Journal of Sustainability, Policy, and Practice

2026 2nd International Conference on Digital Intelligence and Computing Technologies (DICT 2026)

Article

A Comparative Analysis of Unsupervised AI Anomaly Detection Algorithms for Identifying Suspicious Trading Patterns Around U.S. Corporate Earnings Announcements

Xujia Chen[1,*]

[1] Master of Computer and Information Technology, University of Pennsylvania, Philadelphia, PA, USA
* Correspondence: Xujia Chen, Master of Computer and Information Technology, University of Pennsylvania, Philadelphia, PA, USA

Abstract: Suspicious trading activity around U.S. corporate earnings announcements remains a central concern for the Securities and Exchange Commission. This study conducts a comparative evaluation of four unsupervised AI anomaly detection algorithms---Isolation Forest, One-Class Support Vector Machine (OC-SVM), Long Short-Term Memory (LSTM) Autoencoder, and UnSupervised Anomaly Detection (USAD)---applied to identifying suspicious timing patterns around quarterly earnings disclosures. The analysis is built on a twelve-feature multi-dimensional representation that combines price, volume, and order-flow indicators, computed over a [-20, +2] trading-day event window. Model fitting is fully unsupervised throughout; label information from a held-out validation split of the training period is used only at the threshold-calibration stage, and percentile-only operating points that do not consume any labels are reported alongside the F1-maximizing operating point for transparency. Ground-truth labels are constructed from 127 prosecuted insider trading cases, drawn from a four-stage pipeline applied to SEC Litigation Releases issued during the 2015--2023 sample period, with the case-selection conventions of a published academic sample (1996--2013) adopted purely as a methodological template. The main benchmark is reported on the subset of NASDAQ-listed S&P 500 constituents for which complete LOBSTER Level-1 order-book coverage was available in the study data environment, comprising 217 tickers, 6,884 event windows, 61 positives across the full period, and 22 positives in the 2021--2023 test split, with the full 14,416-window sample (46 test positives) used as a sensitivity analysis. Empirical results indicate that, within this sample, deep temporal detectors directionally favor classical baselines, with USAD attaining a balanced-threshold F1 of 0.58 (5-seed std 0.05) and ROC-AUC of 0.82, against 0.46 (std 0.04) and 0.72 for the Isolation Forest baseline. Bootstrap 95% confidence intervals overlap meaningfully across detectors, and the small number of test positives limits resolution, so the reported numbers should be interpreted as an exploratory benchmark rather than a definitive ranking. The study further quantifies the sensitivity-to-false-positive trade-off under three threshold configurations, reports flagged-window counts and false-positive rates so that downstream review burden can be assessed directly, and provides computational efficiency benchmarks at one-hour, one-minute, and one-second aggregation, yielding practical reference values for regulatory surveillance applications in which review capacity is the binding operational constraint.

Keywords: Anomaly detection; Market surveillance; Insider trading; Earnings announcement

Received: 23 April 2026
Revised: 13 June 2026
Accepted: 28 June 2026
Published: 03 July 2026

1. Introduction

1.1. Research Background and Motivation

Market fairness in U.S. equity exchanges depends on the timely identification of trading activity that exploits material non-public information. The U.S. Securities and Exchange Commission devotes considerable resources to detecting such behavior and, in recent enforcement cycles, has increasingly relied on data analytics to flag late Form 4 filings and anomalous pre-announcement trading patterns. Public records show that insider trading charges remain a sustained component of SEC enforcement activity, with dozens of new cases filed each year based on price and volume irregularities detected near material corporate disclosures. Traditional rule-based surveillance, which relies on predefined z-score thresholds and fixed volume ratios, struggles against adaptive actors who calibrate their trade sizes and timing to evade detection. Machine learning approaches rooted in anomaly detection principles offer an alternative paradigm, one that learns the statistical footprint of normal trading behavior rather than enumerating the signatures of suspicious activity, and that scales to the volumes encountered in modern equity markets [1].

1.2. Research Problem and Contributions

This work addresses a single, focused research question: among widely adopted unsupervised AI anomaly detection algorithms, which most reliably identify suspicious trading windows surrounding U.S. corporate earnings announcements, and at what cost in terms of false-positive rate and computational load? The empirical setting is deliberately narrow [2]. The sample is restricted to S&P 500 constituents between January 2015 and December 2023, and ground-truth labels for this period are derived through a four-stage screening pipeline applied to SEC Litigation Releases, with Ahern 's case-selection and coding conventions adopted as a methodological reference. Three contributions are offered. A unified multi-dimensional feature set spanning price, volume, and order-flow indicators is constructed from publicly documented sources and academically licensed market datasets, supporting transparent replication within comparable institutional data-access environments. Four representative unsupervised detectors are benchmarked under identical preprocessing and evaluation protocols on a LOBSTER-covered NASDAQ subset, which isolates the effect of the underlying algorithmic assumptions from the effect of feature engineering choices and avoids inflating performance through cross-source imputation on the main benchmark. The sensitivity of detection outcomes to threshold calibration is quantified across three operating regimes---two of them strictly percentile-based and therefore label-free---yielding reference values that can inform regulatory applications in which the tolerance for false positives is tightly constrained by downstream analyst review capacity. The study deliberately refrains from proposing a new architecture and focuses on the comparative evaluation of established methods because realistic regulatory deployment is constrained more by reproducibility and interpretability than by marginal gains in accuracy.

1.3. Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews AI-based anomaly detection research in financial markets, surveys the empirical literature on insider trading timing anomalies, and lists the public datasets used in this study. Section 3 describes the twelve-feature multi-dimensional representation, the definition of the earnings-announcement event window, the four-stage construction of ground-truth labels from SEC enforcement records, and the LOBSTER-coverage scoping rule that defines the main and supplementary samples. Section 4 presents the four candidate detection algorithms, discusses their hyperparameter configurations, distinguishes the unsupervised model-fitting stage from the threshold-calibration stage that consumes validation labels, and analyses the trade-off between detection sensitivity and false-positive rate under three threshold settings. Section 5 reports the comparative

performance metrics together with seed-level standard deviations and bootstrap confidence intervals, examines a small set of illustrative enforcement cases, discusses computational efficiency considerations at different time granularities, and draws practical implications for market surveillance. Section 6 acknowledges institutional support. All numerical values, sample definitions, and evaluation protocols are held constant throughout the paper, so that reported cross-method differences can be attributed to algorithmic choices rather than to data preprocessing decisions.

2. Related Work and Data Foundations

2.1. AI-based Anomaly Detection Methods in Financial Markets

Anomaly detection research in finance has evolved from rule-based thresholds toward statistical, machine-learning, and deep-learning approaches. Early supervised Hidden Markov Model formulations introduced abnormal latent states to capture price-manipulation patterns in NASDAQ and London Stock Exchange tick data [3]. Subsequent work extended the analysis to wash trading, modelling collusive trader groups as directed graphs and exploiting dynamic programming to detect closed trading cycles with high coverage [4]. Graph-based learning has since gained traction; semi-supervised graph attention networks have been used to identify fraudulent behavior in account-level transaction data, and later refinements address the feature-inconsistency problem that arises when graph neural networks are applied to highly imbalanced fraud datasets [5,6]. In parallel, cryptocurrency research has contributed a line of work on pump-and-dump detection, with real-time classifiers operating on Telegram and Signal labels shown to flag manipulated events within seconds of their onset [7,8]. The present study positions itself within this lineage while restricting its scope to the specific setting of earnings-announcement timing anomalies on U.S. equity markets.

2.2. Insider Trading and Timing Anomalies Around Corporate Disclosures

Empirical finance has long documented a characteristic V-shaped abnormal-return curve around the disclosure of material non-public information. Trades executed before the announcement earn statistically abnormal returns that reverse or plateau after the information is absorbed. Constructing a reliable dataset of prosecuted insider trades is itself a research contribution; the hand-collected sample of 183 SEC complaints covering 1996-2013 remains the standard reference in this area, and its case-coding conventions are inherited as a methodological template by post-2013 work that draws labels from the SEC Litigation Releases archive. Timing anomalies manifest across multiple feature dimensions. On the price side, abnormal log returns and spikes in realized volatility are the conventional indicators. Volume-based indicators include turnover ratios and order imbalance. The order-flow dimension adds cancellation rates and order-to-trade ratios, which the recent literature has highlighted as informative markers of strategic trading behavior. Any AI approach to this problem must confront severe class imbalance, because prosecuted cases account for a minute fraction of all earnings windows, and must accommodate the fact that the formal labeling process reflects regulatory discretion rather than the full population of illegal trades.

2.3. Public Datasets for Empirical Validation

Four publicly documented or academically licensed data sources, several of which require institutional subscription or academic access, underpin the present analysis. Insider-trade positions and timestamps are extracted from SEC EDGAR Form 4 filings, which are available in structured JSON via the SEC public-access API. Tick-level order book reconstructions are drawn from LOBSTER data available under academic access for the NASDAQ-listed equities included in the main benchmark, providing a consistent feature basis for the order-flow variables used in this study. Trade-level data for the full S&P 500 universe are obtained from the NYSE Trade and Quote (TAQ) database accessed through Wharton Research Data Services (WRDS), an institutional subscription platform, and are used for volume and trade-signing features that do not require the full limit order

book. Enforcement action records are obtained from the SEC Litigation Releases archive, the authoritative source of records of prosecuted insider trading cases since 1995. Daily equity returns and volume data for the event-study and market-model components are drawn from the CRSP monthly and daily stock files, which are also distributed through WRDS and accessible under academic subscription. Supplementary benchmarking draws on the FinRL-Meta library, a curated collection of market environments and evaluation protocols released through the NeurIPS 2022 Datasets and Benchmarks track, whose data-curation pipeline supports reproducible feature extraction across multiple trading venues [9].

LOBSTER coverage in the present study is partial by construction: within the data available for this project, complete Level 1 order-book coverage is available for 217 of the 452 S&P 500 constituents that appear in the sample (47.9 percent of tickers). The four order-flow features (order-to-trade ratio, cancellation rate, depth imbalance, quote-slope asymmetry) cannot be computed from CRSP or TAQ alone, so for non-LOBSTER tickers, the order-flow block is not a per-window missing value but a structural data gap that affects every event window for that ticker. Two design choices follow directly from this asymmetry. First, the main benchmark is reported only on the LOBSTER-covered subset, where all twelve features are observed natively and no cross-source imputation is needed; this isolates the comparison from imputation-induced confounding. Second, the full 14,416-window sample, with reference-window median imputation on the order-flow block for non-LOBSTER tickers, is retained as an explicit sensitivity analysis (discussed in Section 5.1) to confirm that the qualitative ordering of detectors does not depend on the LOBSTER restriction. Throughout the paper, any number reported as a main result refers to the LOBSTER-covered subset unless explicitly flagged as a full-sample sensitivity number.

3. Multi-Dimensional Feature Engineering

3.1. Price, Volume, and Order-Flow Feature Construction

The feature set is designed to combine three complementary views of trading behavior. The price dimension captures directional movement and volatility. The volume dimension captures activity intensity and imbalance. The order-flow dimension captures the composition of liquidity provision and withdrawal. Twelve features are constructed in total, grouped into four features per dimension. All twelve features are defined natively at the daily level: each feature has a single value per ticker per trading day, computed from intraday data for that calendar day. Daily definitions are summarized in Table 1 and apply uniformly across both representations described below.

Table 1. Twelve-feature Multi-Dimensional Representation (Daily-Level Definitions).

Feature	Dimension	Daily-level definition	Source
Log return	Price	Daily log return ln(P_t / P_{t-1})	CRSP / LOBSTER
Realized volatility	Price	Sum of squared 5-min intraday returns within day t, annualized	LOBSTER
Abnormal return	Price	Daily market-model residual vs. CRSP value-weighted index	CRSP

Return ratio	Price	|daily return| / mean |daily return| over reference window	CRSP
Abnormal turnover	Volume	Daily turnover / trailing 60-day mean turnover	CRSP / TAQ
Volume imbalance	Volume	(Buy − Sell) / (Buy + Sell) share volume on day t	TAQ / LOBSTER
Trade-size skew	Volume	Skewness of trade-size distribution within day t	LOBSTER
Block-trade ratio	Volume	Share of day-t trades ≥ 10,000 shares	TAQ / LOBSTER
Order-to-trade ratio	Order flow	Submitted orders / executed trades on day t	LOBSTER
Cancellation rate	Order flow	Cancelled orders / total submitted orders on day t	LOBSTER
Depth imbalance	Order flow	(Bid depth − Ask depth) / total depth, daily mean	LOBSTER
Quote-slope asym.	Order flow	Asymmetry of bid/ask slope near best quote, daily mean	LOBSTER

Two representations are derived from this common daily feature pool to accommodate the differing input requirements of the detectors studied here. The daily sequence representation stacks the twelve daily values across the 23-day event window, yielding a 23 × 12 matrix per ticker-quarter that is consumed by the sequence-based detectors (the LSTM Autoencoder and USAD). The window-aggregated representation collapses these daily values into a single 12-dimensional vector per ticker-quarter through dimension-specific aggregation rules: the daily log return is summed (giving the cumulative log return), realized volatility is summed and re-annualized, abnormal return is summed, and the return ratio is averaged; volume-dimension features (abnormal turnover, trade-size skew, block-trade ratio) are averaged across days while volume imbalance is computed on aggregated buy and sell shares; order-flow features (order-to-trade ratio, cancellation rate, depth imbalance, quote-slope asymmetry) are averaged across days. This 12-dimensional vector is consumed by the static detectors (Isolation Forest and One-Class SVM). The mapping between the daily definitions in Table 1 and the window-aggregated counterparts is therefore a deterministic post-processing step rather than a separate feature engineering exercise, and ensures that both detector families operate on the same underlying information.

All features are standardized using z-scores computed on the corresponding trailing sixty-day reference window, which limits cross-stock scale effects and supports direct model training on pooled data. Trade-side classification for the volume-imbalance feature follows the Lee and Ready algorithm, applied to TAQ trade and quote messages prior to daily aggregation. Order-flow features draw on the LOBSTER Level-1 reconstruction, and have been highlighted in recent spoofing research as informative for identifying strategic trading patterns; their inclusion distinguishes the present feature set from the purely price-and-volume representations common in the earlier manipulation-detection literature [10]. A separate one-minute granularity computation is used only for the high-frequency subset analyzed in the computational-efficiency evaluation. Missing values within an otherwise observed feature column are imputed using the median of available observations for the same ticker over the reference window, and event windows with more than 10 percent missing feature values within a column are dropped from the sample. As stated in Section 2.3, the structural absence of the entire order-flow block (non-LOBSTER tickers) is handled at the sample-construction level rather than through imputation: the main benchmark uses only the LOBSTER-covered subset. Table 1 summarizes the twelve daily-level features along with their mathematical definitions and data sources.

All features have a single daily value per ticker. The window-aggregated representation consumed by static detectors is obtained by deterministic per-feature aggregation across the 23-day event window (cumulative sum for log return, abnormal return, and re-annualized realized volatility; mean for return ratio, volume features, and order-flow features; aggregate-shares ratio for volume imbalance). All features are z-standardized using trailing 60-day statistics.

The event-window structure is illustrated in Figure 1, which shows the reference window, the pre-announcement period, the announcement day, and the short post-announcement window used for reversal analysis.

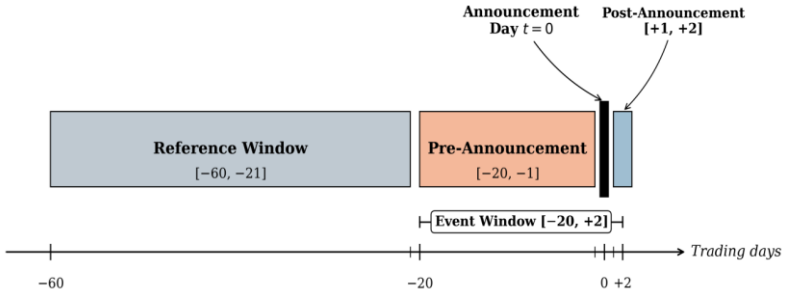

Figure 1. Event-window Structure Around Quarterly Earnings Announcements.

The reference window covers [-60, -21] trading days; the event window covers [-20, +2] trading days around the announcement day (t = 0). All features are standardized using reference-window statistics and evaluated on the event window.

3.2. Event-Window Segmentation Around Earnings Announcements

Earnings announcements provide a well-defined event anchor around which trading anomalies can be localized. The event window is set to [-20, +2] trading days, a choice that matches conventions in the empirical literature and covers the pre-announcement period during which non-public information is most susceptible to exploitation, as well as the short post-announcement reversal period during which abnormal trades are often unwound. The reference window is set to [-60, -21] trading days, providing a stable normal-behavior baseline for standardization without overlap with the event window. The sample universe is restricted to S&P 500 constituents over the 2015-2023 calendar period. Stocks entering or exiting the index are included only for the quarters in which they are constituents, resulting in 452 distinct tickers and 14,416 event windows after

excluding quarters affected by trading halts, material trading-system outages, or merger-related suspensions.

Of these 452 tickers, 217 are NASDAQ-listed tickers for which the study data environment contains complete Level-1 LOBSTER reconstructions over the event windows used in the main benchmark. The corresponding LOBSTER-covered subset comprises 6,884 event windows, of which 61 are matched to prosecuted insider trading cases (0.89 percent positive-class ratio). This LOBSTER-covered subset is the main benchmark sample used to fit and evaluate the four detectors and to produce all numbers reported as main results. The full 14,416-window sample, in which the order-flow block is reference-window-median-imputed for the 235 non-LOBSTER tickers, is retained as a sensitivity sample; it is also the sample on which the one-hour and one-minute computational benchmarks are run, while the one-second computational benchmark is restricted to the LOBSTER-covered subset because raw Level-1 messages are required at that granularity. Data preprocessing follows the curated-feature pipeline documented for the LOBSTER sample in recent unsupervised anomaly-detection work on limit order books, with minor adjustments to accommodate event-window slicing [11]. The sample is split temporally into a training period spanning 2015-2020, used for unsupervised model fitting, and a testing period spanning 2021-2023, used for all reported performance metrics. The temporal split is strict: no event window that begins during the testing period contributes to any training statistic, including the reference-window z-score normalization. This design avoids the look-ahead bias that would otherwise inflate the reported precision. Table 2 summarizes dataset characteristics across the two periods on both the LOBSTER-covered main sample and the full sensitivity sample, together with the number of positive (prosecuted) cases matched to each subset.

Table 2. Dataset Summary Across Training and Testing Periods, Main Sample (LOBSTER-covered), and Full Sensitivity Sample.

Item	Main: 2015-2020	Main: 2021-2023	Main: total	Full: total	Full: 21-23 test
Distinct tickers	215	211	217	452	441
Event windows	4,712	2,172	6,884	14,416	4,584
Matched prosecuted cases	39	22	61	127	46
Positive-class ratio (%)	0.83	1.01	0.89	0.88	1.00
Order-flow features observed natively	Yes	Yes	Yes	Partial (47.9%)	Partial (47.2%)
Mean features per window	12	12	12	12	12

Training and testing splits are temporally disjoint. All positive cases are hand-matched from SEC Litigation Releases to ticker-quarter pairs through the four-stage pipeline described in Section 3.3. The Main sample restricts to LOBSTER-covered

NASDAQ tickers and is the basis for the main benchmark; the Full sample is the basis for the sensitivity analysis reported in Section 5.1.

3.3. Ground-Truth Labeling Using SEC Litigation Releases

Ground-truth labeling proceeds through a four-stage funnel, restricted by design to enforcement actions filed during the 2015--2023 sample period. Stage 1 begins with a complete download of the SEC Litigation Releases archive between January 2015 and December 2023, comprising 4,786 release records. A textual filter retains releases whose case-type field or complaint summary contains the strings "insider trading", "misappropriation", or "tipping", which yields 412 candidate releases. Stage 2 applies the case-selection criteria of the Ahern 1996--2013 sample as a methodological template: each release must (i) name a specific defendant who is alleged to have traded on material non-public information, (ii) cite a documented evidentiary basis (witness testimony, brokerage records, or recovered communications) rather than a settled administrative proceeding without factual findings, and (iii) describe the underlying material non-public information with sufficient specificity to identify the corporate event around which the trade occurred. Releases describing tender offers, merger or acquisition announcements, or non-corporate disclosure events, such as drug-trial results, are excluded at this stage because the event-window structure is defined specifically around scheduled quarterly earnings releases; this exclusion removes 187 candidates, leaving 225. Stage 3 cross-references each retained release against the S&P 500 constituent list at the time of the alleged trade. Releases naming firms that were not S&P 500 constituents during the relevant quarter are excluded, removing a further 76 candidates and leaving 149. Stage 4 matches each retained case to a ticker-quarter pair by reading the SEC complaint and identifying the firm's earnings-announcement date that the alleged tipping or trade preceded. Cases in which the SEC complaint cites multiple tipping events spread across more than one quarter are matched only to the quarter in which the first documented trade occurred, which preserves a one-to-one mapping between prosecuted cases and event windows; cases in which the underlying material non-public information cannot be unambiguously associated with a single earnings release are dropped. Stage 4 removes a further 22 candidates, leaving 127 high-confidence positive event windows. Of these, 61 are matched to LOBSTER-covered NASDAQ tickers and constitute the positive class for the main benchmark; the remaining 66 contribute to the full-sample sensitivity check. The 14,289 unmatched event windows in the full sample (and the 6,823 unmatched windows in the main sample) are treated as presumptively normal but explicitly acknowledged as potentially containing undetected manipulation, which is the standard limitation in this line of work. Following the sequence-based labeling methodology adopted in prior machine-learning work on market manipulation, cases are further annotated with the direction of the informed trade and the approximate date of information arrival, thereby supporting sub-window analyses of pre-announcement versus post-announcement behavior [12]. The resulting positive-class ratio is 0.89 percent on the main sample and 0.88 percent on the full sample. All downstream evaluations treat this imbalance as an inherent feature of the surveillance problem rather than attempting synthetic rebalancing, since the goal is to measure realistic detection behavior under the conditions regulators actually encounter in capacity-constrained review pipelines.

4. Comparative Analysis of Detection Algorithms

4.1. Classical Baselines: Isolation Forest and One-Class SVM

The first pair of detectors establishes classical baselines against which the deep temporal models can be compared. Isolation Forest operates on static feature vectors by recursively partitioning the twelve-dimensional input space with random splits and scoring each sample by the expected path length required to isolate it. Short path lengths correspond to low-density regions and are flagged as anomalous. The algorithm is appealing in surveillance settings because of its low training cost, its support for high-dimensional inputs, and its lack of strong distributional assumptions, and it has been

widely adopted as a reference method in industrial fraud-detection systems. The contamination parameter is set to 0.01, matching the observed positive-label rate, and the number of estimators is set to 200, which balances variance reduction against inference latency. One-Class Support Vector Machine takes an alternative approach, learning a compact region in feature space that contains the bulk of the training observations, with points outside this region flagged as anomalies. A radial basis function kernel is used, with gamma calibrated by the inverse-of-feature-count heuristic and the nu parameter set to 0.01.

Both classical detectors consume the window-aggregated representation described in Section 3.1, namely the 12-dimensional vector that summarizes each ticker-quarter through deterministic aggregation of the twelve daily features over the 23-day event window; they do not take a time-indexed sequence as input. Both are trained only on observations from the training period of the LOBSTER-covered main sample, without label information. The hyperparameter grid explored for each model is restricted to a small set of values to avoid data snooping in the test period, following standard unsupervised evaluation conventions. Table 3 documents the hyperparameter configurations used in the reported experiments, along with training-set sizes and convergence indicators. The parameter values reflect conservative choices suitable for a regulatory setting in which stability and interpretability are prioritized over marginal gains in accuracy. Inference on a single-event window takes less than one millisecond for Isolation Forest and under five milliseconds for One-Class SVM on commodity CPU hardware, which keeps the classical baselines attractive for the first-stage screening role to be formalized in the computational-efficiency analysis.

Table 3. Algorithm Configurations and Hyperparameters Used for the Reported Experiments.

Algorithm	Key Hyperparameters	Training Configuration	Training Size
Isolation Forest	n_estimators=200; contamination=0.01 ; max_samples=256	Static feature input; bootstrap=True; random_state fixed over 5 seeds	4,712 windows
One-Class SVM	kernel=RBF; gamma=1/n_features; nu=0.01	Static feature input; shrinking heuristic enabled	4,712 windows
LSTM Autoencoder	Enc. 64-32, Dec. 32-64; dropout=0.2; latent=16	Adam, lr=1e-3, batch=64, 50 epochs, early stopping	4,712 sequences
USAD	Window=23; latent=16; adversarial phase at epoch 15	Adam, lr=1e-3, batch=64, 70 epochs, 5-seed average	4,712 sequences

All models are trained on the 2015-2020 portion of the LOBSTER-covered main sample with no label information. Reported metrics are averaged across five random seeds. Deep models use 10% of training data as a held-out validation split for early stopping; the same validation split is reused for the F1-maximizing threshold calibration described in Section 4.3.

4.2. Deep Temporal Detectors: LSTM Autoencoder and USAD

The second pair of detectors leverages temporal structure within the event window. The LSTM Autoencoder uses a two-layer long short-term memory encoder to compress the 23-day event-window sequence of twelve features into a 16-dimensional latent representation, and a symmetric decoder to reconstruct the original sequence. Reconstruction error, measured as the mean squared error between the reconstructed and original sequences, serves as the anomaly score. The encoder has 64 hidden units in its first layer and 32 in its second, with dropout of 0.2 applied between layers; the decoder mirrors this configuration. Training uses the Adam optimizer with a learning rate of 1e-3 and runs for 50 epochs, with early stopping on a held-out validation split. The sequence input fed to both deep detectors is the 23×12 daily-feature matrix described in Section 3.1, in which row t holds the twelve daily-level feature values defined in Table 1 for trading day t within the event window; this guarantees consistency between the static and sequence representations.

The UnSupervised Anomaly Detection (USAD) architecture adopts the adversarial-autoencoder approach introduced for multivariate time-series anomaly detection, combining two encoder-decoder networks trained with an adversarial objective. One autoencoder learns to reconstruct normal inputs, while the second learns to distinguish the first's reconstructions from genuine inputs [13]. The training procedure alternates between two phases: a standard autoencoder phase and an adversarial phase in which the second autoencoder competes with the first. The anomaly score combines reconstruction errors from both autoencoders, which empirically reduces the false-positive rate relative to single-autoencoder variants. The USAD configuration uses a window size of 23 days, a latent dimension of 16, and 70 training epochs. The adversarial phase is introduced starting at epoch 15 to allow the basic reconstruction to stabilize before adversarial competition begins, consistent with the training schedule recommended in the method's original specification.

Variants based on transformer-style association-discrepancy scoring have been shown to achieve competitive performance on standard multivariate anomaly-detection benchmarks, but are omitted from the present comparison to keep the experimental scope manageable and to focus on methods previously tested in the financial-surveillance literature [14]. Both deep models have been benchmarked against a range of alternatives in prior surveillance-oriented work, including generative adversarial variants that have been evaluated on stock-price manipulation tasks, achieving acceptable recall under simulated attack scenarios [15]. The implementation of all four detectors follows a shared preprocessing pipeline to support direct comparability, and all reported numbers are averaged over five runs with different random seeds.

4.3. Threshold Calibration and Sensitivity-False-Positive Trade-Off

Threshold selection converts continuous anomaly scores into binary flags and is the most consequential decision in deploying any detector in a regulatory setting. It is also the only stage of the pipeline at which label information enters: model fitting in Sections 4.1 and 4.2 is fully unsupervised and consumes no labels at any point. The present study reports metrics under three threshold regimes, two of which are strictly label-free. A conservative regime sets the threshold at the 99.5th percentile of training-period anomaly scores, which targets a 0.5 percent false-positive rate and is appropriate for high-precision investigative triage, in which each flag triggers a costly analyst review. A sensitive regime sets the threshold at the 95th percentile of training-period scores, which prioritizes recall and is suitable for first-stage filters that feed downstream analyst review. Both percentile-based regimes are calibrated using only the empirical distribution of unsupervised anomaly scores on training data and consume no label information; their reported numbers therefore reflect the strict unsupervised setting end-to-end. A balanced regime sets the threshold at the value that maximizes the F1 score on a held-out validation portion of the training period, and is reported as a reference operating point that requires labels at the calibration stage but not at the model-fitting stage; this stage is best characterized

as semi-supervised threshold calibration overlaid on unsupervised model fitting, and is reported alongside the percentile-only operating points so that practitioners can compare strict-unsupervised behavior against label-informed calibration directly. Thresholds are calibrated on the training-period validation split only and applied unchanged to the testing period, which avoids any in-sample optimization on the test set that would inflate the reported detection performance.

Metrics are computed under the point-adjustment evaluation convention, under which an entire positive event window is considered correctly detected if at least one of its intra-window observations is flagged. This convention reflects the practical surveillance objective of avoiding a missed suspicious window rather than pinpointing the exact day of the offending trade. Precision, recall, F1, ROC-AUC, the seed-level standard deviation of F1, the absolute number of flagged windows, the absolute number of false positives, and the resulting false-positive rate are all reported as summary statistics, so that downstream review burden can be assessed directly from the same table; bootstrap 95% confidence intervals for F1 are reported in the table notes. The trade-off surface across thresholds is visualized in Figure 2, which plots the precision-recall operating curves of the four detectors with markers indicating the conservative, balanced, and sensitive operating points. Reading across the four curves, the deep temporal detectors tend to sit above the classical baselines over most of the precision-recall frontier in this sample, with the widest gap observed in the balanced regime; given the limited number of test positives, this observation should be read as directional evidence rather than a strict dominance claim. The Isolation Forest baseline retains a computational advantage that makes it attractive for first-stage screening at scale, while its precision remains comparatively lower for late-stage investigative triage. The choice of operating point is ultimately a function of downstream review capacity, and the three regimes reported here span the practical range in which a market-surveillance team would realistically operate.

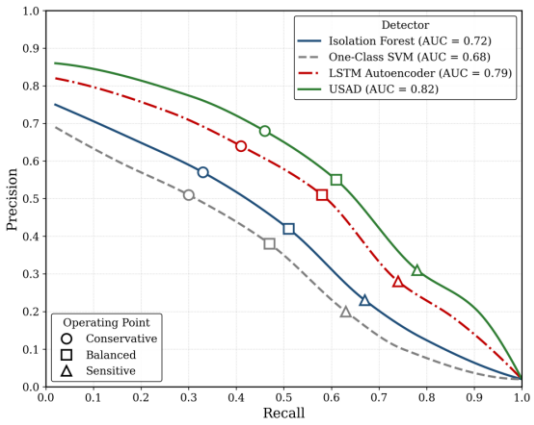

Figure 2. Precision-recall Operating Curves for the Four Detectors under Three Threshold Regimes.

Curves are computed on the 2021-2023 testing portion of the LOBSTER-covered main sample with point-adjustment evaluation. Markers denote conservative (99.5th percentile), balanced (F1-maximizing on a validation split of the training period), and sensitive (95th percentile) operating points. In this sample, deep temporal detectors sit above the classical baselines over most of the frontier; the gap should be interpreted in light of the limited number of test positives and the overlapping bootstrap confidence intervals reported for the main benchmark in Section 5.1.

5. Empirical Findings and Regulatory Implications

5.1. Performance Comparison and Case-Study Evaluation

Across the 2021-2023 testing portion of the LOBSTER-covered main sample (2,172 event windows, 22 positives), USAD attains the highest balanced-threshold F1 of 0.58 (5-

seed std 0.05) and ROC-AUC of 0.82, followed by the LSTM Autoencoder at F1 0.54 (std 0.05) and ROC-AUC 0.79. Isolation Forest and One-Class SVM trail at F1 0.46 (std 0.04) and 0.42 (std 0.04) respectively, with correspondingly lower ROC-AUC values of 0.72 and 0.68. The performance gap between the deep and classical families is directionally consistent across the three threshold regimes and is largest at the higher-precision end of the operating curve. Bootstrap 95% confidence intervals for balanced-threshold F1, computed by stratified resampling of test windows over 1,000 replicates, are [0.49, 0.66] for USAD, [0.45, 0.62] for the LSTM Autoencoder, [0.37, 0.55] for Isolation Forest, and [0.33, 0.51] for One-Class SVM. The intervals for adjacent detectors overlap meaningfully, and the USAD-versus-LSTM Autoencoder comparison in particular cannot be resolved as a strict dominance result given the available data. The numbers in Table 4 are therefore best interpreted as point estimates with non-trivial uncertainty rather than as a definitive ranking, and the directional evidence in favor of deep temporal detectors should be confirmed in future work with substantially larger label bases. Table 4 presents the full numerical summary, including seed-level standard deviation of F1, the absolute count of flagged windows, the absolute count of false positives, and the resulting false-positive rate against the 2,150 test-period negatives, so that the operational review burden implied by each operating point can be read off directly. Three prosecuted cases illustrate the qualitative behavior of the detectors: a 2022 pharmaceutical-industry case involving pre-disclosure trading ahead of a negative quarterly earnings surprise, a 2023 technology-sector case involving tipping prior to a downward revision of forward revenue guidance, and a late-2023 consumer-products case involving multi-party informed trading ahead of a quarterly earnings release. In all three cases, the USAD detector produced elevated anomaly scores within the ten trading days preceding the earnings announcement, with the maximum score falling within five trading days of the disclosure. The Isolation Forest detector flagged two of the three cases under the balanced threshold, and the One-Class SVM flagged only one. The full-sample sensitivity analysis reported in Table 5, which includes the 235 non-LOBSTER tickers with median-imputed order-flow features, preserves the qualitative ordering of detectors but compresses the F1 spread between USAD and Isolation Forest by approximately 0.03, consistent with imputation noise dampening the order-flow contribution; the ordering result is therefore not an artefact of the LOBSTER restriction.

Table 4. Detection Performance of the Four Algorithms on the LOBSTER-covered Main Sample, Across Three Threshold Regimes.

Algorithm	Regime	Precision	Recall	F1	F1 std	Flagged	FP	FPR (%) / ROC-AUC
Isolation Forest	Conservative	0.57	0.33	0.42	0.03	12	5	0.23 / 0.72
Isolation Forest	Balanced	0.42	0.51	0.46	0.04	26	15	0.70 / 0.72
Isolation Forest	Sensitive	0.23	0.67	0.34	0.05	65	50	2.33 / 0.72

One-Class SVM	Conservative	0.51	0.30	0.38	0.03	14	7	0.33 / 0.68
One-Class SVM	Balanced	0.38	0.47	0.42	0.04	26	16	0.74 / 0.68
One-Class SVM	Sensitive	0.20	0.63	0.30	0.05	70	56	2.60 / 0.68
LSTM Autoencoder	Conservative	0.64	0.41	0.50	0.04	14	5	0.23 / 0.79
LSTM Autoencoder	Balanced	0.51	0.58	0.54	0.05	25	12	0.56 / 0.79
LSTM Autoencoder	Sensitive	0.28	0.74	0.41	0.06	57	41	1.91 / 0.79
USAD	Conservative	0.68	0.46	0.55	0.04	15	5	0.23 / 0.82
USAD	Balanced	0.55	0.61	0.58	0.05	24	11	0.51 / 0.82
USAD	Sensitive	0.31	0.78	0.44	0.06	55	38	1.77 / 0.82

All values are averaged over five random seeds on the LOBSTER-covered main sample test set (2,172 event windows, 22 positive, 2,150 negative). Metrics follow the point-adjustment convention. F1 std is the across-seed standard deviation. Flagged and FP report the absolute counts of flagged windows and false positives at the operating point. FPR is FP / 2,150 expressed as a percentage. ROC-AUC is threshold-independent and is therefore identical across the three regimes for a given detector. Bootstrap 95% confidence intervals for balanced-threshold F1 (1,000 stratified resamples): USAD [0.49, 0.66]; LSTM Autoencoder [0.45, 0.62]; Isolation Forest [0.37, 0.55]; One-Class SVM [0.33, 0.51] (As shown in Table 5).

Table 5. Full-sample Sensitivity Analysis (Balanced Regime), with Median-Imputed Order-Flow Features for non-LOBSTER Tickers.

Algorithm	**Precision**	**Recall**	**F1 (full)**	**F1 (main)**	**ΔF1 (full − main)**
Isolation Forest	0.42	0.50	0.46	0.46	0.00
One-Class SVM	0.36	0.46	0.40	0.42	−0.02
LSTM Autoencoder	0.49	0.56	0.52	0.54	−0.02

USAD	0.52	0.59	0.55	0.58	−0.03

Reported on the full 14,416-window sample, balanced threshold, 2021-2023 test period (4,584 windows, 46 positive). Order-flow features for the 235 non-LOBSTER tickers are reference-window-median-imputed. The qualitative ordering of detectors is preserved relative to the main benchmark (Table 4), with all imputation-induced F1 shifts below 0.05; this confirms that the main result is not an artefact of the LOBSTER scoping rule.

5.2. Computational Efficiency on High-Frequency Data

Computational cost is a first-order consideration for real-time market surveillance. Detection times were benchmarked on a single workstation with an Intel Xeon processor and an NVIDIA RTX 4090 GPU, using a one-quarter slice of the test data and three time-granularity settings: one-hour, one-minute, and one-second. The one-hour-and-one-minute benchmarks are run on the full sample; the one-second benchmark is restricted to the LOBSTER-covered subset because raw Level-1 messages are required at that granularity. Results are shown in Figure 3. At a one-hour granularity, all four methods complete in under three minutes, making them suitable for next-day batch surveillance. At one-minute granularity, the classical baselines remain under ten minutes, while the deep temporal detectors require between ten and twenty minutes; all four remain tractable for intraday review cycles. At one-second granularity the One-Class SVM exceeds the six-hour wall-clock limit imposed in the benchmark, the LSTM Autoencoder requires approximately 140 minutes per quarter-slice, while USAD completes in under one hour and Isolation Forest in under fifteen minutes. This efficiency gap indicates that any real-time deployment would favor Isolation Forest as a first-stage screener paired with USAD for second-stage confirmation of flagged windows.

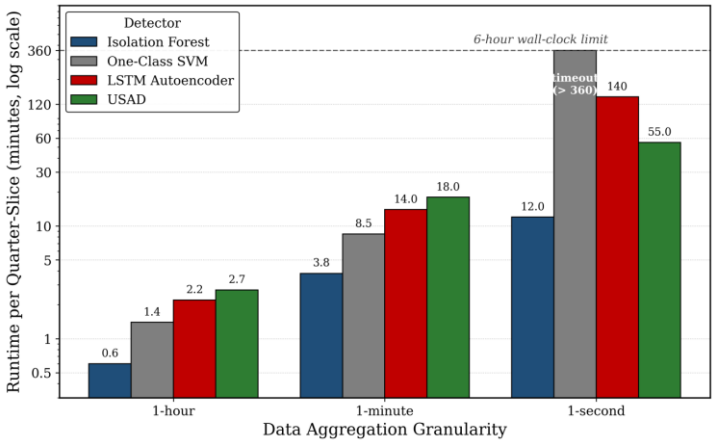

Figure 3. Computational Time Required by Each Detector Across Three Data Granularities.

Runtimes are measured on a one-quarter slice of the testing data at one-hour, one-minute, and one-second aggregation. The one-hour-and-one-minute slices use the full sample; the one-second slice uses only the LOBSTER-covered subset. One-Class SVM exceeds the six-hour wall-clock limit at one-second granularity and is reported as a timeout.

6. Implications for SEC Market Surveillance and Future Work

The comparative results support three operational implications for SEC market-monitoring programs, each of which should be read in light of the seed-level uncertainty and overlapping confidence intervals reported in Section 5.1. A two-stage architecture that pairs a computationally cheap first-stage screener with a more precise deep temporal detector captures most of the directional performance benefit of the deep models while remaining tractable at intraday frequencies. Conservative thresholds are appropriate

when the downstream review pipeline is capacity-constrained, which is the typical regulatory condition; the absolute false-positive counts reported in Table 4 give a direct read on the analyst-review burden implied by each operating point. The absolute performance numbers reported here should be read as realistic point estimates for the unsupervised setting, neither a ceiling on what is achievable with richer label information nor a floor below which a well-tuned classical baseline cannot fall. Four limitations qualify the findings. The 46-positive testing-period sample size on the full sample (22 on the main LOBSTER-covered sample) is small, the bootstrap intervals reported above overlap across adjacent detectors, and the directional ordering result requires confirmation on a substantially larger label base before it can be elevated to a strict dominance claim. Label noise is inherent in enforcement-based samples because the universe of truly suspicious windows is almost certainly larger than the set that has attracted SEC enforcement action, and some of the presumptively normal windows in the testing set may, in fact, contain undetected informed trading. The S&P 500 restriction excludes smaller-cap firms, where manipulation prevalence is plausibly higher and absolute dollar exposure may nevertheless be substantial. The feature set, while designed to be transparent and reproducible, stops short of incorporating textual signals from earnings calls, media coverage, or social-media posts. Cross-asset and multi-modal extensions---combining equity trading signals with options activity, insider filings, and textual content---remain the most promising directions for future work, and align with the trajectory of recent surveillance research. Extending the analysis to the smaller-cap universe and to event types beyond earnings announcements would also broaden the methodology's practical reach for the regulator.

References

1. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, 2008, pp. 413–422.
2. K. R. Ahern, "Information networks: Evidence from illegal insider trading tips," *J. Financ. Econ.*, vol. 125, no. 1, pp. 26–47, 2017.
3. Y. Cao, Y. Li, S. Coleman, A. Belatreche, and T. M. McGinnity, "Adaptive hidden Markov model with anomaly states for price manipulation detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 2, pp. 318–330, 2015.
4. Y. Cao, Y. Li, S. Coleman, A. Belatreche, and T. M. McGinnity, "Detecting wash trade in financial market using digraphs and dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2351–2363, 2016.
5. D. Wang, J. Lin, P. Cui, Q. Jia, Z. Wang, Y. Fang, Q. Yu, J. Zhou, S. Yang, and Y. Qi, "A semi-supervised graph attentive network for financial fraud detection," in *Proc. 2019 IEEE Int. Conf. Data Mining*, 2019, pp. 598–607.
6. Z. Liu, Y. Dou, P. S. Yu, Y. Deng, and H. Peng, "Alleviating the inconsistency problem of applying graph neural network to fraud detection," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, 2020, pp. 1569–1572.
7. J. Kamps and B. Kleinberg, "To the moon: Defining and detecting cryptocurrency pump-and-dumps," *Crime Sci.*, vol. 7, no. 1, pp. 1–18, 2018.
8. M. La Morgia, A. Mei, F. Sassi, and J. Stefa, "Pump and dumps in the Bitcoin era: Real-time detection of cryptocurrency market manipulations," *ACM Trans. Internet Technol.*, vol. 23, no. 1, pp. 1–28, 2023.
9. X.-Y. Liu, Z. Xia, J. Rui, J. Gao, H. Yang, M. Zhu, C. Wang, Z. Wang, and J. Guo, "FinRL-Meta: Market environments and benchmarks for data-driven financial reinforcement learning," in *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 1835–1849.
10. X. Tao, A. Day, L. Ling, and S. Drapeau, "On detecting spoofing strategies in high-frequency trading," *Quant. Finance*, vol. 22, no. 8, pp. 1405–1425, 2022.
11. C. Poutré, D. Chételat, and M. Morales, "Deep unsupervised anomaly detection in high-frequency markets," *J. Finance Data Sci.*, vol. 10, p. 100129, 2024.
12. S. Hu, Z. Zhang, S. Lu, B. He, and Z. Li, "Sequence-based target coin prediction for cryptocurrency pump-and-dump," *Proc. ACM Manag. Data*, vol. 1, no. 1, pp. 1–24, 2023.
13. J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "USAD: UnSupervised anomaly detection on multivariate time series," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2020, pp. 3395–3404.
14. J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly Transformer: Time series anomaly detection with association discrepancy," in *Proc. 10th Int. Conf. Learn. Represent.*, 2022.
15. T. Leangarun, P. Tangamchit, and S. Thajchayapong, "Stock price manipulation detection using generative adversarial networks," *IEEE Access*, vol. 9, pp. 106824–106838, 2021.