Journal of Mathematical Finance and Risk Management
ISSN: 3080-9983 (Print), ISSN: 3080-9991 (Online) | Vol. 1, No. 3, 2026

Robust Multimodal Financial Fraud Detection via Missing-Modality-Aware Adaptive Fusion

Beatrice Reed [1], Olivia Huang [2, *]

[1] Rutgers Business School–Newark and New Brunswick, Rutgers University, USA

[2] Questrom School of Business, Boston University, USA

[*] **Corresponding author**: (Email: 3940402012@qq.com)

Abstract: Financial statement fraud detectors increasingly combine several evidence modalities, such as current statements, accrual ratios, multi-year trajectories, and peer benchmarks, yet in operational screening any of these channels may be unavailable for a given firm-year. Detectors trained on largely complete inputs then emit scores whose scale depends on which modalities happen to be observed, and the pooled ranking used for audit targeting deteriorates sharply. This paper presents MAAF, a missing-modality-aware adaptive fusion framework that (i) condenses each modality into a standardized evidence score produced by a modality expert, (ii) learns cross-modal evidence surrogates that infer the evidence of an absent modality from whichever modalities are observed, using dense distillation targets instead of scarce fraud labels, and (iii) combines observed and surrogate evidence through a fusion layer whose weights, surrogate trust, and calibration are conditioned on the availability pattern and trained with masked-pattern augmentation and cross-pattern consistency. On 146,045 U.S. firm-years (1990–2014) with AAER-based fraud labels and a strict temporal protocol, MAAF attains a test AUC of 0.720 under natural missingness and retains 0.711 when each modality is removed with probability 0.7, whereas the strongest deep fusion baseline falls to 0.653 and the boosted-tree detectors fall to 0.605 or lower. Ablations show that evidence surrogates and masked-pattern training account for most of the robustness gain.

Keywords: Financial statement fraud detection, multimodal learning, missing modality, adaptive fusion, knowledge distillation, AAER.

1. Introduction

Financial statement fraud undermines the informational foundation of capital markets, and regulators, auditors, and investors therefore rely on statistical screening models to decide which firms deserve scrutiny. Most screening research uses the SEC's Accounting and Auditing Enforcement Releases (AAERs) as ground truth and has progressed from parsimonious accounting scores such as the Beneish M-score [1] and the Dechow F-score [2] to machine-learning detectors that capture nonlinear interactions among many predictors [3–6]. Bao et al. [5] showed that ensemble learning on raw financial statement items markedly improves out-of-sample detection, and Bertomeu et al. [6] documented the value of interacting accounting variables with market and audit signals.

A parallel line of work broadens the evidence base beyond financial ratios to narrative disclosures [7] and document layout, and fuses these heterogeneous streams with multimodal architectures (Section II-A). Such detectors are developed on filings for which every evidence stream can be extracted, which leaves open how they should operate when part of the evidence is unavailable at screening time.

Missingness is pervasive and structured in financial data. Newly listed firms have no multi-year history, ratio engines require prior-year statements, peer benchmarks are undefined for thin industry–size cells, and text or layout extraction may fail for individual filings. Robustness to missing modalities has been studied through modality dropout [8], cross-modal imputation [9, 10], Bayesian meta-learning [11], fusion-strategy search [12], missing-aware prompting [13], and shared-specific representations [14], but almost exclusively for vision, language, and affective-computing tasks.

Our experiments reveal a failure mode that is specific to ranking-based fraud screening. When availability varies across firm-years, a detector must rank firms observed through different modality subsets on a single scale. A model can remain informative within every availability pattern and still produce a poor pooled ranking when its scores are not comparable across patterns. For example, a gradient-boosted tree model trained on all modalities attains an AUC of 0.694 when the statement modality is absent for every firm, but only 0.587 when each modality is independently missing with probability 0.5 (Section V).

We address this problem with MAAF (Missing-modality-Aware Adaptive Fusion), which operates on modality evidence rather than raw features. MAAF builds one evidence expert per modality and stacks their out-of-fold evidence [15] through a pattern-conditioned gate in the spirit of mixtures of experts [16]. It further learns cross-modal evidence surrogates that regress the evidence of each modality on the remaining modalities. This distillation [17] turns every training firm-year into a supervised example, rather than only the few hundred fraud cases, and supplies calibrated stand-in evidence whenever a modality is absent. Observed and surrogate evidence are fused with weights, surrogate trust, and calibration offsets that depend on the availability pattern, and the fusion layer is trained on masked availability patterns with a cross-pattern consistency regulariser.

The contributions of this paper are threefold. First, we formalize missing-modality fraud screening as a pooled ranking problem and show that cross-pattern score comparability, not only within-pattern accuracy, governs detection quality. Second, we propose MAAF, which combines modality evidence experts, cross-modal evidence surrogates, and a missingness-aware adaptive fusion layer with fewer than 60 trainable fusion parameters. Third, we

build a missing-modality benchmark on 146,045 AAER-labelled U.S. firm-years with four evidence modalities and natural as well as simulated missingness, and show that MAAF outperforms all ten baselines at every missing rate while achieving the highest AUC under natural missingness.

2. Related Work

2.1. Financial Statement Fraud Detection

Early detectors combine a handful of accrual and performance ratios in discriminant or logistic models [1, 2]. Perols [3] compared six statistical and machine-learning classifiers under realistic cost assumptions, and Abbasi et al. [4] introduced MetaFraud, a meta-learning framework that stacks context-specific base classifiers. Bao et al. [5] demonstrated that RUSBoost trained on 28 raw accounting items outperforms ratio-based logistic models and introduced the ranking metric NDCG@k, which reflects the limited inspection capacity of regulators. Bertomeu et al. [6] showed that accounting variables become informative mainly through interactions with audit and market variables. Beyond numerical data, Craja et al. [7] combined hierarchical attention over MD&A text with financial ratios. Evidence-based multi-agent debate for financial statement fraud offers a complementary approach to combining accounting evidence with traceable narrative reasoning [24].

Ping et al. [18] proposed TriAudit, which encodes financial ratios with a residual network, MD&A text with a FinBERT-based encoder, and page layout with LayoutLMv2 augmented by a Layout Deviation Score (LDS) that measures departures from industry-normative templates. Hierarchical cross-modal attention and a reliability-gated fusion module combine the three streams, and a contrastive module organizes per-modality anomaly scores into auditable evidence chains. On AAER-labelled 10-K filings from the EDGAR corpus, TriAudit reached an AUC of 0.923, and its ablations showed that the layout deviation scoring and the hierarchical attention each contribute to accuracy. TriAudit was evaluated on filings for which all three streams could be extracted; MAAF addresses the complementary screening setting in which streams may be unavailable, and Section IV-A specifies which of its components are adapted to structured financial evidence.

2.2. Multimodal Fusion

Multimodal fusion strategies are commonly organized into early, late, and hybrid schemes [19]. Mixtures of experts [16] and stacked generalization [15] combine specialized predictors through a learned gate or a meta-learner trained on out-of-fold predictions. Dynamic fusion methods adapt the weight of each modality to sample-specific quality: QMF [20] derives fusion weights from energy-based uncertainty with generalization guarantees, and reliability-gated fusion modules weight each evidence stream by per-filing reliability. MAAF follows this reliability-aware line and makes the availability pattern itself a first-class reliability signal.

2.3. Learning with Missing Modalities

ModDrop [8] randomly drops modality channels during training. CRA [9] and MMIN [10] impute missing representations with cascaded residual autoencoders and cross-modal imagination, SMIL [11] uses Bayesian meta-learning for severely missing modalities, Ma et al. [12] search for robust fusion strategies in multimodal transformers, Lee et al. [13] attach missing-aware prompts, and ShaSpec [14] substitutes missing modalities with shared features. These methods impute or align deep representations and are evaluated per sample. Fraud screening differs in three respects: modalities are tabular and subject to strong temporal drift, positives are extremely rare, and decisions are made on a pooled ranking. MAAF therefore imputes at the level of calibrated modality evidence and trains its fusion layer explicitly for cross-pattern comparability.

3. Problem Formulation

Each firm-year i has a fraud label $y_i \in \{0, 1\}$, derived from AAERs, and up to M evidence modalities with feature vectors x_i^m. The availability vector $\pi_i \in \{0, 1\}^M$ records which modalities are observed, and the observed set is

$$O_i = \{m : \pi_{im} = 1\} \tag{1}$$

A screening model $s(\cdot)$ assigns a score to every firm-year in a test period using only the observed modalities and π_i, and the firm-years are ranked jointly. Let p and q index availability patterns, and let n_p^+ and n_q^- be the numbers of fraudulent and clean firm-years observed through p and q. The pooled AUC decomposes as

$$\mathrm{AUC} = \sum_{p,q} \frac{n_p^+ n_q^-}{n^+ n^-} A_{pq} \tag{2}$$

Where A_{pq} is the probability that a fraudulent firm-year observed through p outranks a clean firm-year observed through q. The diagonal terms A_{pp} are within-pattern AUCs, which a pattern-wise accurate model maximizes. The off-diagonal terms compare evidence of different provenance and are maximized only when scores are comparable across patterns, that is, when s is a common monotone transform of $P(y = 1 \mid x^O, \pi)$. Under random missingness with M = 4 modalities the off-diagonal pairs dominate (2): with independent drop probability 0.5, more than 90% of positive–negative pairs straddle different patterns. This observation motivates the design goals of MAAF: calibrated evidence per modality, recovery of the evidence of missing modalities, and pattern-conditioned fusion.

4. Methodology

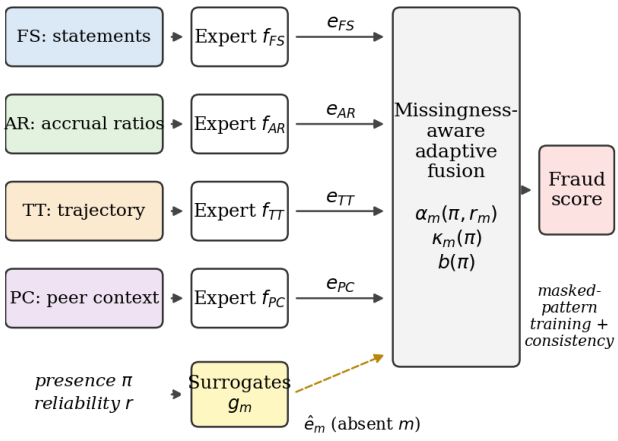


Figure 1. Overview of MAAF. Each available modality is condensed into evidence e_m by a modality expert; evidence surrogates infer the evidence of absent modalities from the observed ones; the missingness-aware fusion layer combines observed and surrogate evidence with weights α_m, surrogate trust κ_m, and calibration $b(\pi)$, all conditioned on the availability pattern π.

Fig. 1 summarizes MAAF. Each observed modality is

condensed into calibrated evidence by a modality expert, the evidence of absent modalities is inferred by cross-modal surrogates, and a missingness-aware fusion layer combines both kinds of evidence with weights that depend on the availability pattern.

4.1. Evidence Modalities and Deviation Measures

We construct four modalities with distinct acquisition pipelines and data structures (Table I). The statement modality (FS) contains the 28 raw accounting items of Bao et al. [5] for fiscal year t. The accrual-ratio modality (AR) contains the 14 accrual, performance, off-balance-sheet, and market-incentive ratios of Dechow et al. [2] and has feature-level missingness. The trajectory modality (TT) stacks the statement items of years t−3 to t−1 and is absent for newly listed firms. Because nominal magnitudes drift strongly across three decades, statement and ratio features are mapped to cross-sectional rank-Gaussian scores within each fiscal year, a standard device in financial machine learning [21]:

$$\tilde{x}_{ij} = \Phi^{-1}\left(\frac{\mathrm{rank}_t(x_{ij}) - 0.5}{n_{ij}}\right) \tag{3}$$

Where n_{tj} is the number of firms with an observed value of feature j in year t. The fourth, peer-context modality (PC) contains robust deviations of 18 common-size items and the 14 ratios from the contemporaneous fiscal-year × size-decile peer group G,

$$z_{ij} = \frac{v_{ij} - \mathrm{med}_G(v_j)}{\mathrm{IQR}_G(v_j)}, \quad \mathrm{PDS}_i = \mathrm{mean}_j\,|z_{ij}| \tag{4}$$

Together with their mean absolute value, the Peer Deviation Score (PDS).

Table I. Evidence modalities and natural availability (%)

Mod.	Dim.	Content	Avail. train / test
FS	28	Raw statement items (balance sheet, income, cash flow, market)	100.0 / 100.0
AR	14	Accrual, performance, off-balance-sheet and market ratios	100.0 / 100.0
TT	3×28	Statement items of years t−3 to t−1	87.0 / 91.6
PC	33	Peer-relative robust z-scores and PDS	99.8 / 99.4

Two deviation measures in MAAF are adapted from TriAudit [18]. Its LDS scores how far the token layout of a filing page departs from an industry-normative template estimated on comparable filings; we transfer this idea from document geometry to financial features, taking the peer-group median as the normative profile and the PDS in (4) as the aggregate departure. Because the panel provides no industry codes, peers are formed from fiscal-year × size-decile cells, and the norm is estimated from all peer firms rather than from non-fraudulent filings only, so that the PDS can be computed at screening time without labels. Second, we implement TriAudit's cross-modal anomaly score, the cosine distance between a modality representation and the centroid of clean training filings, on the feature vector of each modality:

$$d_{im} = 1 - \frac{x_i^m \cdot \bar{x}^m_{\mathrm{clean}}}{\|x_i^m\|\,\|\bar{x}^m_{\mathrm{clean}}\|} \tag{5}$$

MAAF's experts output scores rather than embeddings, so the centroid is a fixed average over clean training firm-years instead of an exponential moving average of learned embeddings, and d_{im} is standardized on the training data. Both quantities serve as reliability information in the fusion layer of Section IV-D, where conditioning the gate on availability extends the per-filing reliability gating of the same framework to incomplete evidence. Unlike TriAudit, which incorporates textual and layout evidence, our implementation fuses four structured financial evidence streams through availability-aware reliability weighting.

4.2. Modality Evidence Experts

For each modality m we fit an expert f_m on the firm-years in which m is observed. We use gradient-boosted decision trees (LightGBM [22]), which handle feature-level missingness natively and are robust to heavy-tailed accounting data. To obtain unbiased evidence for training the fusion layer, the training firm-years are split into K = 5 folds that are stratified by label and grouped by firm, and every firm-year receives the out-of-fold prediction of the expert that did not see it, as in stacked generalization [15]. The expert logits are standardized on the training data,

$$e_{im} = \frac{f_m(x_i^m) - \mu_m}{\sigma_m} \tag{6}$$

So that evidence from different modalities lives on a common, approximately calibrated scale.

4.3. Cross-Modal Evidence Surrogates

When modality m is missing, its evidence is not lost entirely: accounting quantities are strongly coupled, so the observed modalities carry information about what e_m would have been. For each m we learn a surrogate g_m that maps the features of the other modalities to the evidence of m:

$$\min_{g_m} \sum_{i\,:\,\pi_{im}=1} \left(g_m(x_i^{O_i \setminus m}) - e_{im}\right)^2 \tag{7}$$

The targets are the out-of-fold expert evidence, so the surrogate is a student distilled [17] from the expert of m and receives a dense, continuous target for every training firm-year instead of a binary label that is positive for fewer than one firm-year in 150. Surrogates are gradient-boosted regressors trained with availability augmentation, in which each other modality is additionally removed with probability 0.5. Because g_m depends only on which of the other M−1 modalities are observed, its outputs are precomputed for all 2^{M-1} availability codes, and out-of-fold surrogate evidence $\hat{e}_{im}$ is produced with the same folds as the experts.

4.4. Missingness-Aware Adaptive Fusion

The fusion layer combines the evidence of observed modalities with surrogate evidence for absent ones:

$$s(x_i, \pi_i) = b(\pi_i) + \sum_m \pi_{im}\alpha_{im}e_{im} + \sum_m (1-\pi_{im})\kappa_{im}\hat{e}_{im} \tag{8}$$

$$\begin{aligned} \alpha_{im} &= \mathrm{softplus}(\beta_m + \gamma_m^\top \pi_i + \delta_m^\top r_{im}) \\ \kappa_{im} &= \mathrm{softplus}(\eta_m + h_m^\top \pi_i) \end{aligned} \tag{9}$$

With the pattern calibration $b(\pi) = b_0 + b^\top\pi$. The descriptor vector $r_{im} = (o_{im}, d_{im})$ contains the standardized fraction of observed features within the modality and the deviation d_{im} of (5). The evidence weights α thus depend on the availability pattern and on per-filing reliability, the trust κ placed in surrogate evidence depends on which modalities support the

surrogate, and $b(\pi)$ aligns score levels across patterns, as required by the off-diagonal terms of (2). The layer has $M(2M + 5) + 1 = 53$ parameters for $M = 4$, which keeps it well conditioned despite the scarcity of fraud labels.

4.5. Training with Masked Availability Patterns

The natural training data contain few availability patterns, so the fusion layer is trained on an augmented set in which the observed modalities of every training firm-year are additionally masked with rates $\tilde{\eta} \in \{0.2, 0.4, 0.6, 0.7\}$, at least one modality is retained, and surrogate evidence is gathered for the resulting pattern. Let $\tilde{\pi}$ denote a masked pattern and ℓ the logistic loss. The objective combines label fitting with a cross-pattern consistency term that pulls the prediction under a masked pattern toward the stop-gradient prediction under the full pattern:

$$\mathcal{L} = \mathbb{E}_{\tilde{D}}[\ell(y, s(x, \tilde{\pi}))] + \lambda \mathbb{E}_{\tilde{D}}[\ell(\sigma(\mathrm{sg}\, s(x, \pi)),\ s(x, \tilde{\pi}))] + \rho \|\theta\|^2 \quad (10)$$

Where σ is the logistic function, sg denotes stop-gradient, $\lambda = 0.5$, and $\rho = 10^{-3}$. The full procedure is summarized in Algorithm 1. Optimizing the unweighted logistic loss over all patterns encourages s to approximate the log-odds of fraud given the observed evidence, which is precisely the cross-pattern comparability required by (2).

Algorithm 1 Training MAAF

Input: training firm-years (x_i, π_i, y_i); folds K; masking rates H.

1: For each fold k, fit experts f_m on the other folds and compute out-of-fold evidence e_{im} (6).

2: Fit surrogates g_m on out-of-fold evidence with availability augmentation (7); tabulate $\hat{e}_{im}$ for all 2^{M-1} availability codes.

3: Build the augmented set $\tilde{D}$ by masking π_i with every $\tilde{\eta} \in$ H.

4: Optimize (10) over the fusion parameters θ with Adam.

5: Refit f_m and g_m on all training data for deployment.

Output: experts $\{f_m\}$, surrogates $\{g_m\}$, fusion parameters θ.

At inference MAAF evaluates at most M experts and M surrogates followed by an $O(M^2)$ fusion step. On a single CPU core, fitting all experts with out-of-fold evidence took 1.3 min and fitting the surrogates 6.2 min per run, while training the fusion layer took less than half a minute.

5. Experiments

5.1. Data and Protocol

We use the public AAER-labelled firm-year panel released by Bao et al. [5], which covers 146,045 Compustat firm-years from 1990 to 2014, 964 of which are fraudulent. Following their protocol, the data are split chronologically to rule out look-ahead bias. Firm-years from 1991–2001 form the training period (67,166 firm-years, 400 frauds), 2002 is left out as a two-year gap, and 2003–2008 form the test period (35,166 firm-years, 261 frauds, base rate 0.74%). Training frauds belonging to AAER cases that continue into the test period are recoded as non-fraud (99 firm-years), as prescribed in [5]. A firm-grouped 15% subset of the training period serves as validation data for the deep baselines. Natural missingness is substantial: the trajectory modality is absent for 13.0% of training firm-years and 3.8% of ratio values are missing.

We evaluate three missingness protocols on the test period: natural, which uses the observed availability; random, in which every available modality is removed independently with probability $\eta \in \{0.1, 0.3, 0.5, 0.7\}$ while at least one modality is retained, averaged over three draws; and fixed subsets, in which only a given subset of the four modalities is available, for all 15 subsets. We report AUC, average precision (AP), NDCG@1%, and the sensitivity and precision obtained when the top 1% of firm-years are flagged. All results are means ± standard deviations over two runs with different random seeds; the deterministic LR baseline is run once.

5.2. Baselines and Implementation

Ten baselines cover three families. Accounting and tree models: a logistic F-score-style model on the AR ratios (LR) [2]; RUSBoost [23] on the 28 raw items with the configuration of Bao et al. [5]; and LightGBM [22] on all modalities with native missing-value handling. Deep fusion models: early fusion of modality embeddings (EF), late fusion of modality-wise predictions (LF), and cross-modal attention fusion with gated aggregation (XAF). Missing-modality methods: ModDrop [8], MMIN [10], ShaSpec [14], and QMF [20], re-implemented on our modalities. All deep models share the same modality encoders, namely two-layer residual multilayer perceptrons with GELU activations and layer normalization, and a gated recurrent encoder for the trajectory; they use 64-dimensional embeddings, AdamW with learning rate $3\times10-4$, batches of 512 firm-years with eight oversampled frauds, at most ten epochs, and early stopping on validation AUC. The missing-modality methods are trained with modality dropout at rate 0.4.

MAAF uses LightGBM experts with 300 trees, learning rate 0.03, 15 leaves, at least 100 samples per leaf, row and column subsampling of 0.8 and 0.7, and L2 penalty 5. The surrogates use 200 trees, learning rate 0.05, and 31 leaves. The fusion layer is trained by full-batch Adam with learning rate 0.05 for 400 steps.

5.3. Results under Natural Missingness

Table II. Detection performance under natural missingness

Method	AUC	AP (%)	NDCG@1%	Sens@1%	Prec@1%
LR (F-score) [2]	0.659±.000	1.22±.00	0.92±.00	0.77±.00	0.57±.00
RUSBoost [5]	0.711±.004	1.53±.04	1.98±.52	2.49±.57	1.85±.43
LightGBM [22]	0.703±.000	1.69±.00	3.40±.10	4.02±.19	2.98±.14
EF	0.659±.009	1.44±.07	2.84±1.20	2.87±.96	2.13±.71
LF	0.691±.008	1.51±.08	1.97±.19	2.49±.19	1.85±.14
XAF	0.660±.027	1.42±.03	2.77±1.34	2.68±1.15	1.99±.85
ModDrop [8]	0.664±.007	1.46±.03	2.99±.36	3.45±.38	2.56±.28
MMIN [10]	0.671±.004	1.94±.16	5.80±1.26	4.02±.96	2.98±.71
ShaSpec [14]	0.667±.006	1.49±.02	3.67±.63	4.21±1.15	3.13±.85
QMF [20]	0.656±.023	1.34±.11	1.91±.79	1.92±.77	1.42±.57
MAAF (ours)	0.720±.000	1.78±.00	3.56±.39	4.41±.57	3.27±.43

Mean ± std over two runs; NDCG, sensitivity and precision in %; best values in bold.

Table II reports the results with the availability observed in the data. MAAF obtains the highest AUC, sensitivity@1% and precision@1%, with an AUC of 0.720 against 0.711 for the strongest baseline (RUSBoost). MMIN attains the highest AP (1.94%) and NDCG@1% (5.80%), so the ordering at the very top of the ranking under natural availability is shared with this method. The ranking among baselines mirrors the accounting literature: the RUSBoost configuration of [5] remains a strong single-modality detector, deep models trained end-to-end on the four modalities are less accurate under the temporal shift between the 1990s and the post-2002 period, and late fusion is the most accurate deep architecture. MAAF retains the accuracy of tree-based experts while adding the trajectory and peer-context evidence through calibrated fusion.

5.4. Robustness to Missing Modalities

Table III. Test AUC under random and structured missingness

Method	$\eta = 0.1$	$\eta = 0.3$	$\eta = 0.5$	$\eta = 0.7$	No FS
LR (F-score) [2]	0.651±.000	0.623±.000	0.599±.000	0.577±.000	0.659±.000
RUSBoost [5]	0.704±.003	0.680±.001	0.639±.006	0.605±.001	0.517±.000
LightGBM [22]	0.679±.002	0.605±.003	0.587±.006	0.551±.003	0.694±.011
EF	0.622±.008	0.585±.001	0.566±.002	0.552±.002	0.651±.006
LF	0.687±.007	0.684±.010	0.665±.008	0.653±.001	0.677±.012
XAF	0.635±.022	0.591±.021	0.577±.012	0.556±.014	0.637±.022
ModDrop [8]	0.659±.003	0.645±.001	0.642±.002	0.619±.002	0.678±.014
MMIN [10]	0.669±.004	0.666±.003	0.660±.005	0.644±.005	0.665±.001
ShaSpec [14]	0.665±.005	0.657±.001	0.655±.002	0.638±.009	0.674±.011
QMF [20]	0.650±.022	0.628±.032	0.616±.016	0.594±.024	0.641±.005
MAAF (ours)	0.717±.000	0.717±.000	0.715±.000	0.711±.000	0.716±.000

η: independent drop probability per modality; No FS: statement modality absent for all firm-years.

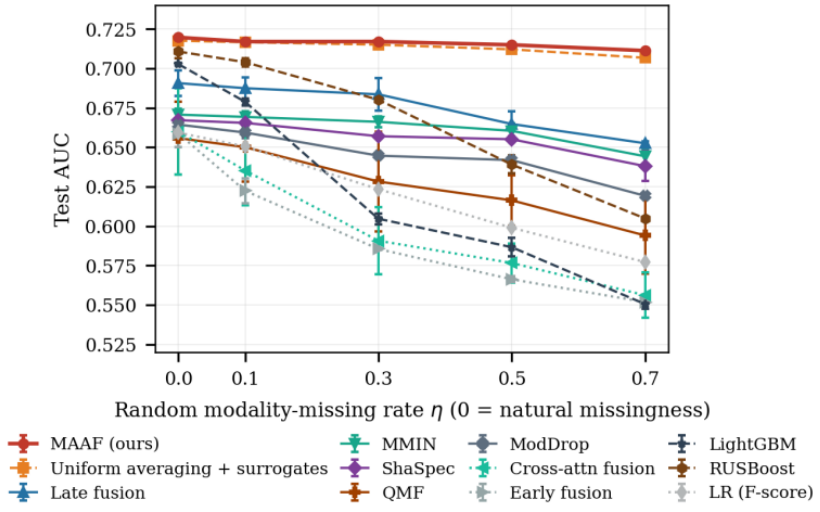


Figure 2. Test AUC as a function of the random missing rate η (mean ± std over two runs). MAAF degrades gracefully, whereas joint models that were trained on largely complete inputs lose most of their advantage once availability becomes heterogeneous.

Table III and Fig. 2 show that missingness separates the methods sharply. MAAF achieves the highest AUC at every missing rate. At $\eta = 0.5$ it reaches 0.715, 5.0 points above the best baseline (LF, 0.665); at $\eta = 0.7$ it reaches 0.711, 5.9 points

above LF. Over the whole range MAAF loses only 0.8 AUC points, whereas LightGBM loses 15.2 and RUSBoost 10.6. The comparison between the no-FS column and the random protocol is instructive: LightGBM still achieves 0.694 when the statement modality is missing for every firm, which confirms that its within-pattern accuracy is intact, yet its pooled AUC collapses when patterns are mixed. This is exactly the cross-pattern comparability problem formalized in (2). RUSBoost relies on the statement modality alone and falls to 0.517 without it, while MAAF retains 0.716.

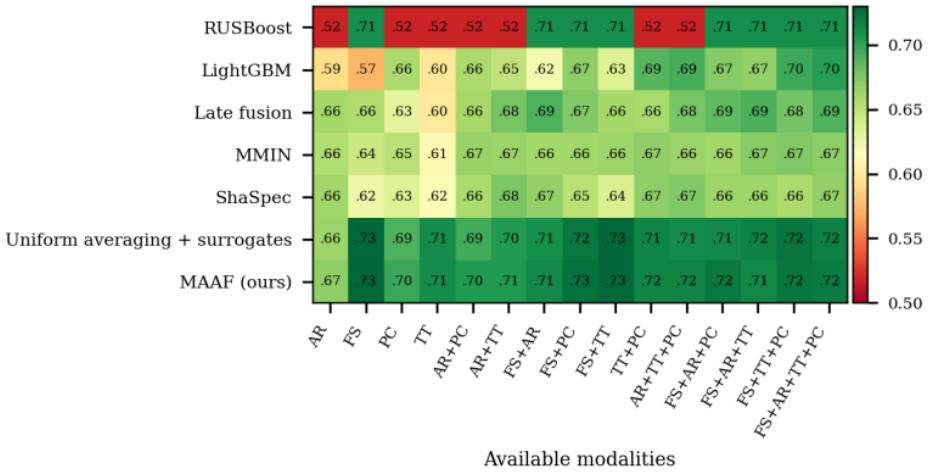


Figure 3. Test AUC for each of the 15 subsets of available modalities (FS: statements, AR: ratios, TT: trajectory, PC: peer context).

Fig. 3 breaks robustness down by availability subset. MAAF is the best of the compared methods in 14 of the 15 subsets and never drops below 0.669, even when a single weak modality is available. Its average margin over the best competing method is 3.4 AUC points on subsets without the statement modality and 1.2 points on subsets that contain it, reflecting the FS surrogate, which recovers much of the statement evidence from the ratio, trajectory, and peer-context modalities.

5.5. Ablation Study

Table IV. Ablation of MAAF components (test AUC)

Variant	Natural	$\eta = 0.5$	$\eta = 0.7$	No FS
MAAF (full)	0.720±.000	0.715±.000	0.711±.000	0.716±.000
w/o evidence surrogates	0.718±.001	0.696±.001	0.681±.000	0.699±.000
w/o masked-pattern training	0.719±.001	0.688±.002	0.688±.002	0.704±.000
w/o pattern-adaptive weights	0.720±.001	0.716±.000	0.710±.000	0.716±.001
w/o pattern calibration	0.720±.000	0.716±.000	0.711±.000	0.716±.000
w/o reliability descriptors	0.719±.000	0.715±.000	0.711±.000	0.716±.000
w/o consistency regulariser	0.719±.000	0.715±.000	0.711±.001	0.716±.001
Uniform expert averaging	0.718±.002	0.697±.001	0.680±.000	0.701±.001
Uniform averaging + surrogates	0.718±.001	0.712±.001	0.707±.000	0.711±.001

Table IV isolates the components. Removing the evidence surrogates lowers AUC by 1.9 points at $\eta = 0.5$ and 3.0 points at $\eta = 0.7$, and training the fusion layer only on natural patterns costs 2.7 and 2.3 points, respectively. These two components account for most of the robustness gain. The finer design choices, namely pattern-adaptive weights, pattern calibration, reliability descriptors, and the consistency regulariser, change AUC by at most 0.1 points in any setting; on this benchmark their information is largely subsumed by the expert and surrogate evidence. Uniform averaging of observed expert evidence is a strong reference under natural missingness (0.718) but degrades to 0.680 at $\eta = 0.7$. Adding surrogate evidence to the uniform average recovers most of the robustness (0.707), and the learned missingness-aware fusion adds a further 0.5 points at $\eta = 0.7$ and 0.5 points without the statement modality.

5.6. Analysis

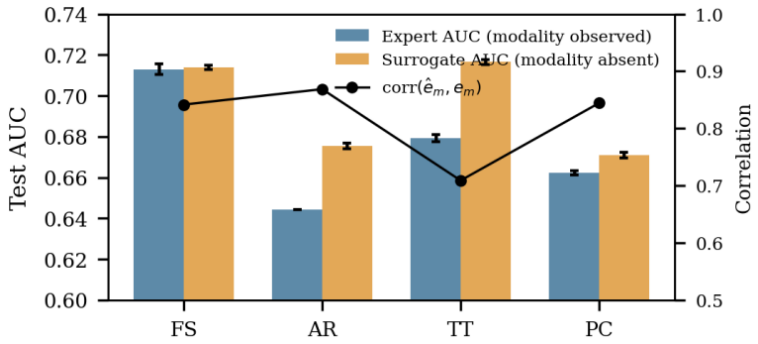


Figure 4. Test AUC of each modality expert (on firm-years where the modality is observed) and of its evidence surrogate computed from the other three modalities, together with the correlation between surrogate and expert evidence.

Evidence surrogates. Fig. 4 compares each expert with its surrogate. The surrogate evidence correlates with the expert evidence at 0.84 for FS, 0.87 for AR, 0.71 for TT, and 0.84 for PC. As fraud detectors, the surrogates reach AUCs of 0.714 (FS), 0.676 (AR), 0.717 (TT), and 0.671 (PC), against 0.713, 0.644, 0.679, and 0.662 for the corresponding experts. Because each surrogate pools three modalities and is trained on dense targets, it can match or exceed the expert of a weak modality, which explains why missing modalities cost MAAF so little. Dynamic heterogeneous graph contrastive learning for collusive fraud further illustrates the predictive value of inter-entity relationships as an additional category of financial-risk evidence [25].

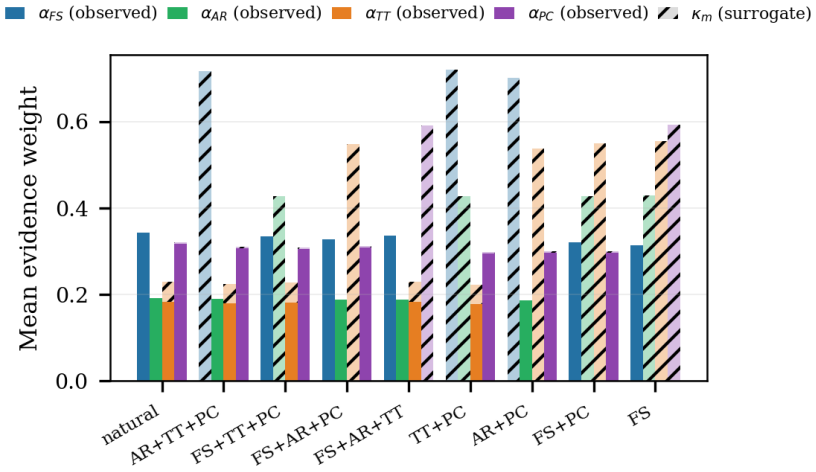


Figure 5. Mean fusion weights of observed evidence (solid) and surrogate evidence (hatched) for the natural pattern and for representative availability patterns.

Fusion weights. Fig. 5 shows how the fusion layer adapts to availability. Under natural availability the largest weight is assigned to FS evidence ($\alpha = 0.34$). When the statement modality is missing, the layer places a trust of $\kappa = 0.72$ on the FS surrogate and redistributes weight over the observed modalities (αAR = 0.19, αTT = 0.18, αPC = 0.31).

Deviation measures. The deviation measures of Section IV-A carry fraud signal of their own, with a modality-dependent direction. The deviation descriptor separates fraudulent from clean firm-years with an AUC of 0.653 in the ratio modality, where fraudulent firm-years lie farther from the clean centroid, and with AUCs of 0.370 and 0.358 in the statement and trajectory modalities, where they lie closer to it; in the peer-context modality it is uninformative (0.505). Because the gate learns signed coefficients, both directions are usable, and their overlap with the expert evidence explains the small ablation effect in Table IV. The scalar PDS alone ranks test firm-years with an AUC of 0.537, whereas the full peer-deviation vector supports an expert with an AUC of 0.662 and strengthens MAAF: withholding the peer-context modality at test time lowers the AUC from 0.720 to 0.714.

Table V. Evidence profiles of three test firm-years

	Case A	Case B	Case C
Fiscal year, label	2008, fraud	2004, fraud	2007, clean
Rank (top %)	66 (0.19%)	1,144 (3.25%)	1 (0.003%)
FS	+4.22 · .35 · −0.9	+1.08 · .34 · −1.4	+4.51 · .36 · −1.6
AR	+1.29 · .19 · +0.8	+1.84 · .19 · +0.7	+2.14 · .19 · +0.8
TT	+4.26 · .19 · −1.6	+0.90† · .55 · –	+3.90 · .19 · −1.8
PC	+0.74 · .32 · +0.1	+2.09 · .31 · +0.1	+2.91 · .32 · −0.3
Score (log-odds)	−2.86	−3.87	−1.93

Entries: evidence e (surrogate ê marked †) · weight α (κ for surrogates) · deviation d; first run.

Evidence profiles. Adapting the evidence chains of TriAudit [18] to incomplete inputs, Table V decomposes the MAAF score of three test firm-years into modality-level evidence, fusion weight, and deviation descriptor; an absent modality appears as surrogate evidence with an explicit trust weight. Case A, a fraudulent firm-year ranked in the top 0.19%, receives its largest contributions from the statement (+1.49) and trajectory (+0.82) evidence. Case B, a fraudulent firm-year without a trajectory record that is ranked in the top 3.25%, draws its second-largest contribution (+0.49) from the TT surrogate inferred from the other three modalities. Case C, the highest-ranked clean firm-year, combines positive evidence in all four modalities, so its profile shows the reviewer that the alert rests on concordant rather than isolated signals.

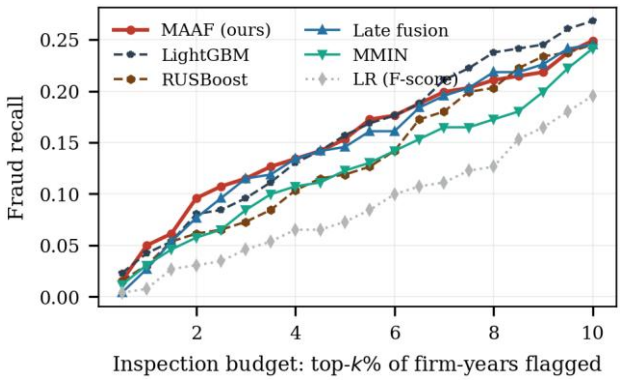


Figure 6. Fraud recall as a function of the inspection budget (share of test firm-years flagged) under natural missingness, first run.

Screening efficiency. Fig. 6 plots fraud recall against the share of firm-years flagged, the quantity of interest for regulators with a fixed inspection budget. With a budget of 5%, MAAF recovers 15.3% of frauds, compared with 15.7% for LightGBM, 11.9% for RUSBoost, and 14.6% for late fusion; with a 10% budget it recovers 24.9%.

6. Conclusion

This paper studied financial statement fraud detection when evidence modalities are missing, a condition that is the norm rather than the exception in operational screening. We showed that pooled fraud rankings depend on cross-pattern score comparability and proposed MAAF, which condenses each modality into calibrated expert evidence, recovers the evidence of absent modalities through distilled cross-modal surrogates, and fuses observed and surrogate evidence with pattern-conditioned weights trained on masked availability patterns. On 146,045 AAER-labelled U.S. firm-years under a strict temporal protocol, MAAF reaches an AUC of 0.720 with natural availability and 0.711 when 70% of modalities are randomly removed, clearly ahead of accounting, tree-based, deep-fusion, and missing-modality baselines under missing evidence.

The evidence-level interface of MAAF extends directly to unstructured modalities such as the MD&A text and document layout used by TriAudit, which can enter as additional evidence experts without architectural changes. Further directions include modelling informative missingness, where outages depend on firm characteristics, evaluating additional runs and test windows, and updating the fusion layer online as new AAERs are released.

References

[1] Beneish, M. D. (1999). The detection of earnings manipulation. Financial Analysts Journal, 55(5), 24–36.

[2] Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting material accounting misstatements. Contemporary Accounting Research, 28(1), 17–82.

[3] Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. Auditing: A Journal of Practice & Theory, 30(2), 19–50.

[4] Abbasi, A., Albrecht, C., Vance, A., & Hansen, J. (2012). MetaFraud: A meta-learning framework for detecting financial fraud. MIS Quarterly, 36(4), 1293–1327.

[5] Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. Journal of Accounting Research, 58(1), 199–235.

[6] Bertomeu, J., Cheynel, E., Floyd, E., & Pan, W. (2021). Using machine learning to detect misstatements. Review of Accounting Studies, 26(2), 468–519.

[7] Craja, P., Kim, A., & Lessmann, S. (2020). Deep learning for detecting financial statement fraud. Decision Support Systems, 139, 113421.

[8] Wang, J., Liu, J., Ge, Y., & Jiang, B. (2023). Risk-aware ensemble learning for credit default prediction under economic uncertainty. World Journal of Economics and Business Research, 1(1), 18–32.

[9] Tran, L., Liu, X., Zhou, J., & Jin, R. (2017). Missing modalities imputation via cascaded residual autoencoder. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1405–1414).

[10] Zhao, J., Li, R., & Jin, Q. (2021). Missing modality imagination network for emotion recognition with uncertain missing modalities. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (pp. 2608–2618).

[11] Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., & Peng, X. (2021). SMIL: Multimodal learning with severely missing modality. In Proceedings of the AAAI Conference on Artificial Intelligence, 35(3), 2302–2310.

[12] Ma, M., Ren, J., Zhao, L., Testuggine, D., & Peng, X. (2022). Are multimodal transformers robust to missing modality? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18177–18186).

[13] Lee, Y. L., Tsai, Y. H., Chiu, W. C., & Lee, C. Y. (2023). Multimodal prompting with missing modalities for visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14943–14952).

[14] Chen, J., Han, X., Zhang, X., & Yang, Y. (2023). Exact concept-level explanations for early detection of financial misstatements. Eurasia Journal of Science and Technology, 5(2), 34–48.

[15] Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259.

[16] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. Neural Computation, 3(1), 79–87.

[17] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. https://doi.org/10.48550/arXiv.1503.02531

[18] Ping, W., Jiao, Y., Fan, H., & Zhang, X. (2026). Multimodal fraud detection in financial statements: A trimodal attention network with contrastive evidence chain construction. IEEE Access, 14, 80456–80468. https://doi.org/10.1109/ACCESS.2026.3695466

[19] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.

[20] Chen, J., Zhang, X., & Han, X. (2023). Explainable identification of tax-related financial reporting anomalies in U.S. public companies: Cross-fitted tax-consistency residuals and Owen-value Shapley propagation. World Journal of Economics and Business Research, 1(1), 33–45.

[21] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. Review of Financial Studies, 33(5), 2223–2273.

[22] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.

[23] Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 40(1), 185–197.

[24] Yue, X., Yang, J., & Liu, W. (2026). FRAUDDEBATE-AGENT: A Multi-Agent LLM Framework with an Evidence-Based Debate Mechanism for Financial Statement Fraud Detection. Mathematics, 14(15), 2695. https://doi.org/10.3390/math14152695

[25] Jiao, Y., Fan, H., Yue, X., Ping, W., Sun, T., & Wang, J. (2026). Dynamic heterogeneous graph contrastive learning for uncovering collusive financial fraud. Scientific Reports. https://doi.org/10.1038/s41598-026-58938-5