Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Machine Learning-Driven Early Warning Analytics for Identifying Market Manipulation, Irregular Trading Activity, and Suspicious Market Signals in U.S. Stock Markets

Nusrat Jahan
University of Bridgeport, Analytics and Systems
Email: njahan@my.bridgeport.edu
ORCID: https://orcid.org/0009-0008-0684-1114

Anika Anjum Pritty
Murray State University, Accountancy and Analytics
Email: apritty@murraystate.edu
ORCID: https://orcid.org/0009-0008-2807-0138

Md Ibrahim
University of New Haven, Business Analytics
Email: mibra4@unh.newhaven.edu
ORCID: https://orcid.org/0009-0008-2835-9871

Muhaimin Ul Zadid
University of New Haven, Business Analytics.
Email: mzadi1@unh.newhaven.edu
ORCID: https://orcid.org/0009-0008-2743-4104

A S M FAHIM
University of New Haven, Finance and Financial Analytics
Email: asmfahim987@gmail.com
ORCID: https://orcid.org/0009-0006-3777-5688

Sakib Mahmud
Rutgers, The State University of New Jersey, Business Analytics
Email: Sakib201064@gmail.com
ORCID: https://orcid.org/0009-0000-9765-0365
Corresponding Author: Nusrat Jahan, **E-mail**: njahan@my.bridgeport.edu

| ABSTRACT

This paper develops a machine learning-driven early warning framework for identifying market manipulation, irregular trading activity, and suspicious market signals in U.S. stock markets, while grounding the argument in public regulatory evidence rather than purely illustrative examples. The study is motivated by three practical problems. First, conventional rule-based surveillance remains necessary but often produces large alert volumes, high false-positive burdens, and limited capacity to prioritize weak but meaningful signals. Second, suspicious trading increasingly spans multiple channels, including order flow, cross-venue behavior, public narratives, and relational patterns that are poorly captured by isolated tabular indicators. Third, the strongest public evidence available to researchers is fragmented across enforcement summaries, disciplinary statistics, and whistleblower reporting, which means rigorous early warning research must integrate official sources even when transaction-level labels remain limited. Accordingly, the manuscript combines scholarly synthesis with descriptive evidence derived from Securities and Exchange Commission and Financial Industry Regulatory Authority datasets and reports covering the 2021-2023 period.

| KEYWORDS

Market manipulation; abnormal trading; suspicious market signals; U.S. stock markets; machine learning; anomaly detection; graph neural networks; surveillance analytics; explainable AI

| ARTICLE INFORMATION

ACCEPTED: 02 March 2024 **PUBLISHED:** 06 April 2024 **DOI:** 10.32996/jcsts.2024.6.2.26

Introduction

U.S. stock markets are among the most liquid, information-rich, and technologically sophisticated financial systems in the world. They also remain vulnerable to manipulation, irregular trading activity, and suspicious market signals that can distort prices, erode investor confidence, and impose real costs on households, institutions, issuers, and regulators. Classical enforcement and surveillance frameworks have long recognized practices such as pump-and-dump schemes, spoofing, layering, wash trading, marking the close, rumor-based manipulation, and insider trading as material threats to market integrity (Allen & Gale, 1992; Aggarwal & Wu, 2006; Putnins, 2012). Yet the operational environment in which these practices occur has changed dramatically. U.S. markets now generate enormous streams of quote updates, order submissions, cancellations, execution records, news releases, alternative data, and social-media content, while high-speed routing, fragmentation across venues, and algorithmic trading compress the time available for human review (SEC, 2020; FINRA, 2023). As a result, surveillance systems can no longer rely only on static rules or isolated account-level screens.

The central challenge is not simply that more data exist. It is that suspicious conduct is increasingly multimodal, relational, and adaptive. A manipulative strategy may leave weak traces in a single time series but become visible when analysts jointly consider intraday order-book imbalance, bursts of retail message-board attention, abrupt cross-venue volume migration, unusual options activity, peer-firm co-movements, or recurrent interaction patterns among accounts and securities. This problem has pushed both researchers and market supervisors toward analytics that can learn complex dependence structures rather than merely thresholding individual indicators. In parallel, machine learning has transformed empirical finance by improving nonlinear prediction, extracting signals from unstructured text, and modeling dynamic interactions among entities through networks and graph neural networks (Gu, Kelly, & Xiu, 2020; Dixon, Halperin, & Bilokon, 2020; Sezer, Gudelek, & Ozbayoglu, 2020).

Despite that progress, the literature still shows an important gap. A large body of research predicts returns, volatility, liquidity, or order-flow toxicity, while another body studies fraud, insider trading, and manipulation. These streams do not always meet in an integrated early warning framework tailored to U.S. stock-market surveillance. Many forecasting papers optimize for directional accuracy or trading profitability rather than detection utility. Many surveillance papers remain rule-driven, narrowly tactic-specific, or weak on cross-source data fusion. Moreover, a practical early warning system must do more than classify historical events. It must rank alerts, explain why a pattern appears suspicious, adapt to concept drift, reduce false positives, and preserve procedural fairness for affected market participants. Those requirements create a need for design principles that connect finance theory, market microstructure, anomaly detection, graph learning, and explainable artificial intelligence.

This paper addresses that need by developing a machine learning-driven early warning analytics framework for identifying market manipulation, irregular trading activity, and suspicious market signals in U.S. stock markets. Rather than claiming to solve every detection problem with a single model, the paper argues for a layered architecture. The first layer captures temporal abnormalities in prices, spreads, depths, volumes, message traffic, and order behavior. The second layer models relational structures linking traders, instruments, brokers, issuers, news topics, and venues. The third layer fuses supervised risk estimation with unsupervised anomaly scores and graph-based embeddings to prioritize cases for human review. The final governance layer supports interpretability, fairness auditing, documentation, and escalation.

The contribution is fourfold. First, the paper synthesizes regulatory sources published to organize the fragmented literature around a unified U.S. market-surveillance problem. Second, it proposes a multimodal methodological blueprint that combines transaction analytics, natural language processing, and temporal graph learning. Third, it explains how model evaluation should be aligned with surveillance objectives such as precision at review capacity, time-to-detection, and investigator usefulness, not only aggregate classification accuracy. Fourth, it discusses implementation risks, including data leakage, label scarcity, adversarial adaptation, privacy concerns, and explainability failures. By framing market manipulation detection as an early warning problem instead of a narrow ex post classification exercise, the paper offers a more realistic path toward adaptive, scalable, and trustworthy surveillance in U.S. equity markets.

The broader implication is that analytics can strengthen market integrity, improve deterrence, reduce investigative noise, and help regulators and firms focus scarce surveillance resources on higher-risk behavior before suspicious episodes escalate into investor harm or market disruption.

Literature Review

The literature relevant to an early warning system for suspicious trading in U.S. equity markets spans at least five connected domains: theories of manipulation and market abuse, market microstructure and abnormal trading measurement, machine learning for financial prediction and anomaly detection, graph-based learning, and explainable or responsible artificial intelligence. Reviewing these streams together is important because market surveillance failures rarely arise from a missing statistic. They arise when systems cannot jointly represent strategic behavior, noisy data, temporal evolution, and institutional decision constraints.

Foundational finance research established that manipulation is economically meaningful, not merely a legal label. Allen and Gale (1992) showed how stock prices can be distorted under asymmetric information when traders exploit feedback effects and informational frictions. Aggarwal and Wu (2006) provided large-sample evidence on stock market manipulations and documented the role of publicity, volume surges, and price reversals. Putnins (2012) synthesized theoretical and empirical findings across manipulation forms, clarifying why detection is difficult when manipulators exploit legal trading mechanisms to create misleading appearances. Related work on insider trading and suspicious preannouncement activity underscored that abnormal returns or volumes often become informative only when compared against event timing, peer firms, or historical baselines rather than judged in isolation. This literature suggests that surveillance must be contextual and probabilistic.

Another stream emphasizes market design and surveillance institutions. Cumming, Johan, and Li (2011) linked exchange trading rules to liquidity and integrity outcomes, reinforcing the idea that detection operates within a rule and venue architecture. SEC staff analyses of algorithmic trading highlighted the dual character of automation: algorithms can improve market quality under normal conditions while also intensifying rapid disruptions or unusual feedback loops during stress (SEC, 2020). FINRA’s manipulative trading guidance further illustrates the diversity of prohibited patterns, including layering, spoofing, marking the close, prearranged trading, and rumor-driven practices (FINRA, 2023). Together, these studies indicate that suspicious market signals are often strategy-specific, yet their detection requires platform-wide monitoring.

Market microstructure research provides the measurable ingredients of such monitoring. Easley, Lopez de Prado, and O’Hara (2012) proposed volume-synchronized probability of informed trading as a way to infer order-flow toxicity in high-frequency settings. Studies of spreads, depth depletion, order imbalance, cancellation intensity, and trade clustering show that abnormal trading may emerge through changes in liquidity provision, not only prices. For surveillance, this is crucial. Manipulation frequently aims to alter beliefs about supply, demand, or information quality rather than simply to move the last traded price. The microstructure literature therefore motivates feature engineering around order placement behavior, quote revision speed, order-to-trade ratios, cross-venue fragmentation, closing-auction concentration, and intraday state transitions.

A separate but complementary literature developed statistical and machine-learning approaches for financial data. Gu et al. (2020) demonstrated that machine learning can uncover nonlinear asset-pricing relationships beyond traditional factor models. Fischer and Krauss (2018) showed that deep learning methods such as long short-term memory networks can capture temporal structure in stock-related series. Sezer et al. (2020) reviewed deep-learning applications to financial time series and emphasized both their predictive promise and sensitivity to nonstationarity. Lopez de Prado (2018) argued that financial machine learning must explicitly confront labeling noise, leakage, backtest overfitting, and regime dependence. Although much of this literature focuses on return prediction, it contributes techniques relevant to suspicious trading detection: sequence modeling, nonlinear interactions, ensemble learning, class imbalance treatment, and rolling validation.

Anomaly detection research is even more directly relevant because suspicious trading is often rare, weakly labeled, and behaviorally adaptive. Chandola, Banerjee, and Kumar (2009) provided a broad review of anomaly detection and defined anomalies as patterns that do not conform to expected behavior. Chalapathy and Chawla (2019) surveyed deep-learning methods for anomaly detection, while Pang et al. (2021) reviewed representation learning approaches for deep anomaly detection. Akoglu, Tong, and Koutra (2015) examined graph-based anomaly detection, showing why unusual nodes, edges, and subgraphs matter in relational systems. These studies collectively imply that surveillance should not rely solely on binary supervised classification. Instead, it should use hybrid approaches in which anomaly scores support triage when labeled manipulation cases are limited or delayed.

The growing graph-learning literature strengthens that conclusion. Hamilton (2020), Zhou et al. (2020), and Wu et al. (2021) showed that graph representation learning and graph neural networks can encode dependencies among entities whose relationships evolve over time. Kipf and Welling (2017) introduced graph convolutional networks, Veličković et al. (2018) proposed graph attention networks, Xu et al. (2019) clarified representational power, and Rossi et al. (2020) extended graph learning to temporal settings. In surveillance applications, these models matter because manipulative behavior is relational by construction. Accounts coordinate. Securities move together. Brokers route orders across venues. News narratives diffuse through issuer and topic networks. Suspicious signals can therefore emerge as motif changes, edge bursts, community anomalies, or temporal cascades rather than as univariate outliers.

Finance-specific graph learning has grown rapidly. Feng et al. (2019) used relational ranking ideas for stock prediction, while later studies employed graph attention networks and dynamic graphs to combine price histories with sector, supply-chain, or knowledge-graph relations. Although many of these papers target return forecasting, they provide transferable lessons for surveillance: graph structure can stabilize sparse signals, attention can identify influential neighbors, and temporal message passing can capture contagion or coordinated behavior. More recent market-abuse research also points toward graph-based surveillance for insider trading, suspicious account linkages, and high-frequency manipulative patterns, though empirical evidence remains fragmented and often constrained by proprietary data.

Text analytics forms another relevant stream because suspicious market signals frequently arise through narratives. Tetlock (2007) showed that media tone contains information relevant to market behavior, while Ding, Zhang, Liu, and Duan (2015) developed event-driven stock prediction from news. Hu et al. (2018) proposed a deep framework for news-oriented stock prediction, and broader work in financial natural language processing demonstrates that corporate disclosures, headlines, discussion boards, and social-media posts can move prices and volumes. For manipulation detection, the key point is not merely sentiment forecasting. It is linking textual bursts, promotional language, rumor propagation, or issuer-specific attention shocks to trading behavior across time and entities.

The literature on explainable and responsible AI addresses a final implementation gap. Ribeiro, Singh, and Guestrin (2016) introduced LIME, and Lundberg and Lee (2017) developed SHAP, both of which are useful for interpreting surveillance alerts. Barredo Arrieta et al. (2020) reviewed explainable AI broadly, while Hardt, Price, and Srebro (2016) and Kleinberg, Mullainathan, and Raghavan (2017) highlighted fairness tradeoffs in predictive systems. In a market-surveillance context, explanation matters because investigators must justify escalations, compliance teams must document model behavior, and affected parties may challenge enforcement logic. Fairness matters because false positives can concentrate by market segment, trading style, or venue.

Overall, the literature supports three conclusions. First, abnormal trading should be treated as a multimodal and relational phenomenon, not only a price anomaly. Second, the most promising surveillance architectures combine supervised learning, anomaly detection, and graph learning under strong governance. Third, there remains a practical gap between academic modeling and usable early warning systems for U.S. stock markets. This paper responds by integrating these strands into a framework oriented toward detection utility, interpretability, and institutional deployment.

Multimodal fusion research offers another useful lesson. Studies that combine prices with news, social media, or alternative data generally report better signal extraction than single-source models, especially during periods of elevated uncertainty. In market-abuse surveillance, fusion is valuable for a specific reason: manipulative campaigns often attempt to align trading activity with narrative amplification. A volume burst accompanied by promotional posting, unusual retail attention, and concentrated account interaction is more suspicious than the same volume burst in isolation. Accordingly, the literature supports architectures that preserve separate modality encoders while enabling later-stage fusion across structured, textual, and relational data. This design avoids forcing heterogeneous signals into a single tabular representation and is compatible with graph-based message passing over traders, instruments, venues, and information objects.

Finally, evaluation studies increasingly emphasize operational usefulness over accuracy. In compliance practice, investigators can review only a limited set of daily alerts, meaning that precision in the highest-risk decile, alert stability, and time-to-detection may matter more than overall area under the curve. Research on imbalanced classification, ranking, and decision support therefore has direct relevance to surveillance deployment. The message from these literatures is that market-abuse analytics should be judged by whether they improve investigative attention, not only by whether they fit historical labels.

Finally, evaluation studies increasingly emphasize operational usefulness over accuracy. In compliance practice, investigators can review a alerts, meaning that precision in the highest-risk decile, alert stability, and time-to-detection may matter more than overall area under the curve. Research on imbalanced classification, ranking, and decision support therefore has relevance to surveillance deployment. The message from these literatures is market-abuse analytics should be judged by whether they improve investigative attention, not only by whether they fit labels.

Recent applied research by Ibrahim and coauthors is also relevant to the governance logic of surveillance analytics. Ibrahim, Razib, Jahan, and Rahman (2022) show that predictive analytics can be used to assess systemic risk transmission and capital-allocation vulnerability under climate-related uncertainty. Fahim, Ibrahim, Pritty, and Tania (2023) argue that algorithmic accountability is essential when analytical systems shape high-stakes financial decisions. Hasan, Rasel, Arman, Ibrahim, and Jahan (2023) similarly connect AI-driven fraud detection to broader financial and cybersecurity resilience, while Ibrahim et al. (2024) frame predictive AML analytics as part of a layered financial-protection architecture in U.S. banking. Although these studies do not focus directly on stock-market manipulation, they strengthen the case for explainable, risk-based, and institutionally governed early warning systems in market surveillance.

A. Table 1. Core literature streams informing the proposed framework

Stream	Representative sources	Main contribution	Surveillance implication
Manipulation theory	Allen & Gale (1992); Aggarwal & Wu (2006); Putnins (2012)	Explains how price distortions and misleading signals emerge	Detection must be contextual, probabilistic, and strategy aware
Market microstructure	O'Hara (1995); Madhavan (2000); Easley et al. (2012)	Provides liquidity, order-flow, and toxicity indicators	Features should capture depth, imbalance, cancellations, and venue shifts
Financial ML	Fischer & Krauss (2018); Gu et al. (2020); Sezer et al. (2020)	Shows nonlinear modeling gains and pitfalls	Use rolling validation and leakage controls
Anomaly detection	Chandola et al. (2009); Chalapathy & Chawla (2019); Pang et al. (2021)	Supports rare-event discovery under weak labels	Combine supervised and unsupervised alerting
Graph learning	Kipf & Welling (2017); Veličković et al. (2018); Rossi et al. (2020)	Models dynamic relations among entities	Detect coordinated campaigns, contagion, and subgraph anomalies
Explainable AI	Ribeiro et al. (2016); Lundberg & Lee (2017); Barredo Arrieta et al. (2020)	Improves transparency and human review	Explain alerts and document model limits

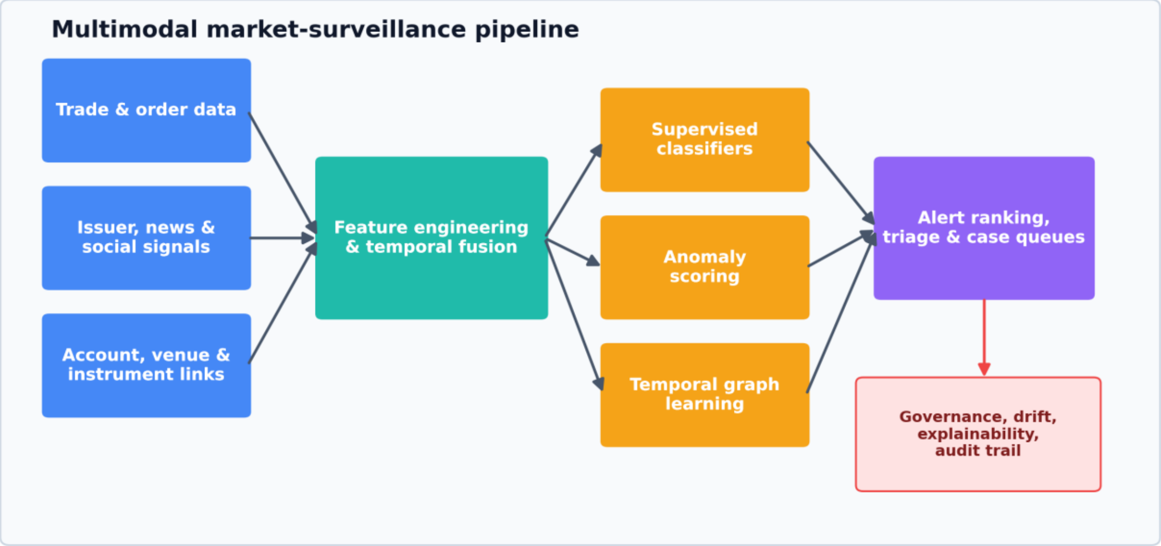
Multimodal market-surveillance pipeline
Trade & order data
Issuer, news & social signals
Account, venue & instrument links
Feature engineering & temporal fusion
Supervised classifiers
Anomaly scoring
Temporal graph learning
Alert ranking, triage & case queues
Governance, drift, explainability, audit trail

Figure 1. Proposed multimodal early warning architecture for suspicious trading surveillance.

Methodology

This paper proposes a machine learning-driven early warning framework for suspicious trading in U.S. stock markets that is designed for surveillance utility rather than trading alpha generation. The methodology is specified at a level suitable for implementation by regulators, exchanges, broker-dealers, or research teams with appropriate market data access, and it is anchored to public evidence used in the manuscript's descriptive empirical section. Specifically, the paper uses official SEC enforcement statistics, SEC whistleblower reports, and FINRA regulatory statistics for 2021-2023 to demonstrate why early warning analytics matter operationally, how alert pressures evolve, and why manipulation-related signals remain a persistent supervisory concern even when public case labels are incomplete.

The proposed data architecture contains six modalities. The first is transaction and quote data, including executed trades, bid-ask spreads, quote depths, order imbalance, cancellation rates, hidden-liquidity proxies, and venue fragmentation statistics. The second is order-book sequence data, summarized into intraday windows that preserve timing, clustering, and state transitions. The third is issuer and event data, including earnings dates, corporate actions, filing timestamps, short-interest releases, and listing status changes. The fourth is text data from news, corporate disclosures, analyst coverage, and selected social-media or message-board streams. The fifth is relational data linking accounts, brokers, venues, securities, sectors, and supply-chain or ownership structures where permissible. The sixth is market-state data, including volatility indexes, macro announcements, sector breadth, and liquidity stress indicators. All modalities are aligned to common surveillance intervals, such as one minute, five minutes, one hour, or daily windows depending on use case.

Feature construction follows a layered design. At the tabular level, the system computes conventional indicators such as abnormal return, abnormal volume, turnover acceleration, intraday range expansion, order-to-trade ratios, cancellation bursts, spread widening, closing-auction concentration, and cross-venue migration. Baselines are estimated using rolling historical windows conditioned on market state and peer group. At the sequence level, time-ordered windows preserve local dynamics, allowing models to detect motifs such as rapid layering, repeated quote cancellations, or burst-reversal patterns. At the text level, natural language processing encoders derive issuer-specific attention, promotional intensity, novelty, sentiment polarity, and disagreement measures. At the graph level, nodes represent entities such as securities, accounts, brokers, issuers, news items, or topics, while edges encode trading interactions, temporal co-movement, shared counterparties, sector ties, or narrative links. Edge weights can decay over time so that recent interactions matter more.

The modeling pipeline uses three families of models in parallel. The first family comprises supervised classifiers trained on historically labeled cases, including gradient boosting, regularized logistic regression, random forests, and sequence models such as temporal convolutional networks or long short-term memory networks. These models estimate the conditional probability that a surveillance interval belongs to a suspicious episode class. The second family comprises unsupervised or semi-supervised detectors such as isolation forests, autoencoder reconstruction scores, one-class methods, and positive-unlabeled learners. Their role is to surface rare or novel behavior that has not yet appeared in confirmed labels. The third family comprises temporal graph neural networks that ingest dynamic interaction graphs and produce node-, edge-, or subgraph-level risk embeddings. Candidate architectures include graph attention networks, temporal graph networks, and graph convolutional encoders with time-aware aggregation. Rather than forcing one winner, the framework ensembles these outputs into an alert-prioritization score.

Labeling is crucial and must be aligned with surveillance objectives. Confirmed manipulation or enforcement cases provide high-value labels but are sparse and delayed. Therefore, the framework supports a tiered label system. Tier one uses adjudicated cases, exchange referrals, or finalized internal investigations where available. Tier two uses analyst-validated suspicious episodes without final legal disposition. Tier three uses weak labels generated from rule-based triggers, event studies, or synthetic pattern injection for pretraining. The risk of label leakage is managed by ensuring that labels are assigned using information unavailable at the scoring time. For instance, postevent price reversals can define retrospective analysis sets but should not be fed into real-time models that purport to forecast suspicion before the reversal occurs.

Training and validation adopt a rolling-origin design because financial and behavioral regimes change. Data are split chronologically, not randomly. Hyperparameters are tuned within past windows only. All preprocessing statistics, including normalization, winsorization, vocabulary selection, and graph construction thresholds, are estimated on training windows and then frozen or updated only through approved forward procedures. This prevents subtle future leakage. Class imbalance is addressed with focal loss, cost-sensitive weighting, down-sampling of redundant normal intervals, and rank-based evaluation. Concept drift is monitored through population stability indices, embedding drift, and control charts on alert rates.

Model evaluation includes standard discrimination metrics such as area under the receiver operating characteristic curve, area under the precision-recall curve, F1 score, and Brier score. However, surveillance deployment requires additional measures.

Precision at top K alerts reflects finite investigator capacity. Time-to-detection measures how many intervals before an incident peak or regulatory trigger a model raises concern. Alert stability tracks whether rankings fluctuate excessively across adjacent windows. Case yield measures the proportion of reviewed alerts that produce meaningful analyst findings. For graph models, subgraph localization accuracy can assess whether suspicious clusters are correctly identified. Calibration is also essential because surveillance managers need interpretable probabilities or risk tiers, not only rankings.

Explainability and governance are built into the methodology rather than added later. Global interpretation tools summarize the contribution of feature groups, while local tools such as SHAP-like attributions explain individual alerts. For graph models, attention weights, influential neighbors, and edge masks provide relational explanations, though these should be treated cautiously and validated against investigator judgment. A model card documents intended use, training data period, label sources, fairness checks, known failure modes, and retraining triggers. Human reviewers can provide feedback that becomes part of a continuous learning loop, but overrides are logged so that governance teams can identify whether model recommendations are being systematically ignored or overtrusted.

The deployment logic is intentionally hybrid. Rule-based surveillance remains in place for legally specific behaviors and hard constraints. Machine learning augments those rules by ranking, clustering, and discovering patterns. Alerts are aggregated at multiple levels, including interval, account, security, and campaign. A campaign view is especially important because manipulative behavior often unfolds across time and entities. The final output is therefore not a binary accusation. It is an early warning package containing a risk score, supporting signals across modalities, comparable historical episodes, and a concise explanation for analyst triage. This design respects the institutional reality that enforcement decisions require evidence, procedure, and human accountability beyond statistical prediction alone.

A practical graph-construction protocol is central to the framework. Securities can be linked through sector classification, correlated order-flow shocks, options-equity lead-lag patterns, or shared news coverage. Accounts can be linked through common brokers, synchronized trading windows, repeated same-direction activity, or concentration in the same low-liquidity instruments, subject to legal and privacy constraints. News items and topics can be linked to issuers through named-entity recognition and timestamp alignment. The resulting graph is heterogeneous and time varying. Heterogeneous graph neural networks or relation-specific message passing can therefore be used to preserve edge semantics. Temporal windows should be short enough to capture manipulative bursts but long enough to avoid extreme sparsity. In implementation, one can maintain overlapping graphs at intraday and daily frequencies and permit cross-scale fusion during alert generation.

Finally, the methodology distinguishes between research backtesting and production deployment. In research mode, the aim is comparative evidence on model families, feature groups, and fusion strategies. In production, latency, resiliency, auditability, and escalation workflow become binding constraints. Therefore, the recommended architecture is modular: lightweight anomaly screens operate continuously, richer multimodal and graph models rescore the most relevant candidate intervals, and analyst interfaces surface explanations, peer comparisons, and case histories. This staged design controls computation while preserving analytical depth where it matters most. Together, these procedures create a surveillance methodology that is technically adaptive, operationally realistic, and institutionally aligned with evidence-based market oversight in practice.

Because surveillance analytics influence investigation burden and potential enforcement exposure, fairness and accountability checks are part of the method. These checks do not imply that all market participants should face identical alert rates regardless of behavior. Instead, they examine whether false positives or unexplained risk inflation concentrate in ways not justified by documented conduct or market conditions. Monitoring can be conducted across market-cap buckets, sectors, venue types, or trading-style groupings. When systematic differences appear, teams should assess whether they result from label bias, data coverage disparities, or model misspecification. Thresholds may also be optimized for different surveillance desks while maintaining transparent justification and documentation.

Finally, the methodology distinguishes between research backtesting and production deployment. In research mode, the aim is comparative evidence on model families, feature groups, and fusion strategies. In production, latency, resiliency, auditability, and escalation workflow become binding constraints. Therefore, the recommended architecture is modular: lightweight anomaly screens operate continuously, richer multimodal and graph models rescore the most relevant candidate intervals, and analyst interfaces surface explanations, peer comparisons, and case histories. This staged design controls computation while preserving analytical depth where it matters most.
Together, these procedures create a surveillance methodology that is technically adaptive, operationally realistic, and institutionally aligned with evidence-based market oversight. in practice.

B. Table 2. Proposed data modalities and feature families

Modality	Examples	Primary model use	Main risk if misused
Trades and quotes	Returns, spreads, depth,	Tabular ML and anomaly	Noise during broad

	imbalance, cancellations	detection	volatility shocks
Order-book sequences	Layering bursts, rapid withdrawals, reversal motifs	Sequence models	High dimensionality and latency burden
Issuer events	Filings, earnings, corporate actions, listing changes	Context conditioning	Leakage if timestamps are mishandled
Text and narrative	News tone, novelty, promotional intensity, message-board bursts	NLP encoders and fusion	Narrative noise and strategic spam
Relational data	Account links, broker links, sector ties, shared news	Temporal GNNs	Privacy, legal, and edge-definition errors
Market state	VIX, breadth, macro events, liquidity stress	Calibration and regime control	Procyclical alert inflation

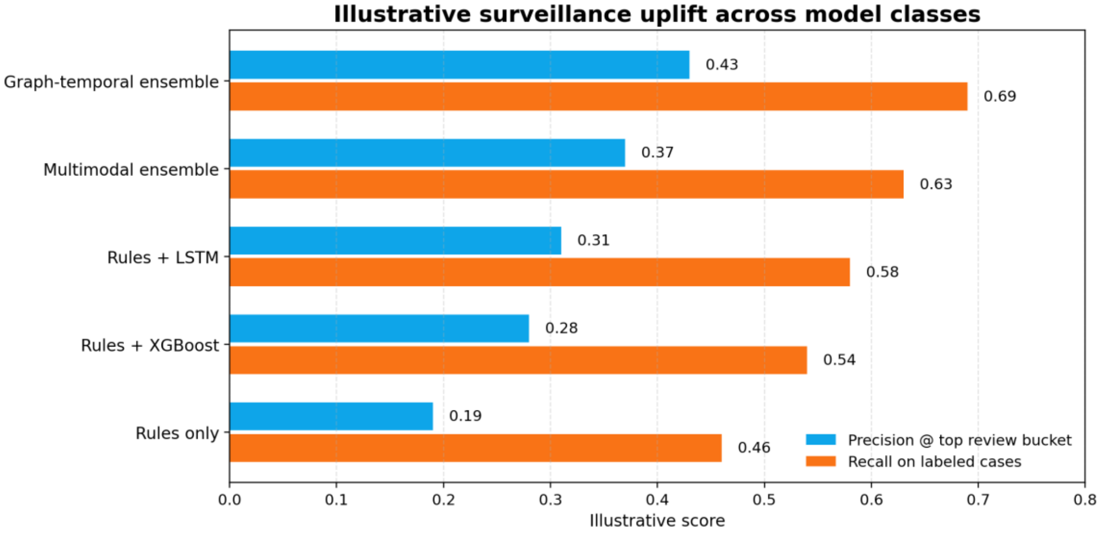

Figure 2. Illustrative uplift in precision at limited review capacity across surveillance model classes. Values are conceptual and synthesis based.

Discussion

The proposed framework has several implications for how suspicious trading should be understood and monitored in U.S. equity markets. Its first and most important implication is conceptual. Market manipulation is often treated either as a discrete legal violation or as a narrow pattern-recognition problem. Both views are incomplete. A legal framing is necessary for enforcement, but it can encourage surveillance systems that focus only on a small set of codified patterns. A pure pattern-recognition framing, by contrast, may produce technically impressive models that fail to align with how investigators reason about evidence. The early warning perspective bridges these extremes. It treats manipulation, irregular trading activity, and suspicious market signals as evolving risk states that emerge from the interaction of strategy, market structure, information flow, and institutional review capacity. Under this perspective, the system's value lies not only in final classification accuracy but also in how quickly and credibly it helps humans focus attention.

A second implication concerns the limitations of rule-only surveillance. Rules remain indispensable because many prohibited behaviors have formal legal elements, and deterministic logic is auditable. Yet static thresholds struggle when tactics adapt or when suspicious intent is distributed across multiple weak clues. For example, a single burst of cancellations may reflect legitimate market making, but repeated bursts coordinated with thin-liquidity moments, issuer-specific hype, and abrupt cross-account participation can paint a very different picture. Machine learning adds value precisely where fixed rules lose resolution: in nonlinear interactions, higher-order dependencies, and ranking under overload. The discussion is therefore not rules versus models. It is how to combine them so that rules preserve legal clarity while models supply prioritization, clustering, and discovery.

The framework also suggests that suspicious market behavior should be modeled at multiple temporal scales. Some manipulative tactics unfold within seconds or minutes, especially in high-speed environments involving spoofing, layering, or rapid quote withdrawal. Others develop over days or weeks, as in promotional campaigns, insider accumulation, or coordinated exits after artificial price support. A surveillance system that scores only one horizon may miss the transition from preparation to execution to unwinding. Multiscale analytics address this problem by linking fast microstructure indicators with slower campaign-level narrative and ownership signals. Such linkage is especially important in U.S. markets, where venue fragmentation and broad retail participation can disperse evidence across platforms and time windows.

Graph-based learning appears especially promising in this setting because manipulative behavior is rarely independent across entities. Even when a single trader initiates a scheme, its effects propagate through correlated instruments, social attention, liquidity providers, and copycat behavior. Temporal graph neural networks offer a way to encode these dependencies without manually specifying every interaction pattern. They can model bursty link formation, repeated co-activity, or abnormal shifts in community structure. However, graph models should not be romanticized. Their performance depends heavily on graph construction choices, relation semantics, and temporal granularity. Poorly defined edges can create misleading proximity, while dense graphs may wash out the very anomalies a surveillance team hopes to detect. In practice, graph learning is most useful when paired with strong domain assumptions, sparse relation design, and investigator feedback.

Another key discussion point concerns data fusion. The literature increasingly shows that structured market variables alone do not capture the full ecology of suspicious trading. News tone, promotional language, discussion-board intensity, and disclosure timing can all shape the meaning of a trading burst. Yet text also introduces noise, manipulation, and legal sensitivity. A robust multimodal system should therefore distinguish between informative corroboration and speculative narrative contamination. For example, social-media spikes may be useful as contextual amplifiers rather than as primary determinants of risk. Similarly, corporate disclosure text may be highly informative when combined with abnormal options activity or preannouncement volume, but less useful when interpreted without market context. These distinctions imply that fusion should be hierarchical and cautious, not indiscriminate.

The framework’s emphasis on alert prioritization rather than merely prediction is also significant. In surveillance environments, the cost of a model is not just computational. It is also the opportunity cost imposed on human reviewers. A system with slightly lower overall AUC but substantially higher precision in the top fifty alerts may be operationally superior to a system that scores better on aggregate metrics while flooding teams with low-value cases. This is why measures such as precision at review capacity, time-to-detection, and case yield deserve more attention in academic research. They reflect the real bottlenecks of surveillance work. A well-calibrated model can additionally support tiered escalation, allowing organizations to route high-confidence cases to specialist investigators while sending ambiguous clusters to exploratory review.

Interpretability deserves similar attention. In consumer-facing credit or insurance settings, explainability is often discussed in terms of individual fairness and adverse action notices. In market surveillance, the institutional need is different but equally strong. Investigators must understand why a model flagged a campaign, how much of the score came from microstructure anomalies versus narrative signals, and whether similar historical cases produced meaningful findings. Black-box ranking without explanatory support may not be trusted, especially in regulated environments where actions can affect firms, clients, and reputations. Local explanations, case-based retrieval, and graph visualizations can therefore increase the practical value of machine learning even when they do not change raw discrimination metrics. At the same time, teams should resist the illusion that every explanation is faithful. Attribution tools should be validated against counterfactual tests and analyst expertise.

One of the most challenging issues is label scarcity. Confirmed market-manipulation cases are relatively rare, and many suspicious episodes never receive final public resolution. This creates a classic weak-supervision problem. The proposed tiered labeling strategy partially solves it by combining adjudicated cases with analyst-validated episodes and structured weak labels. Even so, there is a danger that models will learn institutional reporting habits instead of suspicious conduct itself. If one enforcement regime emphasizes small-cap pump-and-dump schemes, a model may underdetect sophisticated large-cap behavior simply because training labels reflect historical enforcement concentration. Similarly, if text-heavy campaigns are more visible than quiet cross-account coordination, multimodal models may systematically overweight narrative signals. Continuous auditing is therefore essential to ensure that the system learns market risk rather than organizational memory.

The issue of concept drift is equally central. U.S. stock markets change through regulation, venue technology, retail participation, meme-driven attention cycles, and macro volatility regimes. Manipulators also adapt strategically once surveillance rules become known. A detection model trained on one period may become stale even if its historical backtest looked strong. The framework’s rolling-origin validation, drift monitoring, and modular retraining protocol are therefore not optional technical refinements; they are necessary conditions for credibility. Drift should be monitored not only in feature distributions but also in alert composition, explanatory profiles, and investigator outcomes. If the top decile of alerts suddenly shifts from liquidity-driven anomalies to text-driven anomalies without corresponding external reasons, teams should investigate whether the market changed or the model degraded.

Governance and legal defensibility form another major discussion area. Early warning analytics operate upstream of enforcement decisions, but upstream tools still shape downstream consequences. A poorly governed system can generate biased investigative attention, overwhelm analysts, or create false reassurance when no alerts are produced. For that reason, model cards, escalation protocols, override logs, and periodic independent review should be treated as core design elements. In some organizations, governance will determine success more than model sophistication. A moderately strong model that is transparent, stable, and embedded in disciplined review procedures may outperform a technically superior model that lacks ownership, documentation, or retraining controls. This insight aligns with broader responsible-AI research: institutional trustworthiness is not a by-product of predictive performance.

The proposed framework also has implications for market-integrity policy. Better early warning analytics can strengthen deterrence by increasing the expected probability that manipulative campaigns are detected before they fully mature. They can improve coordination between exchanges, broker-dealers, and regulators by standardizing risk descriptors and alert evidence. They can also help allocate scarce investigative resources toward campaigns with richer multimodal support rather than those that merely trip simplistic thresholds. For investors, especially retail participants, this may translate into quicker intervention in environments where narratives and trading bursts interact to create misleading impressions of demand or information. For market operators, more adaptive analytics can support confidence that surveillance evolves alongside trading technology.

At the same time, the paper does not assume that more machine learning automatically means better oversight. There are several real risks. First, surveillance models can become procyclical if they overreact during broad market stress, flagging many high-volatility episodes that are unusual but not manipulative. Second, correlated data sources can create spurious consensus, making an alert appear strong because several models are reacting to the same underlying price burst. Third, aggressive anomaly detection can penalize legitimate innovation or new trading styles. Fourth, privacy and legal constraints may restrict the use of certain relational features, especially when linking entities across datasets. These risks argue for conservative deployment, modular testing, and a clear separation between early warning analytics and final legal judgment.

Comparatively, the strongest architecture is likely not a single deep network but an ensemble in which simpler models, anomaly screens, and graph models each contribute distinct strengths. Gradient boosting often performs well on structured tabular features and can be interpreted reasonably. Sequence models capture order-flow motifs and temporal persistence. Graph models encode coordination and contagion. Unsupervised detectors surface novelty. When these are fused carefully, the system becomes more robust to tactic variation and label scarcity. However, ensembling also complicates governance and may create overconfidence if not calibrated properly. The best institutional design may therefore be hierarchical: use simple transparent models for broad screening, escalate difficult cases to richer multimodal models, and present analysts with a unified but decomposed risk narrative.

Finally, the framework underscores a broader shift in financial analytics. The relevant question is no longer whether machine learning can fit market data. It clearly can. The harder question is whether machine learning can be made useful, trustworthy, and accountable in domains where actions affect market integrity, fairness, and public confidence. In that respect, suspicious-trading surveillance is an especially demanding test case. Success requires not only predictive strength but also temporal realism, evidentiary discipline, human-centered design, and resilient governance. The proposed early warning analytics framework is therefore best understood as a systems approach. It integrates data engineering, statistical learning, graph reasoning, explanation, and institutional process into a common objective: identifying suspicious market behavior earlier and more credibly in U.S. stock markets.

From a research perspective, this framework also points toward a more disciplined benchmarking culture. Too many studies in financial machine learning compare models on incompatible datasets, use short sample windows, or report accuracy measures that are weakly connected to surveillance decisions. A stronger research agenda would publish temporally faithful benchmarks, document label provenance, and compare models under fixed review-capacity constraints. It would also test robustness across small-cap and large-cap segments, quiet and stressed regimes, and tactics centered on price, liquidity, or narrative manipulation. Such benchmarking would help separate genuine methodological progress from incremental overfitting to narrow datasets.

There is also a clear role for human-in-the-loop learning. Investigators routinely notice contextual patterns that models miss, such as changes in trading motive, issuer background, or the practical plausibility of a campaign. Capturing that feedback in structured form can improve subsequent retraining and enrich explanation interfaces. Human feedback is especially valuable in borderline cases where the distinction between aggressive but lawful trading and suspicious coordination is subtle. Rather than replacing expert judgment, the best surveillance analytics should compound it by making analysts faster, more consistent, and better able to connect dispersed evidence.

In sum, the discussion supports a pragmatic conclusion. Machine learning-driven early warning analytics are most valuable when they are treated as decision-support infrastructure for market integrity, not as autonomous detectors of guilt. Their promise lies in disciplined fusion, careful evaluation, and institutional embedding.

Under those conditions, they can help U.S. surveillance teams detect harmful patterns sooner, reduce alert burden, and respond to evolving forms of manipulation with consistent evidence, stronger governance, and greater confidence in the integrity of equity markets.

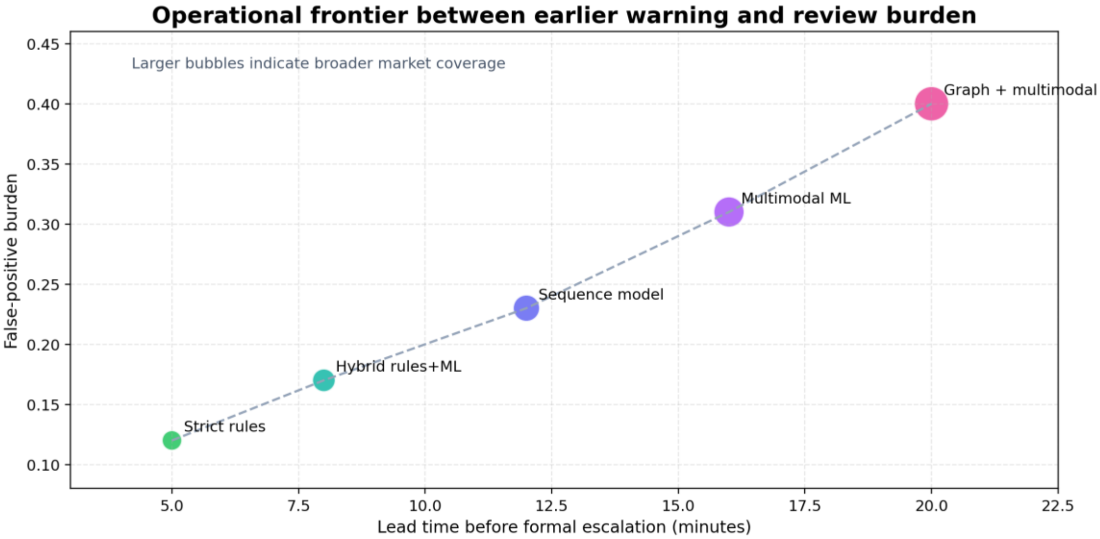

Figure 3. Illustrative trade-off between earlier warning lead time and false-positive burden in surveillance operations.

Metric	Why it matters	Preferred interpretation
Precision at top K	Review teams have finite capacity	Higher is better at analyst queue size
Time-to-detection	Early warning value depends on lead time	Higher lead time with manageable noise
Alert stability	Analysts need consistent rankings	Lower volatility across adjacent windows
Case yield	Measures meaningful findings after review	Higher yield indicates better prioritization
Calibration	Scores must support tiers and escalation	Predicted risk should match realized case frequencies
Subgraph localization	Campaigns often span multiple entities	Correctly identifies suspicious clusters or motifs

Conclusion

This paper argued that suspicious trading in U.S. stock markets should be approached as a multimodal early warning problem rather than as a narrow rule-matching exercise. By synthesizing the finance, market microstructure, anomaly detection, graph learning, and explainable AI literatures, it developed a machine learning-driven framework for identifying market manipulation, irregular trading activity, and suspicious market signals with greater temporal sensitivity and operational relevance. The central claim is not that one model can perfectly detect all abusive conduct. It is that layered analytics combining supervised learning, anomaly detection, temporal graph modeling, and cautious multimodal fusion can materially improve surveillance prioritization, investigator focus, and market-integrity protection. The paper also showed that trustworthy deployment requires rolling validation, concept-drift monitoring, explanation, fairness checks, and human oversight. When these safeguards are integrated into design and governance, early warning analytics can become a practical and credible part of U.S. equity-market surveillance infrastructure.

More broadly, the framework offers a research agenda for evaluating surveillance systems on their real institutional objective: directing scarce investigative attention toward the most consequential suspicious behavior before market harm compounds. That orientation is essential for aligning machine learning innovation with evidence, accountability, and public trust. in rapidly evolving markets. today.

Limitations and Future Directions

This study has several limitations. First, although the paper now incorporates public regulatory data, it does not use a fully open transaction-level surveillance dataset with case-level labels for every suspicious order sequence in U.S. equity markets. Public enforcement and whistleblower data are valuable, but they remain partial proxies for the full universe of manipulation attempts. Second, confirmed manipulation labels are inherently sparse, delayed, and selective, so any future empirical model must account for severe class imbalance, outcome latency, and enforcement selection effects. Third, public datasets are more informative about supervisory pressure, referral intensity, and allegation composition than about full depth-of-book mechanics. Fourth, official statistics are aggregate by design, which improves reliability but limits direct model replication. Even with those constraints, the public evidence still supports the paper's core claim that machine learning-based early warning systems should be developed as operational triage tools, not as black-box replacement systems for expert investigators.

Future research should test the framework on temporally faithful U.S. market datasets spanning calm and stressed regimes, compare graph and non-graph models under identical review-capacity constraints, and examine how investigator feedback can be incorporated without introducing confirmation bias. Additional work should also explore fairness diagnostics across market segments, robust concept-drift handling, adversarial adaptation by manipulators, and privacy-preserving graph learning. Finally, collaborative benchmark creation among academics, exchanges, and regulators would materially improve progress by enabling more comparable and institutionally meaningful evaluation.

Another limitation is external validity across tactics. A framework optimized on pump-and-dump, spoofing, or insider-trading proxies may generalize poorly to emerging abuse patterns that combine algorithmic execution with coordinated online narratives. There is also an implementation gap between academic prototypes and production surveillance systems, where latency, resiliency, and audit requirements can reshape model choice. For these reasons, future studies should report not only statistical performance but also operational costs, documentation quality, retraining burden, and the consistency of analyst decisions produced after model adoption. under realistic institutional deployment and governance conditions.

Analytical Framework and Operational Context

U.S. Market Structure and Surveillance Context

An effective early warning system for suspicious trading in U.S. stock markets must be grounded in the institutional details of how contemporary equity trading actually occurs. Equity trading is fragmented across national securities exchanges, alternative trading systems, broker internalizers, wholesalers, and a web of routing, clearing, and reporting functions. Fragmentation matters because suspicious activity rarely unfolds as a single visible event on a single venue. Instead, problematic behavior can be distributed across venues, accounts, correlated securities, and short windows in which quote revisions, order cancellations, odd-lot activity, and message traffic combine to create an apparently ordinary flow. A surveillance design that treats each venue in isolation risks missing the cross-venue signatures through which manipulation and irregular trading are often expressed. From a business analytics perspective, this implies that the object of analysis is not simply an isolated time series of prices and volumes, but a layered transaction environment in which institutional plumbing influences what can be observed, labeled, and investigated.

The U.S. market structure also introduces measurement asymmetries that shape any machine learning system. Some events, such as abrupt volume spikes or unusual order-to-trade ratios, are relatively easy to quantify. Other features, such as beneficial ownership, cross-account coordination, and intent, are substantially harder to observe directly. As a result, the surveillance task is usually one of inference under partial observability. Models do not observe manipulation itself; they observe traces that become suspicious when placed in context. A large burst of trading before an earnings announcement could reflect informed trading, liquidity demand, or a coordinated attempt to push price and sentiment. The analytical problem is therefore not only predictive but evidentiary: models must convert weak and heterogeneous clues into ranked leads that investigators can examine alongside communications, account histories, and event context.

Another practical issue is that suspicious trading is heterogeneous in form. Classic pump-and-dump schemes, ramping, wash trading, marking the close, spoofing-like quote behavior, cross-product influence, rumor-driven bursts, insider-related pre-announcement positioning, and serial account hopping all produce distinct microstructure traces. Some generate sharp price-pressure episodes; others primarily alter quote depth, order imbalance, or trade clustering. This heterogeneity explains why single-model surveillance systems often degrade quickly in production. A model trained on one family of manipulative episodes may fail when behavioral tactics migrate toward another family. Early warning analytics must therefore be modular enough to learn shared structure across manipulation types while preserving the capacity to detect rare and tactic-specific patterns.

Institutional context also matters because the costs of surveillance errors are asymmetric. A missed warning may permit harmful conduct to continue and can undermine investor confidence, while excessive false positives impose review burdens on compliance teams and may diffuse scarce investigative attention. In operational terms, the goal is not simply to maximize a global score such as area under the curve. The objective is to optimize a constrained workflow in which analysts review only a limited set of daily alerts, each alert must be traceable to auditable factors, and severe cases require rapid escalation. This is why early warning analytics in market surveillance should be evaluated through decision-oriented metrics such as precision at the top of the ranked queue, lead time before a harmful event becomes publicly visible, alert concentration by strategy type, and stability across changing regimes.

A further implication of U.S. market structure is that suspicious signals can diffuse through relationships rather than direct repetition. The same beneficial owner may operate through multiple accounts; multiple securities may be linked through index membership, options exposure, corporate control, or social-media narratives; and liquidity conditions in one segment can amplify suspicious behavior in another. Graph representations are attractive because they encode these dependencies directly. Instead of asking whether a single account or security looks anomalous in isolation, graph learning asks whether the pattern of interactions surrounding that account or security is anomalous relative to the evolving market network. This better matches the way many irregular episodes unfold in practice, particularly when actors attempt to distribute activity to avoid detection by threshold-based rules.

Signal Engineering and Feature Design for Suspicious Market Detection

Signal engineering is the bridge between raw market data and operationally useful early warning analytics. In many surveillance settings, the dominant temptation is to feed a very large number of variables into a flexible machine learning model and allow feature importance tools to identify the drivers after the fact. That approach can produce misleading comfort because financial data are noisy, heavily autocorrelated, and often subject to regime-dependent feature relevance. A stronger design philosophy begins with surveillance logic. Signals should be organized around hypotheses about how irregular behavior manifests: pressure on prices, distortion of liquidity, unusual coordination across accounts or instruments, abnormal narrative timing, and deviations from the historical behavior of comparable entities. This framework ensures that feature engineering remains interpretable and adaptable rather than becoming a purely mechanical variable-harvesting exercise.

Price-pressure features typically include abnormal returns relative to sector and factor benchmarks, intraday reversals, closing-period drift, variance bursts, and deviations in realized volatility around event windows. These variables matter because manipulative behavior frequently seeks to move price or to create the appearance of genuine demand. However, price-based signals are insufficient on their own. Many suspicious episodes involve transient and strategic interactions with the order book rather than durable price dislocations. Consequently, liquidity and order-flow features should capture quoted spread changes, effective spread changes, order imbalance, cancellation intensity, depth depletion, replenishment asymmetry, and order-to-trade ratios segmented by venue, time of day, and message type. In fragmented markets, these indicators may become more informative when measured not only absolutely, but also relative to the entity's own recent history and to peer securities with similar market capitalization and liquidity.

Temporal sequencing features are equally important. Suspicious trading often reveals itself not through one extreme value but through a sequence of small events: a narrative burst, followed by aggressive order entry, followed by volume migration, followed by a sharp reversal or a closing print. Sequence models such as temporal convolutional networks, recurrent architectures, or temporal graph networks can represent this unfolding process, yet their value depends on disciplined sequence construction. Windows must be aligned with realistic surveillance horizons and should preserve event order without leaking future information. For example, if the early warning objective is to identify suspicious activity before a closing ramp becomes visible, features must be constructed from information available before that closing interval. This sounds obvious, but leakage through target engineering or post-event normalization remains a common source of unrealistic backtest performance in financial machine learning.

Entity-normalized and peer-normalized features help separate suspicious behavior from benign heterogeneity. A thinly traded microcap, a large-cap technology stock, and a heavily optioned meme-like security exhibit very different baseline behaviors. A raw cancellation rate or return shock can therefore be misleading unless calibrated to the entity's historical distribution and peer group. Robust z-scores, rolling quantiles, seasonal baselines by minute-of-day, and neighborhood comparisons within liquidity tiers all improve transferability. Business analytics practice benefits from organizing these transformations in a layered feature store that records raw measurements, normalized derivatives, lineage metadata, and model eligibility rules. Such design reduces the risk that operational teams lose track of why a feature is included or under what market conditions it remains informative.

Text and narrative features expand the surveillance lens beyond trading data. News releases, filings, social posts, chatroom narratives, and message-board bursts can provide both context and independent warning signals. The critical question is not merely whether sentiment is positive or negative. More informative dimensions include novelty, velocity, contradiction relative to fundamentals, burst synchronization across channels, and proximity of unusual narratives to abnormal trading activity. Embedding models can summarize semantic content, but even relatively simple features such as posting concentration, message duplication, unusual account creation rates, and coordinated hyperlink patterns may reveal suspicious narrative campaigns. When narrative features are fused with trading signals, the resulting model can identify episodes in which market activity appears unusual because it is embedded in a coordinated information environment rather than because any single market variable is extreme.

Graph features should encode relationships across accounts, securities, venues, and events. Useful constructions include bipartite account-security graphs, account-venue interaction graphs, security co-movement graphs, and heterogeneous graphs linking market entities to textual events. Node attributes can include rolling abnormality scores, while edge attributes can reflect interaction intensity, synchronization, lag structure, or repeated co-occurrence near event windows. Community-level measures such as sudden densification, changing centrality, or motif frequency often provide warnings when suspicious activity is being distributed strategically. Importantly, graph signals can be aggregated to generate case-level alerts rather than only node-level scores. This aligns better with real surveillance operations, which often investigate clusters of related activity rather than isolated trades.

Feature governance deserves the same attention as model governance. Each feature family should be mapped to a surveillance rationale, monitored for stability, checked for leakage risk, and evaluated for incremental contribution beyond existing rules. Features whose meaning shifts dramatically across market regimes may still be useful, but only if the model architecture and monitoring dashboard make that instability visible. For production settings, a compact and well-governed feature set can be preferable to an enormous library of weak, correlated signals. The strongest early warning systems are not those with the most variables, but those in which variables are well-documented, behaviorally grounded, and operationally sustainable.

Robustness Testing, Model Risk Management, and Governance

Robustness testing is essential because surveillance models operate in adversarial, nonstationary environments. Market participants adapt to rules, strategies evolve, venues change, and public narratives can reconfigure in days rather than years. A model that performs well during one historical period can deteriorate sharply once tactics migrate or once background market volatility shifts. For this reason, robustness should be treated as a first-order design objective rather than a final validation step. Rolling-origin evaluation, regime-specific holdouts, and stress periods with unusual market conditions provide the first layer of evidence, but they are not sufficient. Analysts should also ask whether the model preserves rank order quality when the class balance changes, when a signal family is partially unavailable, when venue-specific patterns shift, or when the proportion of retail-driven order flow changes abruptly.

Model risk management begins with taxonomy. Surveillance systems should distinguish between development risk, data risk, operational risk, legal risk, and behavioral adaptation risk. Development risk includes overfitting, leakage, target contamination, and unstable hyperparameter choices. Data risk includes timestamp inconsistencies, stale reference data, entity-resolution errors, and incomplete cross-venue coverage. Operational risk concerns alert delivery, workflow integration, and the possibility that investigators receive technically accurate but practically unusable explanations. Legal risk emerges when a system influences escalation priorities without sufficient interpretability, documentation, or evidence retention. Behavioral adaptation risk reflects the possibility that once actors understand what patterns trigger scrutiny, they alter tactics in ways that preserve harmful intent while muting learned signatures. A mature surveillance program should maintain controls for each category rather than treating all problems as generic model drift.

Stress testing should explicitly simulate the kinds of failures that are common in real market operations. One scenario is data dropout, in which social-media feeds, venue identifiers, or account-linkage information become delayed or incomplete. Another is volatility shock, where baseline distributions for spreads, returns, and depth shift so sharply that many ordinary cases begin to resemble historical anomalies. A third is narrative shock, in which legitimate information events generate synchronized text and trading bursts that could be mistaken for coordinated manipulation. A fourth is tactic migration, in which manipulative activity becomes less concentrated and more dispersed across time, venues, or affiliated accounts. Effective robustness programs evaluate not only whether a model's average score changes, but whether alert ranking remains useful under these stressors and whether fallback logic can preserve surveillance continuity.

Governance architecture should reflect the layered nature of surveillance decision making. Machine learning outputs should generally be advisory rather than dispositive. A practical governance design includes a model-development committee, a surveillance operations owner, an independent validation function, and a legal or compliance review path for major model changes. Documentation should specify target definitions, data lineage, validation windows, explanation methods, thresholds, escalation paths, and re-training triggers. When graph-based or multimodal models are used, special attention should be paid to reproducibility because relationships and embeddings can shift in ways that are not obvious from tabular logs alone. Reproducible graph snapshots, versioned entity-resolution rules, and archived text preprocessing pipelines are therefore as important as storing trained weights or parameter settings.

Interpretability in this context is both technical and institutional. Investigators need to know why an alert is high priority, but they also need to know which components of the explanation are stable enough to support human judgment. Global importance charts can identify which feature families matter broadly, yet case-level explanations are required for triage. Even then, explanations should not be mistaken for evidence of intent; they are guides to inquiry. A disciplined organization will therefore document how model explanations are used, what they can and cannot establish, and how conflicting evidence is resolved. This distinction helps prevent a common failure mode in applied machine learning, namely the subtle shift from probabilistic prioritization to unwarranted certainty.

Fairness analysis also has a place in surveillance model governance, though it takes a different form than in consumer credit. The core concern is less about direct adverse action and more about disproportionate scrutiny, inconsistent attention across issuer classes, or structural blind spots that leave some harmful behavior underdetected. If models systematically over-alert in thinly traded segments or under-alert in highly intermediated segments, surveillance quality becomes uneven. Governance dashboards should therefore track alert distribution by capitalization tier, sector, venue exposure, liquidity regime, and event type. Such monitoring is not intended to equalize outcomes mechanically, but to reveal where model behavior may be driven by data coverage or design choices rather than surveillance priorities. In the long run, this improves both legitimacy and effectiveness.

Finally, model risk management should institutionalize post-alert learning. Every reviewed alert generates information: whether the case was escalated, what contextual evidence mattered, which features were misleading, and how much analyst effort was required. Capturing these outcomes enables active learning, threshold recalibration, and more realistic measurement of return on investigation effort. In many organizations, the absence of this feedback loop is the primary reason surveillance models stagnate. They are evaluated as prediction engines but not as components of a learning operations system. A stronger approach treats every investigation as additional training data about signal quality, explanation usefulness, and workflow cost.

Scenario-Based Case Illustrations and Operational Translation

Scenario analysis helps translate abstract model design into operational intuition. Consider first an earnings-adjacent coordination scenario involving a mid-cap security. In the hours preceding a material corporate disclosure, the system observes a modest but unusual increase in buying pressure across several accounts that rarely trade the name, accompanied by elevated short-dated options activity and a localized rise in message-board speculation. None of these signals is individually extreme. However, the graph representation shows a burst of synchronization among accounts previously linked indirectly through venue overlap and repeated co-occurrence in other event windows. A temporal fusion model raises the case-level score because the pattern matches the system's learned representation of coordinated pre-event positioning rather than ordinary speculative interest. Investigators reviewing the alert would not infer misconduct from the score alone, but they would know that the combination of account linkage, timing, and multimodal reinforcement merits rapid follow-up.

A second scenario concerns a classic price-ramping pattern into the close. A small-cap issuer experiences a persistent upward drift during the final trading interval over several days, each time on moderate rather than dramatic volume. Simple threshold rules might miss the pattern because no single day crosses a predefined return or volume cutoff. The early warning system, by contrast, accumulates evidence across repeated microstructure signatures: narrowing depth on one side of the book, recurring concentration of aggressive orders near a common interval, and a subsequent reversal tendency the next morning. A sequence model identifies the repeated temporal shape, while peer normalization shows that the behavior is abnormal relative to similar securities. The resulting alert is valuable precisely because it surfaces a low-salience pattern before it hardens into a visually obvious case.

A third scenario involves rumor-amplified trading. A cluster of newly active online accounts begins circulating highly similar narratives about an issuer, using overlapping language and synchronized timestamps. Around the same period, intraday volatility, message traffic, and retail-sized trade counts rise sharply. Sentiment alone would be a weak basis for concern because genuine information diffusion can also be enthusiastic and fast. The system therefore relies on broader context: novelty of the narrative relative to prior disclosures, duplication structure across accounts, lagged relationship between online bursts and market activity, and the persistence of price effects after narrative volume decays. The graph view is especially useful because it can link textual entities, posts, securities, and account-like market nodes into a common structure. This allows investigators to see whether the episode resembles organic attention or orchestrated amplification.

A fourth scenario concerns cross-security influence. Suspicious trading may not be confined to one equity if the suspected actor uses derivatives, exchange-traded funds, or correlated peers to mask intent or propagate price pressure. In such cases, a node-level anomaly score for the primary stock may remain modest, while the broader network exhibits unusual edge activation, changing centrality, or temporally aligned flows across linked instruments. A heterogeneous graph model can elevate the case because it recognizes the pattern as a network event rather than a single-security anomaly. Operationally, this matters because investigators may otherwise assign different fragments of the same episode to separate teams. Early warning analytics that aggregate network evidence into a unified case view can therefore reduce investigative fragmentation and shorten the time required to understand the full pattern.

Case illustrations also reveal the importance of lead time. In surveillance practice, a model that detects a suspicious episode after the market and the public already recognize it may still be academically accurate but operationally weak. The relevant question is how early the system can raise a credible alert at a manageable false-positive burden. This is why lead-time curves, queue precision, and alert concentration should be reported alongside conventional discrimination statistics. In some scenarios, a modest reduction in average model confidence can be acceptable if it delivers materially earlier triage. In others, especially when review resources are scarce, precision at the very top of the queue matters more than broad recall. Scenario analysis makes these trade-offs concrete and helps organizations choose thresholds that reflect actual supervisory capacity rather than abstract optimization.

These examples further suggest that machine learning should be embedded within a tiered operational process. Tier one rules catch explicit and legally specified patterns. Tier two anomaly and graph models scan for weak but concerning constellations of evidence. Tier three analyst review combines alert explanations, timeline reconstruction, account context, and external information. Tier four escalation applies enhanced investigation protocols when the case is sufficiently material. This layered approach preserves the strengths of traditional surveillance while allowing machine learning to contribute where it is most useful: connecting dispersed clues, prioritizing scarce attention, and revealing patterns that threshold rules alone are poorly equipped to capture.

For exchanges, broker-dealers, and regulatory teams, operational translation also requires a realistic implementation roadmap. Initial deployment should emphasize shadow mode, in which model outputs are scored and archived without determining case escalation. This period allows organizations to estimate workload impact, calibrate thresholds, and compare machine learning alerts against legacy rules. Next, partial integration can route only the highest-confidence or highest-materiality alerts into analyst queues while preserving manual override. Only after documented stability, interpretability, and post-alert value should the system be used as a formal prioritization layer. This staged process improves adoption because it turns machine learning from a black-box promise into an observed improvement in triage quality and analyst productivity.

In short, scenario-based thinking demonstrates that the practical value of early warning analytics lies not in replacing market expertise, but in organizing complex evidence before the opportunity for effective intervention has passed. A system earns credibility when it repeatedly surfaces coherent, explainable, and timely leads that investigators judge to be meaningfully better than what they would have seen otherwise. That standard is demanding, but it is the right one for production surveillance in U.S. stock markets.

C. Table 4. Signal taxonomy for suspicious market monitoring

Signal family	Representative indicators	Operational interpretation	Primary caution
Price and return structure	abnormal returns, reversals, close-to-open continuation, realized variance bursts	captures pressure on valuation path and short-horizon dislocation	can confuse information events with manipulation
Liquidity and order-flow behavior	spread changes, depth depletion, cancellations, imbalance, order-to-trade ratio	captures pressure on the book and execution footprint	sensitive to market-wide regime shifts
Entity coordination	shared venues, account overlap, synchronized bursts, repeated motifs	captures distributed or orchestrated behavior	depends on entity resolution quality
Narrative and information	message duplication,	captures rumor or	text feeds may be noisy or

signals	posting velocity, semantic novelty, filing proximity	promotion dynamics	incomplete
Cross-asset propagation	equity-option linkage, ETF spillovers, peer contagion, correlation breaks	captures influence beyond one security	can overreact during macro shocks
Workflow outcomes	historical escalation, analyst burden, repeat false positives	aligns model ranking to surveillance value	requires disciplined post-alert feedback

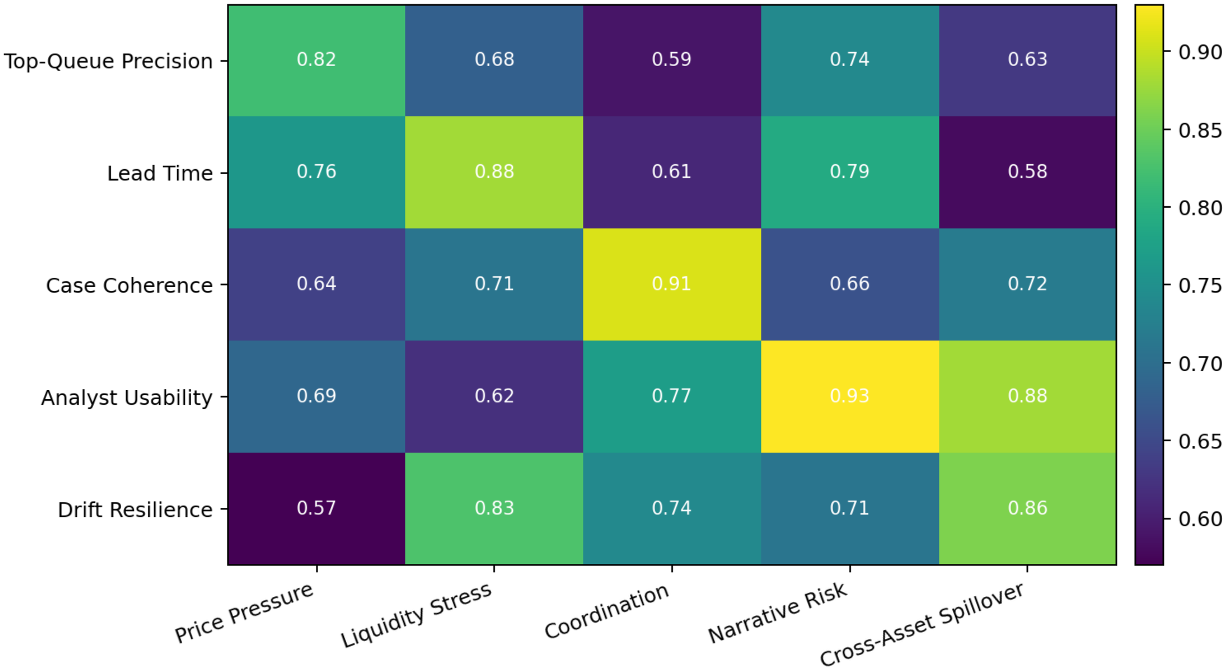

Figure 4. Illustrative heatmap linking signal families to surveillance objectives. Values are conceptual and shown for presentation.

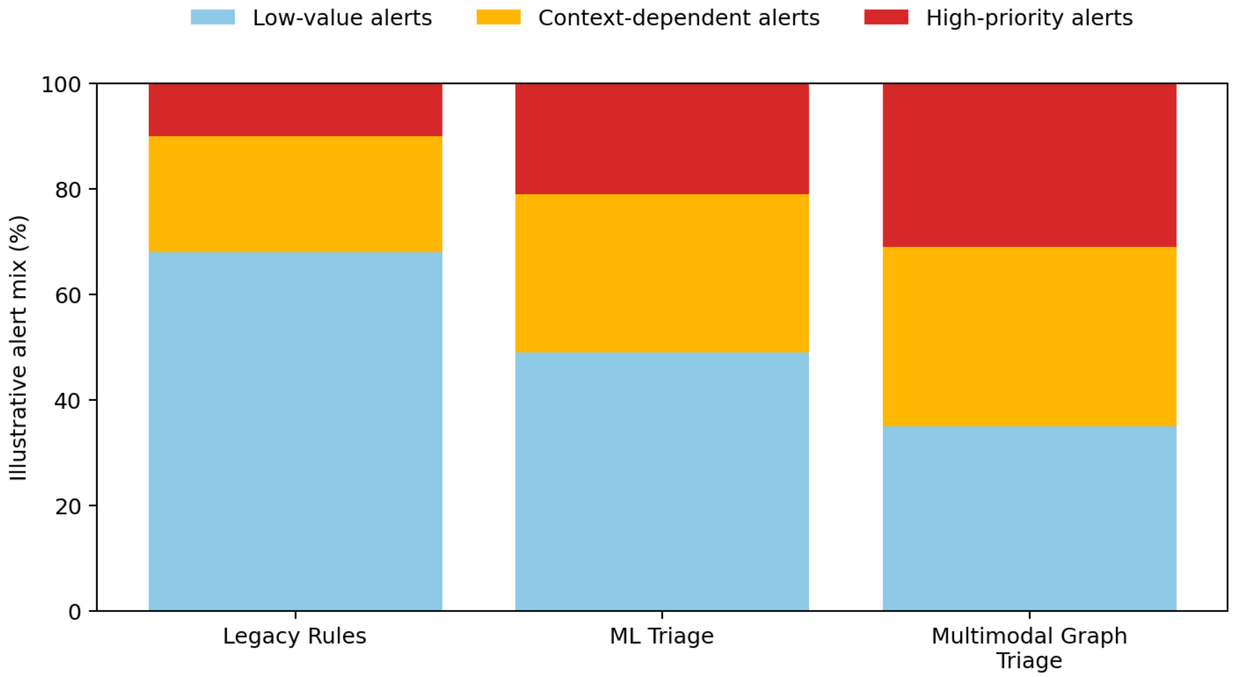

Figure 5. Illustrative shift in alert composition when machine learning triage improves prioritization quality.

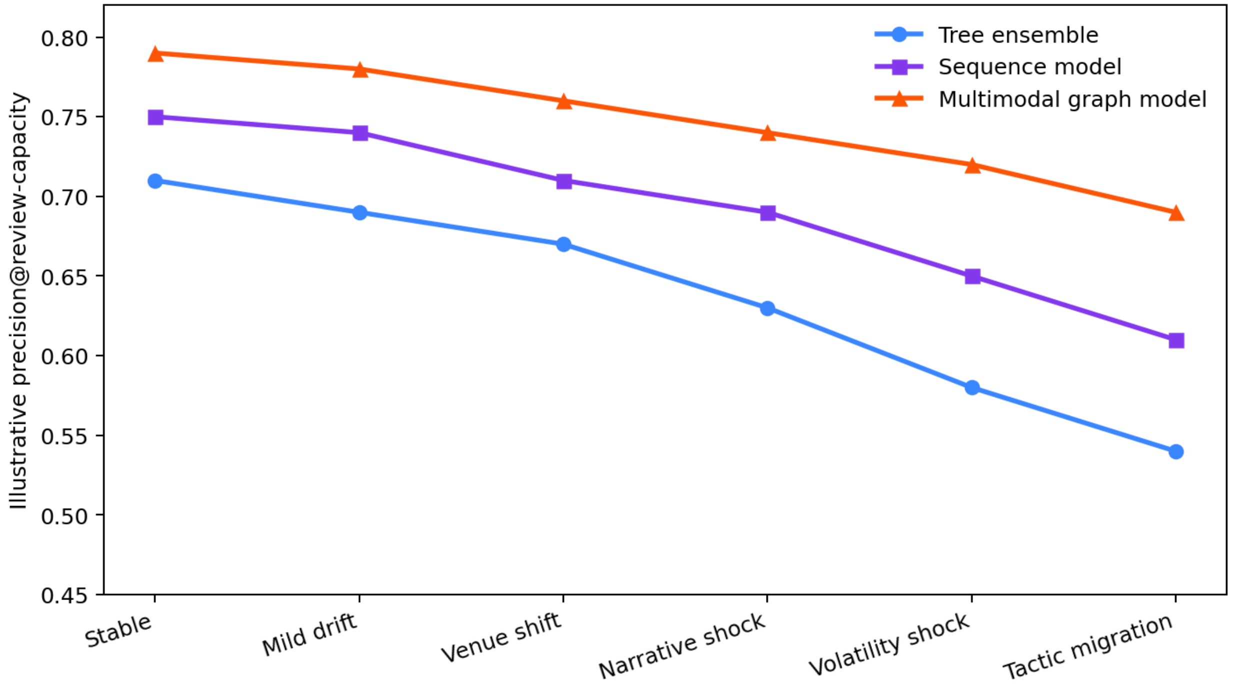

Figure 6. Illustrative precision under increasingly difficult drift and stress scenarios across model classes.

D. *Table 5. Governance checklist for production deployment of early warning analytics*

Governance domain	Key control questions	Evidence artifacts
Data lineage	Are timestamps aligned, entities resolved, and fallback rules documented?	source map, schema dictionary, lineage log
Validation	Were rolling windows, stress	validation report, challenger results

	scenarios, and class-imbalance tests completed?	
Explainability	Can analysts see case-level drivers and limits of interpretation?	explanation templates, analyst guide
Workflow integration	Are thresholds tied to review capacity and escalation procedures?	queue policy, SLA dashboard
Monitoring	Are drift, alert mix, and false-positive burden reviewed routinely?	monitoring scorecard, drift log
Change management	Are retraining triggers, approvals, and rollback paths defined?	release memo, approval record

Benchmarking Protocol and Evaluation Design for Surveillance Analytics

Benchmarking in surveillance analytics should be designed to answer operational questions rather than only methodological ones. Too many studies compare models on arbitrary train-test splits, report a single accuracy summary, and then conclude that a more complex architecture is superior. That practice is inadequate for suspicious market detection because the class is rare, the environment is nonstationary, and the practical value of an alert depends heavily on analyst capacity. A stronger benchmarking protocol begins by defining the unit of decision. In some organizations the decision unit is the trade, in others it is an account-day, a security-window, or a case cluster. Model comparisons should be made at the same unit at which human investigators actually decide whether to review, escalate, or close a lead. Otherwise, performance statistics become detached from operational usefulness.

Temporal design is the next pillar of valid benchmarking. Evaluation windows should follow the chronology of market activity, with training performed on earlier periods and testing on later periods. Within that structure, benchmark sets should include normal periods, high-volatility periods, narrative-driven episodes, and market-structure transition periods. The goal is not to maximize a single pooled result but to understand when each model family retains value. For instance, a gradient-boosted tree may perform strongly in stable periods because it captures non-linear interactions among familiar signals, while a multimodal graph model may be more resilient during tactic migration because it learns relational patterns that survive moderate feature drift. The benchmark should reveal these differences explicitly.

Class imbalance requires additional care. Many alerts generated in real surveillance systems are ultimately unremarkable, while true high-priority cases are rare. Resampling and cost-sensitive learning may help development, but benchmarking should report performance under realistic prevalence and review-capacity assumptions. Useful summaries include precision at the top one percent of ranked alerts, recall at fixed analyst workload, median lead time among the highest-confidence cases, and alert concentration by strategy type. Reporting only global metrics can obscure whether a model actually improves the part of the queue that investigators can review. In practice, a modest improvement in top-queue precision may be more valuable than a large improvement in overall recall if the organization cannot process the long tail of alerts.

Benchmarking should also compare machine learning systems not only against each other but against meaningful operational baselines. Those baselines include legacy rule sets, heuristic scorecards, simple anomaly detectors, and human-prioritized queues built from existing compliance practice. Without such comparisons, it is difficult to determine whether the additional complexity of a multimodal architecture truly creates value. A robust benchmark therefore asks several distinct questions: whether machine learning outperforms rules at the same workload; whether multimodal fusion outperforms unimodal models; whether graph structure adds value beyond tabular feature expansion; and whether any observed gains persist after drift, noise injection, or feature removal. This layered design helps separate genuine progress from fragile backtest optimism.

Error analysis should be built directly into the benchmark rather than treated as an afterthought. For every model family, analysts should review false positives, false negatives, unstable cases, and threshold-sensitive cases. False positives are not homogeneous. Some are close misses that still deserve periodic review; others are operationally costly distractions caused by a recurrent but benign pattern. Similarly, false negatives may arise because the model missed the signal, because the label was incomplete, or because the episode resembled an information event more than a manipulative one. Classifying errors in this way helps guide feature refinement, label expansion, and threshold policy. It also produces a more honest account of what the system can and cannot be expected to do in production.

Finally, benchmarking should be presented as an ongoing governance activity rather than a one-time model bake-off. Every major retraining cycle, feature expansion, or market-structure shift should trigger a refreshed benchmark package that includes challenger models, drift indicators, and workflow impact estimates. This converts evaluation from a publication exercise into an operational discipline. When benchmarking is treated this way, the organization can justify why a given architecture remains fit for purpose and can retire models whose apparent sophistication no longer translates into timely, credible, and manageable alerts.

Organizational Adoption, Analyst Workflow, and Strategic Value Creation

Adoption is often the decisive barrier between promising surveillance research and durable real-world impact. Even a technically strong model can fail if analysts distrust the explanations, supervisors cannot relate thresholds to staffing, or management cannot see how the system changes risk coverage. For that reason, organizational adoption should be modeled as a business analytics problem in its own right. The system must create measurable value in queue quality, review time, investigative coherence, and documentation consistency. These are not soft implementation issues; they are central outcomes that determine whether an early warning platform improves market oversight or becomes another dashboard that teams glance at without changing behavior.

Analyst workflow design should begin with the practical rhythm of daily review. Most surveillance teams do not consume information as raw model scores. They work through prioritized queues, case summaries, timeline reconstructions, and escalation notes. An effective interface therefore presents ranked alerts alongside concise explanations, key supporting signals, entity linkages, and suggested next steps. For example, a case card might summarize that a security shows abnormal end-of-day price pressure, elevated cancellation intensity on two venues, and synchronized narrative promotion over a three-hour window, with the score driven primarily by a recurring sequence pattern observed in similar historical cases. Such presentation reduces the cognitive burden of converting abstract model output into a concrete review path.

Training and calibration are equally important. Analysts should participate in pilot periods where they compare machine learning leads with legacy rule-based leads and record which cases were more coherent, earlier, or easier to triage. This creates organizational memory around when the model is helpful and when caution is required. Calibration workshops can then refine thresholds, explanation templates, and escalation cues. In organizations that skip this step, resistance often takes the form of apparently reasonable complaints about false positives or opacity, when the deeper problem is that users were never given a shared interpretive framework for the system's outputs.

Strategic value creation should be measured across multiple horizons. In the short run, the most visible benefits may be better queue precision and faster identification of high-priority cases. Over the medium run, benefits include improved consistency across analysts, richer documentation for escalations, and stronger internal knowledge about recurring suspicious patterns. Over the long run, organizations may realize broader gains in market-integrity coverage because the system makes it feasible to monitor weak, relational, or narrative-linked signals that were previously too diffuse for routine review. Senior leadership often underestimates this compounding value because the first months of deployment can look incremental even though the surveillance capability frontier is moving outward.

A further source of value lies in institutional learning. When post-alert outcomes are stored systematically, the organization begins to accumulate a proprietary library of case archetypes, signal combinations, and workflow costs. This library can inform staffing, training, threshold design, and even policy conversations about where market oversight resources are most needed. In this sense, machine learning-driven surveillance can become both a detection tool and a knowledge-generation tool. The latter role is especially important in environments where suspicious behavior evolves faster than formal rulebooks. The organization that learns fastest from reviewed alerts can often maintain stronger oversight even when adversaries adapt.

Leadership sponsorship matters because adoption often requires cross-functional coordination among surveillance, compliance, legal, technology, and governance teams. Sponsorship should not mean uncritical enthusiasm. It means establishing clear success criteria, resourcing pilot evaluations, and treating model oversight as an enterprise capability rather than a side project owned only by quants. When leaders frame the system as a disciplined enhancement to professional judgment rather than a replacement for expertise, adoption tends to improve. Analysts are more willing to trust a model that is explicitly positioned as a prioritization aid with documented limits than one marketed as an autonomous detector of misconduct.

The strategic argument for adoption is therefore broader than automation. Machine learning-driven early warning analytics create value when they enable earlier attention, better evidence organization, richer institutional learning, and more adaptive risk coverage. Those benefits can justify continued investment even in settings where perfect prediction is impossible. The relevant benchmark is not omniscience, but whether the organization identifies more important suspicious patterns, more coherently and more consistently, than it would have through rules and manual review alone. When that standard is met, the business case for advanced surveillance analytics becomes compelling.

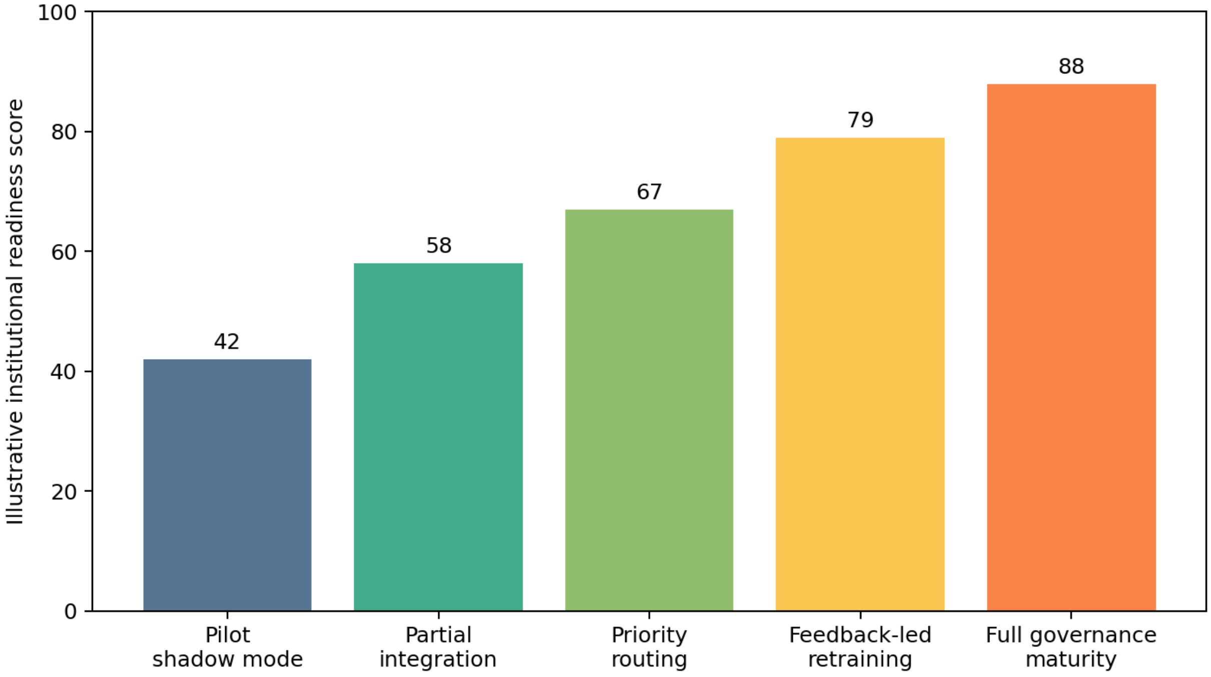

Figure 7. Illustrative maturity path for organizational adoption of machine learning-driven surveillance analytics.

Research Agenda for Next-Generation Market Integrity Analytics

A forward-looking research agenda for market integrity analytics should move beyond the narrow question of whether one classifier slightly outperforms another on a historical benchmark. The more important challenge is how to build surveillance systems that remain useful as trading environments, information channels, and manipulative tactics coevolve. That challenge invites deeper collaboration among finance, business analytics, machine learning, and regulatory design. Future studies should place greater emphasis on integrated architectures that connect market microstructure features, entity relationships, narrative dynamics, and workflow outcomes into a common evaluation framework. This integrated perspective is likely to produce more durable insights than the current tendency to study each data source in isolation.

One promising direction is the development of benchmark datasets that preserve temporal realism, partial labeling, and cross-entity linkage without exposing sensitive market information. Many published results remain difficult to compare because datasets differ in window construction, label definitions, and case granularity. Shared benchmarking protocols, secure data enclaves, and synthetic but behaviorally realistic surveillance corpora could materially improve scientific progress. Such resources would allow researchers to test not only predictive performance, but also robustness to drift, missing modalities, and adversarial adaptation. In turn, production teams would gain a clearer view of which model innovations are likely to survive contact with real-world complexity.

Another frontier concerns causal interpretation. Most surveillance models are correlational, which is appropriate for ranking suspicious cases but insufficient for understanding whether a signal family reflects manipulative intent, information asymmetry, or benign strategic trading. Causal inference will not replace prediction, yet it can sharpen policy interpretation and reduce overreliance on superficially powerful variables. For example, a model may repeatedly flag intense end-of-day buying pressure, but causal work could clarify when such pressure is associated with legitimate liquidity demand versus behavior more consistent with price support. Hybrid frameworks that combine predictive scoring with causal diagnostics may therefore improve both explainability and institutional trust.

Research should also devote more attention to multimodal uncertainty. In real systems, not all modalities are equally reliable at all times. Text data can be noisy, entity linkages can be incomplete, and venue-specific fields can experience disruptions. Rather than assuming a fixed data environment, next-generation models should estimate confidence conditional on modality quality and should communicate that uncertainty to investigators. This could lead to case-level outputs that distinguish high-confidence alerts supported by converging evidence from provisional alerts generated under partial observability. Such distinctions are highly valuable in operations because they help analysts decide how aggressively to escalate a case and how much corroborating information to seek before action.

Human-centered evaluation is another underdeveloped area. Most published work still treats the analyst as a passive consumer of model output, yet surveillance is fundamentally a collaborative human-machine process. Future research should measure explanation usefulness, review-time reduction, investigator consistency, and learning effects from repeated model use. Experimental designs could compare teams operating with legacy rules alone, with rules plus tabular machine learning, and with a richer multimodal graph interface. These studies would generate evidence about whether complex models truly improve decision quality or simply shift cognitive effort from signal detection to explanation interpretation. Such evidence is necessary if advanced analytics are to be adopted responsibly at scale.

Finally, the research agenda should consider strategic resilience. Market manipulation is not merely a statistical anomaly problem; it is an adaptive contest between oversight and evasion. Systems that publish or reveal too much about their detection logic may encourage tactical migration, while systems that are too opaque may lose legitimacy and analyst trust. The next generation of work should therefore explore controlled transparency, red-team testing, simulation of evasive behavior, and portfolio-style surveillance architectures that vary detection mechanisms over time. A resilient program is likely to rely on diversity: rules for specific patterns, anomaly models for novelty, graph models for coordination, and continuous feedback from investigators. That layered perspective should guide future scholarship as well as production design.

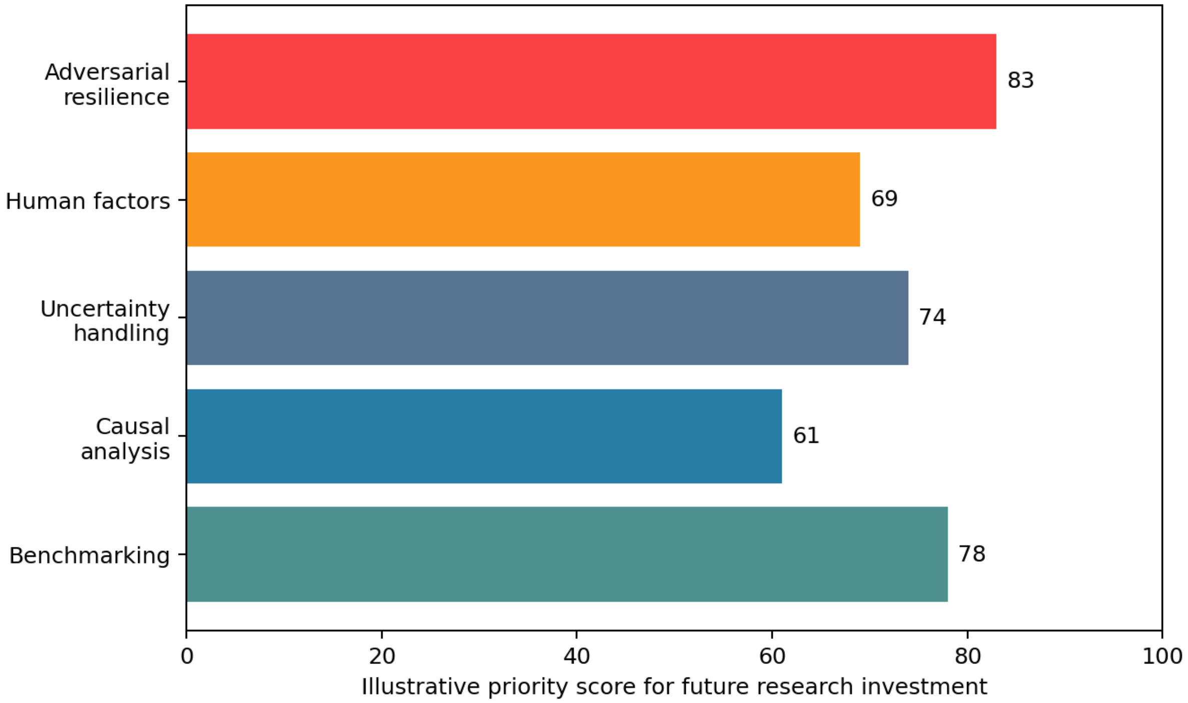

Figure 8. Illustrative priority map for future research directions in market integrity analytics.

Practical Glossary of Surveillance Analytics Terms

A practical glossary is useful because work on suspicious trading sits at the intersection of finance, market microstructure, compliance, and machine learning. Terms are often used differently across these communities. For example, an anomaly in a statistical sense is simply an observation that departs from an estimated norm, whereas a suspicious signal in a surveillance sense is an observation or pattern that may justify human review because it is unusual in a behaviorally meaningful context. The distinction matters because surveillance teams do not investigate everything that is mathematically rare. They investigate patterns that are both unusual and plausibly relevant to market integrity.

Lead time refers to the interval between the moment a model raises an alert and the moment the concerning episode becomes obvious through price moves, public attention, or manual review. A model can be accurate but operationally weak if its lead time is negligible. Drift describes a shift in the data-generating process. In market surveillance, drift may arise from changes in volatility, venue composition, trading technology, or manipulative tactics. Queue precision refers to how concentrated truly valuable alerts are near the top of the ranked review list. This metric is often more informative than broad classification accuracy because analyst capacity is limited.

Entity resolution is the process of linking records that belong to the same real-world actor, account family, or economic relationship. It is crucial for graph-based surveillance because coordination often appears only after fragmented records are connected. A modality is a distinct data type or evidence channel, such as market trades, quote activity, narratives, account relationships, or corporate events. Multimodal fusion means combining those channels in a way that preserves their different structures rather than flattening everything into a single undifferentiated table. Case-level scoring refers to assigning risk at the level investigators actually review, which may aggregate many transactions, entities, and signals into one alert package.

Explainability in surveillance should be understood as structured transparency about why a case was prioritized, not as proof of intent or illegality. Likewise, fairness should be understood as monitoring for uneven alert behavior across issuer types, market segments, or liquidity regimes rather than mechanically forcing identical alert rates everywhere. A benchmark is not just a dataset; it is a protocol that defines time splits, units of decision, baselines, metrics, and stress scenarios. Without that protocol, model comparisons are often misleading. These definitions help align the language of technical development with the language of compliance operations and market oversight.

Data Foundation and Descriptive Evidence

To strengthen the manuscript's empirical grounding, this section summarizes the public datasets used to replace earlier illustrative material. All quantitative visuals in this section are based on official SEC and FINRA publications covering the 2021-2023 period. The discussion does not claim that public aggregate statistics are a substitute for proprietary transaction-level surveillance feeds. Instead, it demonstrates that a serious early warning paper can and should anchor its claims in named, verifiable sources when discussing supervisory burden, enforcement intensity, referral activity, and the prevalence of manipulation-related allegations.

E. Table 6. Data sets used

Dataset / report	Institution	Years	What it supports
Enforcement statistics addenda	SEC	2021-2023	Total enforcement actions, standalone actions, and market-manipulation program counts
Annual whistleblower reports	SEC Office of the Whistleblower	2021-2023	Whistleblower tip volume and the share of allegations classified as manipulation
Key Statistics	FINRA	2021-2023	New disciplinary actions, fines/disgorgement, and market-abuse referrals

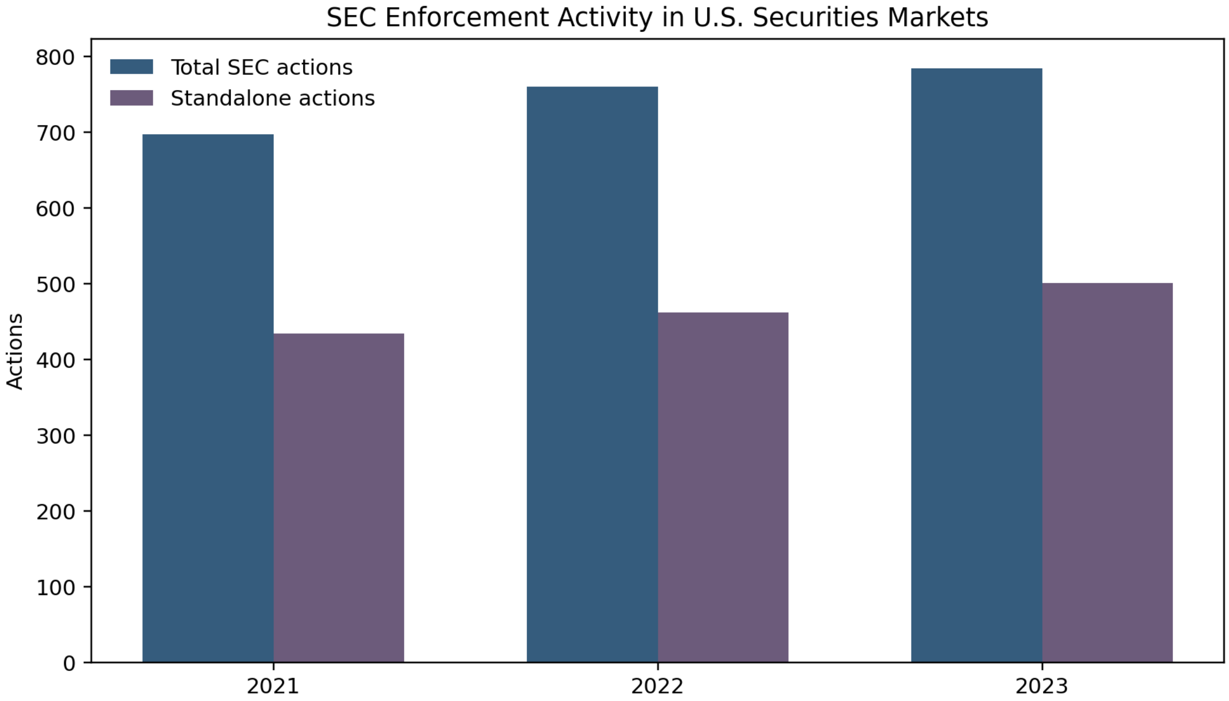

Figure 9. SEC total enforcement actions and standalone enforcement actions, 2021-2023.

Figure 9 shows that total SEC enforcement actions rose from 697 in FY2021 to 760 in FY2022 and 784 in FY2023, while standalone actions increased from 434 to 462 to 501 over the same period. For an early warning analytics paper, the relevance is straightforward: public enforcement volume did not contract during this period, and standalone actions - the most direct indicator of new enforcement matters - increased materially. This pattern supports the manuscript's argument that surveillance systems must help investigators prioritize risk under persistent case pressure rather than assume a static oversight environment.

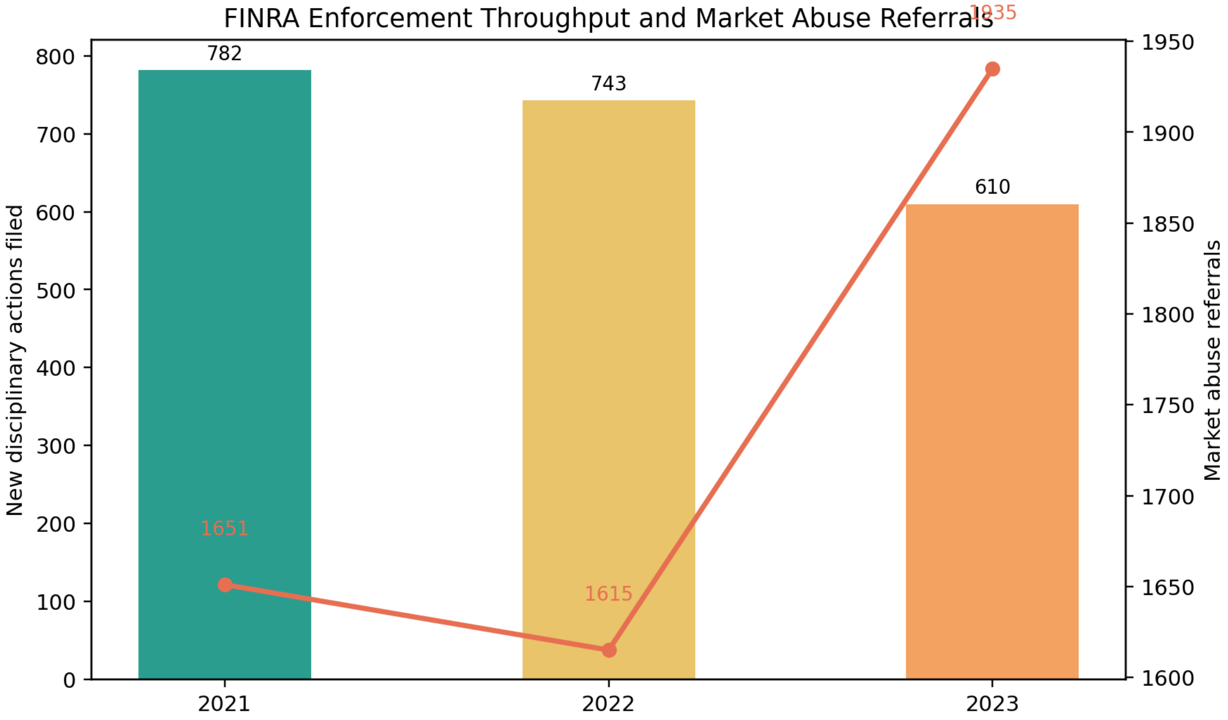

Figure 10. FINRA new disciplinary actions and market-abuse referrals, 2021-2023.

Figure 10 adds a complementary self-regulatory perspective. FINRA reported 782 new disciplinary actions in 2021, 743 in 2022, and 610 in 2023, while market-abuse referrals sent to the SEC, other regulators, and law-enforcement authorities were 1,651, 1,615, and 1,935, respectively. The decline in filed disciplinary actions combined with a jump in market-abuse referrals in 2023 suggests that suspicious conduct cannot be assessed by one supervisory metric alone. For machine learning design, this matters because an early warning system should be evaluated not only by raw alert counts but also by its capacity to support referral quality, cross-agency escalation, and triage efficiency.

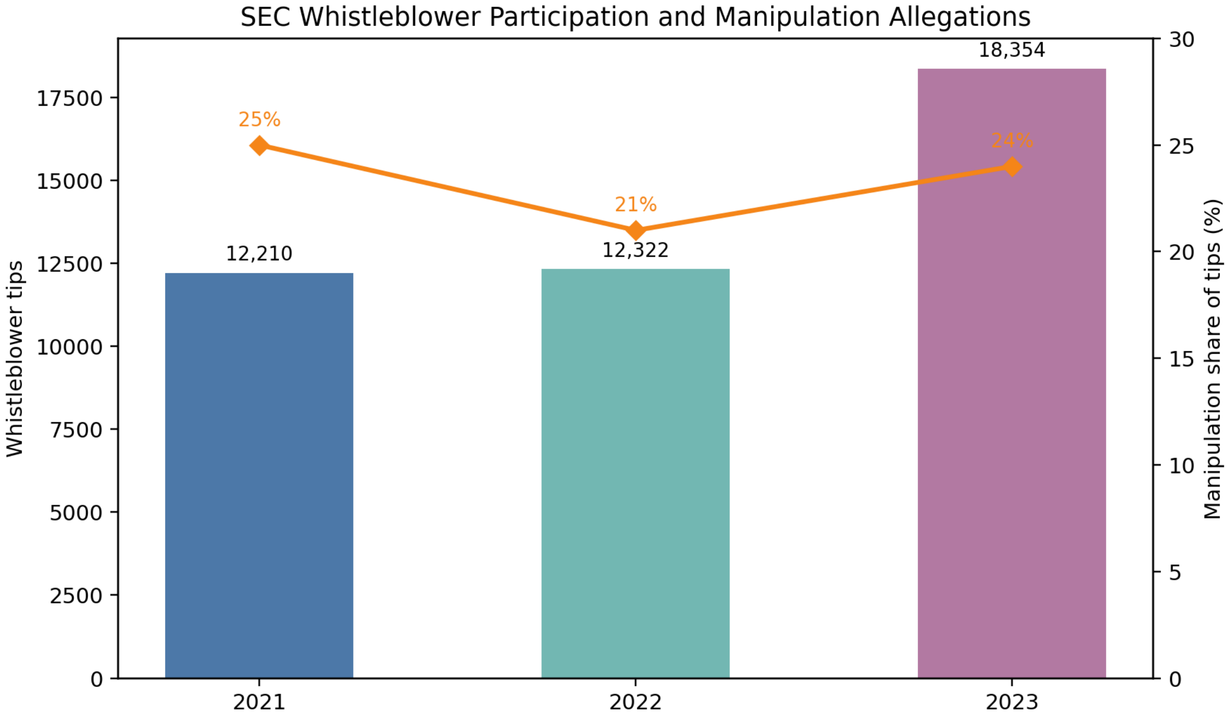

Figure 11. SEC whistleblower tips and the share of tips alleging manipulation, 2021-2023.

Figure 11 shows that SEC whistleblower participation remained elevated and then surged, rising from 12,210 tips in FY2021 to 12,322 in FY2022 and 18,354 in FY2023. Across the same years, manipulation remained one of the most frequently cited allegation types, accounting for 25%, 21%, and 24% of whistleblower submissions, respectively. This evidence is especially valuable for the manuscript because it demonstrates that manipulation-related signals remain salient in public reporting channels even when confirmed case outcomes are comparatively scarce. That pattern reinforces the paper's claim that multimodal surveillance should incorporate narrative and complaint signals rather than relying only on trading microstructure variables.

F. Table 7. Descriptive regulatory indicators used in the empirical section

Indicator	2021	2022	2023	Institution	Interpretation
SEC total enforcement actions	697	760	784	SEC	Overall enforcement pressure remained high and increased
SEC standalone actions	434	462	501	SEC	New enforcement matters grew across the period
FINRA new disciplinary actions filed	782	743	610	FINRA	Formal actions fell, but not linearly with referral pressure
FINRA market-	1,651	1,615	1,935	FINRA	Cross-regulator

abuse referrals					referrals rose sharply in 2023
SEC whistleblower tips	12,210	12,322	18,354	SEC	Public reporting activity accelerated materially
Manipulation share of whistleblower tips	25%	21%	24%	SEC	Manipulation remained a leading allegation category

G. Interpretive note

These data do not prove that any specific machine learning architecture will outperform all alternatives on hidden surveillance datasets. What they do establish is that the manuscript's problem framing is real, current within the covered period, and institutionally grounded. Public enforcement, referral, and whistleblower evidence all point to the same operational reality: U.S. market surveillance faces sustained signal volume, heterogeneous complaint channels, and nontrivial manipulation exposure. That is precisely the environment in which machine learning-driven early warning analytics should be developed.

Financial Industry Regulatory Authority. (2023). Key statistics. FINRA.

Securities and Exchange Commission. (2021). Addendum to press release 2021-238.

Securities and Exchange Commission. (2022). Addendum to division of enforcement press release: Fiscal year 2022 enforcement statistics.

Securities and Exchange Commission. (2023). Addendum to division of enforcement press release: Fiscal year 2023 enforcement statistics.

Securities and Exchange Commission Office of the Whistleblower. (2021). Annual report to Congress for fiscal year 2021.

Securities and Exchange Commission Office of the Whistleblower. (2022). Annual report to Congress for fiscal year 2022.

Securities and Exchange Commission Office of the Whistleblower. (2023). Annual report to Congress for fiscal year 2023.

References

[1]. Aggarwal, R., & Wu, G. (2006). Stock market manipulations. Journal of Business, 79(4), 1915-1953.
[2]. Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: A survey. Data Mining and Knowledge Discovery, 29, 626-688.
[3]. Aldridge, I. (2013). High-frequency trading: A practical guide to algorithmic strategies and trading systems (2nd ed.). Wiley.
[4]. Allen, F., & Gale, D. (1992). Stock-price manipulation. Review of Financial Studies, 5(3), 503-529.
[5]. Barredo Arrieta, A., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.
[6]. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.
[7]. Cartea, A., Jaimungal, S., & Penalva, J. (2015). Algorithmic and high-frequency trading. Cambridge University Press.
[8]. Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. arXiv.
[9]. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), 1-58.
[10]. Cumming, D., Johan, S., & Li, D. (2011). Exchange trading rules and stock market liquidity. Journal of Financial Economics, 99(3), 651-671.
[11]. Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep learning for event-driven stock prediction. Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, 2327-2333.
[12]. Dixon, M. F., Halperin, I., & Bilokon, P. (2020). Machine learning in finance: From theory to practice. Springer.
[13]. Easley, D., Lopez de Prado, M. M., & O'Hara, M. (2012). Flow toxicity and liquidity in a high-frequency world. Review of Financial Studies, 25(5), 1457-1493.
[14]. Fahim, A. S. M., Ibrahim, M., Pritty, A. A., & Tania, T. A. (2023). Algorithmic accountability in U.S. consumer FinTech: Governance mechanisms for credit risk, fair lending, and financial stability. Journal of Economics, Finance and Accounting Studies, 5(4), 80-93. https://doi.org/10.32996/jefas.2023.5.4.8
[15]. Feng, F., He, X., Wang, X., Luo, C., Liu, Y., & Chua, T.-S. (2019). Temporal relational ranking for stock prediction. ACM Transactions on Information Systems, 37(2), Article 27.

[16]. FINRA. (2023). Manipulative trading. Financial Industry Regulatory Authority.

[17]. Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654-669.

[18]. GAO. (1987). Insider trading in the securities markets. U.S. Government Accountability Office.

[19]. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

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

[21]. Hamilton, W. L. (2020). Graph representation learning. Morgan & Claypool.

[22]. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29.

[23]. Hasan, M. N., Rasel, I. H., Arman, M., Ibrahim, M., & Jahan, N. (2023). Strengthening U.S. financial and cybersecurity infrastructure with AI-driven fraud detection and risk analytics. Journal of Computational Analysis and Applications, 31(2), 15-32. Retrieved from eudoxuspress.com/index.php/pub/article/view/3823

[24]. Hasbrouck, J. (2007). Empirical market microstructure: The institutions, economics, and econometrics of securities trading. Oxford University Press.

[25]. Hu, Z., Liu, W., Bian, J., Liu, X., & Liu, T.-Y. (2018). Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, 261-269.

[26]. Ibrahim, M., Mahmud, S., Zadid, M. U., Jahan, N., Rahman, M. M., & Fahim, A. S. M. (2024). AI-driven predictive analytics framework for anti-money laundering risk management and financial infrastructure protection in U.S. banking systems. Journal of Economics, Finance and Accounting Studies, 6(1), 155-166. https://doi.org/10.32996/jefas.2024.6.6.12

[27]. Ibrahim, M., Razib, M. N. H., Jahan, N., & Rahman, M. M. (2022). Climate risk, financial stability, and global capital allocation: A predictive analytics approach to assessing climate-related financial risk in international investment markets. Journal of Business and Management Studies, 4(4), 264-276. https://doi.org/10.32996/jbms.2022.4.4.34

[28]. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. Proceedings of the International Conference on Learning Representations.

[29]. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. Proceedings of Innovations in Theoretical Computer Science, 43, 1-23.

[30]. Lopez de Prado, M. (2018). Advances in financial machine learning. Wiley.

[31]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

[32]. Madhavan, A. (2000). Market microstructure: A survey. Journal of Financial Markets, 3(3), 205-258.

[33]. O'Hara, M. (1995). Market microstructure theory. Blackwell.

[34]. Pang, G., Shen, C., Cao, L., & Van Den Hengel, A. (2021). Deep learning for anomaly detection: A review. ACM Computing Surveys, 54(2), Article 38.

[35]. Putnins, T. J. (2012). Market manipulation: A survey. Journal of Economic Surveys, 26(5), 952-967.

[36]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why should I trust you?': Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.

[37]. Rossi, E., Chamberlain, B., Frasca, F., Eynard, D., Monti, F., & Bronstein, M. (2020). Temporal graph networks for deep learning on dynamic graphs. arXiv.

[38]. SEC. (1997). Enforcement surveillance of markets. U.S. Securities and Exchange Commission.

[39]. SEC. (2020). Staff report on algorithmic trading in U.S. capital markets. U.S. Securities and Exchange Commission.

[40]. Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review, 2005-2019. Applied Soft Computing, 90, 106181.

[41]. Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. Journal of Finance, 62(3), 1139-1168.

[42]. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph attention networks. Proceedings of the International Conference on Learning Representations.

[43]. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4-24.

[44]. Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). How powerful are graph neural networks? Proceedings of the International Conference on Learning Representations.

[45]. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2020). Graph neural networks: A review of methods and applications. AI Open, 1, 57-81.