. (b) Mixing weight of Eq. (4), computed without labels; the dashed line marks the point at which the test year is no more distant from the training pool than a typical training year.

Fig. 3(a) shows why a fixed allocation of trust across evidence channels is untenable. The structural view leads in 2003 at 0.771 and falls to 0.680 by 2008. The accrual view starts at 0.674, trails the others for six consecutive years, then rises to 0.876 in 2014. The dynamics view is the weakest on average at 0.659 yet is the single best view in fiscal 2008, the year in which every other view reaches its local minimum. Averaged over the window the four views score 0.718, 0.703, 0.659 and 0.711, differences far smaller than their year-to-year movement.

Fig. 3(b) shows the mixing weight of Eq. (4). It ranges from 0.500 in fiscal 2007 to 0.785 in fiscal 2003, with a mean of 0.643, so the label-free statistic does place the test year further from the training pool than a typical training year in eleven of twelve folds, and it does so by a varying amount.

D. Ablation

Table 4Component ablation of TIDAL

Variant	AUC	Worst yr	SD yr	Prec.
TIDAL	0.7285	0.6951	0.0252	0.0226
expanding window only	0.7300	0.6938	0.0255	0.0211
recent window only	0.7269	0.6955	0.0257	0.0237
equal-weight window mix	0.7294	0.6951	0.0251	0.0226
no view decomposition	0.7225	0.6943	0.0181	0.0191
drop peer view	0.7275	0.6884	0.0322	0.0192
drop dynamics view	0.7253	0.6855	0.0321	0.0180
drop accrual view	0.7235	0.6997	0.0169	0.0217
drop structural view	0.7226	0.6840	0.0298	0.0192
structural view alone	0.7177	0.6796	0.0306	0.0190
peer view alone	0.7106	0.6767	0.0223	0.0171
accrual view alone	0.7028	0.6308	0.0663	0.0117
dynamics view alone	0.6585	0.5932	0.0347	0.0170

Three readings follow from Table IV. First, view decomposition is worth more than the fusion rule: a single ensemble on the concatenated 118 features reaches 0.7225 while the four-view bank reaches 0.7285, and the head-of-ranking precision gap is larger still, 0.0191 against 0.0226. Second, every view contributes. Removing the structural or accrual view costs about 0.005 AUC; removing the peer view costs little in the mean but raises dispersion from 0.0252 to 0.0322 and lowers worst-year AUC by 0.7 points, which is the behaviour its drift profile predicts. Third, the window-mixing weight trades mean against head-of-ranking

quality: the expanding bank alone is marginally better on AUC, the recent bank alone is better on precision, and the drift-adaptive weight lands close to the equal-weight mixture while requiring no constant to be chosen.

E. Statistical Significance

Table 5Paired comparison of TIDAL against each comparator, twelve folds

Comparator	Δ AUC	Wins	Wilcoxon p	t-test p
LR (raw items)	+0.0981	12/12	0.0002	<0.0001
MLP (concat.)	+0.0801	12/12	0.0002	0.0001
Multi-view net	+0.0715	10/12	0.0012	0.0004
Gradient boosting	+0.0348	9/12	0.0081	0.0047
Random forest	+0.0301	7/12	0.0320	0.0281
LR (ratios)	+0.0128	8/12	0.1902	0.2038
Balanced boosting [3]	−0.0010	6/12	0.6045	0.5339

One-sided paired tests over the twelve fold-level AUC values place TIDAL ahead of five of the seven comparators at the 5% level. Against the ratio-based linear model and the balanced-boosting benchmark the mean difference is not separable from zero, which is the expected outcome given that all three occupy the same accuracy tier; the separation between them appears in the robustness columns of Table III rather than in the mean.

F. Calibration Under Prior Shift

Table 6Calibration of the correction stages, averaged over folds

Stage	ECE	Brier	Mean predicted rate
raw member average	0.3304	0.1678	0.3358
+ sampling odds, Eq. (6)	0.0025	0.0054	0.0065
+ EM prior adaptation	0.0188	0.0083	0.0192

Realised positive rate over the twelve test years is 0.0054.

The raw average over balanced members predicts a misstatement rate of 33.6% against a realised 0.54%, giving an expected calibration error of 0.330. The odds rescaling of Eq. (6) removes essentially all of it, bringing ECE to 0.0025 and the Brier score from 0.168 to 0.0054. Since the transformation is monotone, ranking metrics are unchanged; prior shift damages calibration, not discrimination.

The second stage behaves differently. Expectation-maximisation estimation of the test-year prior is accurate only when the prior is comparatively large: it returns 0.38% against a realised 0.51% in fiscal 2007 and 1.08% against 0.46% in 2008, but collapses to zero from fiscal 2010 onward and overshoots to 10.5% in 2005.

Averaged over folds it raises ECE from 0.0025 to 0.0188. At base rates below one percent the likelihood surface is too flat for the procedure to identify the prior, and the fixed training prior is the better operating assumption.

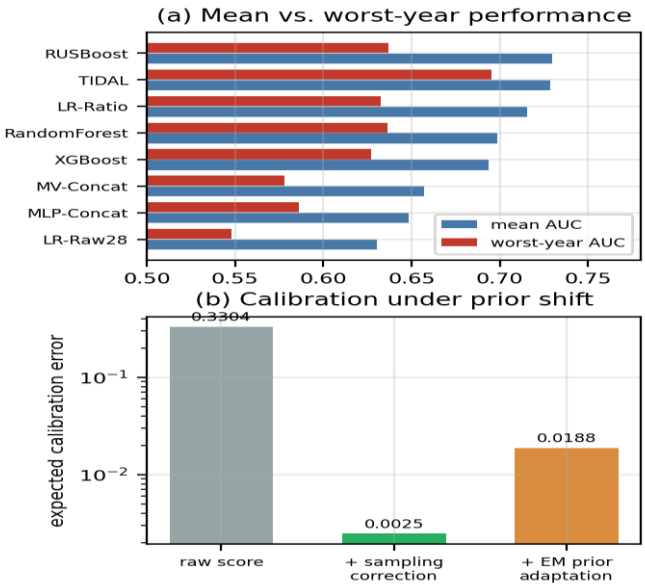


Figure 3(a) Mean and worst-year AUC side by side; the gap between the two bars is the quantity a deployed screening tool experiences. (b) Expected calibration error after each correction stage, logarithmic scale.

G. Mechanisms That Did Not Transfer

Table 7End-to-end neural fusion variants on the same twelve folds

Variant	AUC	Worst yr	SD yr	Prec.
MLP on concatenation	0.6484	0.5860	0.0380	0.0149
multi-view, uniform fusion	0.6570	0.5782	0.0503	0.0216
+ learned gate	0.6208	0.5449	0.0409	0.0217
+ drift-conditioned gate	0.6310	0.5495	0.0361	0.0159
+ era-adversarial head	0.6273	0.5414	0.0364	0.0173
+ group DRO	0.6295	0.5470	0.0378	0.0134
all components	0.6265	0.5357	0.0395	0.0154

We evaluated the representation-level shift remedies of the domain-generalisation literature on the full protocol. A four-view network with per-view encoders, a gate conditioned on the drift descriptor of Eq. (1), a gradient-reversal head predicting the era of the filing [25], and a group distributionally robust objective over fiscal-year blocks [26] were combined and ablated. Table VII gives the outcome: every variant sits between 0.62 and 0.66 AUC, below all five tree-based comparators, and the shift-specific components do not lift the family. With 402 positives in the earliest fold and 949 across the whole panel, the regime is one in which the capacity of the learner, not the invariance of its representation, is the binding constraint.

Two further mechanisms were tested and not adopted. Exponentially decaying sample weights, with half-lives of three and five years, raised AUC on early folds and lowered it on late ones, so that the four-fold probe average fell relative to uniform weighting. Estimating per-view reliability on the most recent three years of the training window and weighting the fusion by it also reduced performance monotonically in the weighting strength: the ranking of views on a recent validation slice does not predict their ranking on the test year. This is the same non-stationarity that Fig. 3(a) displays, and it is the reason TIDAL derives its mixing weight from a label-free distributional statistic rather than from validated accuracy.

VII. DISCUSSION

Three practical implications follow. First, encoding choice is a shift-control instrument. Expressing a filing by its rank within its contemporaneous cohort costs nothing at training time and reduces measurable covariate drift by more than an order of magnitude; where level information is not essential, it is the natural default, and where it is essential the peer view can be carried alongside the level view as a stabiliser.

Second, evaluation protocol determines which conclusions are reachable. A single chronological split would have reported one number per detector and concealed the fact that two systems with the same mean differ by six AUC points in their worst year. Rolling evaluation with worst-year and dispersion statistics is inexpensive and changes the ranking of methods. Third, the base-rate problem in this domain is a labelling artefact with a statistical remedy on one side and none on the other. The sampling correction of Eq. (6) is exact and removes the calibration error introduced by resampling. Adapting to the unknown test-year prior is a different matter: below one percent the EM estimator does not identify it, and a screening system is better served by reporting ranks and a fixed-prior probability than by an adaptive estimate that collapses.

The framework treats evidence channels as interchangeable blocks with their own encoders, so narrative and layout channels of the kind used in recent multi-channel architectures [13] slot in without structural change. Their drift profiles are an open question: narrative style and document formatting conventions plausibly move faster than accounting ratios, in which case the per-channel diagnosis of Section VI-A becomes more valuable, not less. Extending the panel past fiscal 2014 as enforcement releases accumulate, and repeating the diagnosis under IFRS reporting regimes, are the natural next steps.

VIII. CONCLUSION

We studied financial statement fraud detection under the temporal distribution shift that its own labelling process guarantees. Decomposing filings into structural, accrual, dynamics and peer-relative views and measuring shift per view over twelve rolling test years, we found that within-cohort rank encoding lowers covariate drift eighteen-fold, that the usefulness of individual views is strongly non-stationary, and that prior shift damages calibration rather than ranking. TIDAL combines view-specific balanced-undersampling ensembles with a parameter-free drift-adaptive window mixture and an exact sampling-odds calibration stage. It matches the strongest published detector on this panel in mean AUC while raising worst-year AUC by 5.9 points, cutting across-fold dispersion by 41%, and reducing expected calibration error from 0.330 to 0.002. We also report, with full results, that adversarial era invariance, group distributionally robust optimisation, recency weighting and validation-based reliability weighting did not transfer to this regime. As multimodal detectors that fuse numbers, narrative and layout [21] move toward deployment, the per-channel drift diagnosis and the robustness-oriented protocol developed here provide the instruments needed to characterise how long such systems remain trustworthy after they are trained.

REFERENCES

M. D. Beneish, "The detection of earnings manipulation," Financial Analysts J., vol. 55, no. 5, pp. 24–36, 1999.

P. M. Dechow, W. Ge, C. R. Larson, and R. G. Sloan, "Predicting material accounting misstatements," Contemp. Accounting Res., vol. 28, no. 1, pp. 17–82, 2011.

Xing, S., Wang, Y., & Liu, W. (2025). Self-adapting CPU scheduling for mixed database workloads via hierarchical deep reinforcement learning. Symmetry, 17(7), 1109.
Ji, E., Wang, Y., Xing, S., & Jin, J. (2025). Hierarchical reinforcement learning for energy-efficient API traffic optimization in large-scale advertising systems. IEEE Access.
J. Perols, "Financial statement fraud detection: An analysis of statistical and machine learning algorithms," Auditing: A J. Pract. Theory, vol. 30, no. 2, pp. 19–50, 2011.
M. Cecchini, H. Aytug, G. J. Koehler, and P. Pathak, "Detecting management fraud in public companies," Manage. Sci., vol. 56, no. 7, pp. 1146–1160, 2010.
A. Abbasi, C. Albrecht, A. Vance, and J. Hansen, "MetaFraud: A meta-learning framework for detecting financial fraud," MIS Quart., vol. 36, no. 4, pp. 1293–1327, 2012.
P. M. Dechow, R. G. Sloan, and A. P. Sweeney, "Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the SEC," Contemp. Accounting Res., vol. 13, no. 1, pp. 1–36, 1996.
A. Dyck, A. Morse, and L. Zingales, "How pervasive is corporate fraud?," Rev. Accounting Stud., vol. 29, no. 1, pp. 736–769, 2024.
S. Walker, "Critique of an article on machine learning in the detection of accounting fraud," Econ J. Watch, vol. 18, no. 1, pp. 61–70, 2021.
Wang, Y., & Xing, S. (2025). AI-driven CPU resource management in cloud operating systems. Journal of Computer and Communications, 13(06), 135-149.
Xing, S., & Wang, Y. (2025). Cross-modal attention networks for multi-modal anomaly detection in system software. IEEE Open Journal of the Computer Society.
E. Zavitsanos, D. Mavroeidis, K. Bougiatiotis, E. Spyropoulou, L. Loukas, and G. Paliouras, "Financial misstatement detection: A realistic evaluation," in Proc. 2nd ACM Int. Conf. AI in Finance (ICAIF), 2021, art. 34.
T. Loughran and B. McDonald, "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," J. Finance, vol. 66, no. 1, pp. 35–65, 2011.
L. Purda and D. Skillicorn, "Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection," Contemp. Accounting Res., vol. 32, no. 3, pp. 1193–1223, 2015.
P. Craja, A. Kim, and S. Lessmann, "Deep learning for detecting financial statement fraud," Decis. Support Syst., vol. 139, art. 113421, 2020.
T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, 2019.
J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, Eds., Dataset Shift in Machine Learning. Cambridge, MA, USA: MIT Press, 2009.
J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," ACM Comput. Surv., vol. 46, no. 4, art. 44, 2014.
P. W. Koh et al., "WILDS: A benchmark of in-the-wild distribution shifts," in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 5637–5664.
W. Ping, Y. Jiao, H. Fan, and X. Zhang, "Multimodal fraud detection in financial statements: A trimodal attention network with contrastive evidence chain construction," IEEE Access, vol. 14, pp. 80456–80468, 2026.
H. Yao, C. Choi, B. Cao, Y. Lee, P. W. Koh, and C. Finn, "Wild-Time: A benchmark of in-the-wild distribution shift over time," in Proc. NeurIPS Datasets and Benchmarks Track, 2022.
G. Bai, C. Ling, and L. Zhao, "Temporal domain generalization with drift-aware dynamic neural networks," in Proc. Int. Conf. Learn. Represent. (ICLR), 2023.
A. Nasery, S. Thakur, V. Piratla, A. De, and S. Sarawagi, "Training for the future: A simple gradient interpolation loss to generalize along time," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 34, 2021, pp. 19198–19209.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," J. Mach. Learn. Res., vol. 17, no. 59, pp. 1–35, 2016.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2020.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," arXiv:1907.02893, 2019.

S. Rabanser, S. Günnemann, and Z. C. Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2019.

J. Cao, A. Ganesh, J. Cai, R. Southwell, E. M. Perkoff, M. Regan, K. Kann, J. H. Martin, M. Palmer, and S. D'Mello, "A comparative analysis of automatic speech recognition errors in small group classroom discourse," in Proc. 31st ACM Conf. User Modeling, Adaptation and Personalization (UMAP), 2023, pp. 250–261.

Chen, T., Shang, W., & Guo, Z. (2024). Xsla: An SLA-aware heterogeneous XPU scheduling framework for edge AI inference. World Journal of Information Technology, 2(2), 48–59.

Zou, J., Qin, Y., & Rhee, M. (2024). Beyond model scale: A task-aware reliability benchmark for large language models in clinical decision support. Innovation and Technology Studies, 1(1), 39–47.

Fang, Q. (2024). Bias-aware large language models for equitable Fang, Q. (2024). Bias-aware large language models for equitable educational feedback and assessment. Trends in Social Sciences and Humanities Research, 2(1), 41–50.

Qin, Y., Zou, J., & Rhee, M. (2022). An AI-driven intelligent transaction execution framework for internet finance platforms: Cost-sensitive real-time integrity interdiction. Eurasia Journal of Science and Technology, 4(2), 36–44.

Wang, Y., & Zhang, C. (2023). AERIS: Adaptive executor scoped resource management for privacy preserving data processing in Apache Spark. Eurasia Journal of Science and Technology, 5(2), 25-33.

Zhao, W., & Wang, B. (2022). CAPA: A prediction-driven autoscaling framework for SLO-aware cloud-native machine learning inference. Journal of Computer Science and Electrical Engineering, 4(1), 8-16.

Guo, Z., Chen, T., & Shang, W. (2022). Class-conditional intermediate-domain adversarial adaptation for cross-device acoustic scene classification. Journal of Computer Science and Electrical Engineering, 4(1), 26-35

Shen, A. (2023). Compute-efficient scaling of large sequential recommendation models under dynamic user behavior. World Journal of Information Technology, 1(1), 62–71.

Fang, Q. (2024). True-EWS: Trustworthy early warning with equalized conformal guarantees for equitable student support in U.S. education. Educational Research and Human Development, 1(2), 44–54.

Shen, A., & Huang, S. (2022). A systems-aware framework for intelligent large-scale recommendation models. Journal of Computer Science and Electrical Engineering, 4(1), 36–44.