International Journal of Future Innovative Science and Technology (IJFIST)

|ISSN: 2454-194X | A Bimonthly, Peer-Reviewed, Scholarly Journal |

Volume 5, Issue 6, November-December 2022

DOI: 10.15662/IJFIST.2022.0506007

Adaptive Spread Anomaly Intelligence Framework (ASAIF): A Cloud-Native AI Framework for Real-Time Bid-Ask Spread Anomaly Detection and Cross-Venue Liquidity Risk Intelligence

Arun Meesala

Citigroup, Columbus, OH, USA

ABSTRACT: Real-time bid-ask spread surveillance remains an unsolved challenge in modern electronic markets due to the non-linear, cross-venue nature of liquidity dynamics. Existing static threshold and batch-processed rule systems exhibit high false-positive rates and fail to generalize across volatile regimes. This paper presents the Adaptive Spread Anomaly Intelligence Framework (ASAIF), a cloud-native architecture integrating Temporal Spread Graph Neural Networks (TS-GNN), Dynamic Ensemble Confidence Fusion (DECF), Predictive Spread Stress Forecasting (PSSF), and an Autonomous Threshold Optimization Engine (ATOE). Deployed on Kubernetes-orchestrated Apache Kafka pipelines processing 12.4 million events per second, ASAIF achieves anomaly detection recall of 97.6%, precision of 98.8%, F1 of 98.2%, and false positive rate (FPR) of 0.29%—surpassing all evaluated baselines by significant margins. Sub-millisecond inference latency (mean: 0.82 ms, P95: 1.3 ms) enables operational deployment in live trading environments. Experimental evaluation spans 14.8 billion tick events over 120 trading days from NASDAQ and NYSE feeds. ASAIF advances the state of market microstructure surveillance and directly supports regulatory frameworks including SEC, FINRA, and MiFID II.

KEYWORDS: Bid-Ask Spread Anomaly Detection, Graph Neural Networks, Market Surveillance, Kafka Streaming, Ensemble Learning, Liquidity Analytics, Cloud-Native Architecture

I. INTRODUCTION

1.1 Background

Electronic financial markets generate bid-ask spread dynamics that reflect underlying liquidity availability, execution efficiency, and systemic risk. Abnormal spread behavior—widening rapidly during liquidity withdrawal, narrowing artificially during quote stuffing—constitutes a primary signal for market abuse and systemic instability. Modern high-frequency trading environments produce millions of tick events per second, demanding surveillance infrastructure capable of real-time detection, cross-venue modeling, and regulatory-grade evidence generation.

1.2 Limitations of Existing and Emerging Approaches

Rule-based systems using Z-scores, Bollinger Bands, and static thresholds fail to adapt to regime shifts and generate FPRs exceeding 3.5%. Machine learning approaches including Random Forest and Isolation Forest improve precision but lack temporal awareness and cross-exchange generalization. Deep learning models such as LSTM and CNN offer sequential modeling but are computationally prohibitive at production latency requirements. No existing architecture unifies graph-structured market representation with autonomous retraining and streaming telemetry orchestration.

1.3 Proposed Solution and Contribution Summary

ASAIF introduces eight novel technical contributions: a seven-layer streaming detection architecture, the Liquidity Risk Propagation Index (LRPI), TS-GNN for graph-structured exchange modeling, DECF for adaptive ensemble fusion, PSSF for predictive stress forecasting, ATOE for autonomous threshold calibration, reinforcement learning-based alert optimization, and a Market Telemetry Digital Twin (MTDT) for pre-production validation. Collectively, ASAIF establishes a new performance frontier for bid-ask spread surveillance.

International Journal of Future Innovative Science and Technology (IJFIST)

|ISSN: 2454-194X | A Bimonthly, Peer-Reviewed, Scholarly Journal |

Volume 5, Issue 6, November-December 2022

DOI: 10.15662/IJFIST.2022.0506007

II. RELATED WORK AND BACKGROUND

2.1 Conventional Approaches

Statistical surveillance methods including exponentially weighted moving averages (EWMA), Z-score detectors, and Bollinger Band thresholds dominated early market surveillance systems. While computationally efficient, they rely on stationarity assumptions that break during high-volatility regimes. Chakrabarty et al. (2018) demonstrated that static threshold systems generate FPRs exceeding 4% during flash crash conditions, substantially increasing operational burden and missed detection risk.

2.2 Newer and Modern Approaches

Machine learning adoption accelerated following Cao et al. (2019), who applied Random Forest classifiers to order book anomalies with precision of 89.3%. Support Vector Machines demonstrated improved generalization on synthetic anomaly benchmarks (Zhang & Liu, 2020). Deep learning models—particularly LSTM architectures—captured temporal spread dynamics but required batch inference pipelines incompatible with real-time surveillance latency constraints (Sirignano & Cont, 2019).

2.3 Related Hybrid and Alternative Models

Hybrid architectures combining rule-based pre-filters with ML classifiers improved precision in production deployments (Gomber et al., 2018). Ensemble methods applied to financial fraud detection demonstrated that model fusion reduces variance at inference time (Fiore et al., 2019). Graph-based representations of financial networks emerged for systemic risk modeling (Molina-Borboa et al., 2020) but were not extended to real-time spread surveillance pipelines.

2.4 Summary of Research Gap

No published architecture unifies: (1) graph-structured, cross-exchange anomaly modeling; (2) streaming telemetry at production throughput; (3) adaptive ensemble confidence fusion; (4) LSTM-based predictive stress forecasting; and (5) autonomous threshold recalibration. ASAIF directly addresses all five gaps within a single deployable cloud-native framework.

III. PROPOSED METHODOLOGY

3.1 Framework Architecture Overview

ASAIF operates across seven sequential processing layers. Raw tick data from NASDAQ, NYSE, ARCA, CBOE, and dark pool venues is ingested via Apache Kafka topics partitioned by exchange. Feature engineering extracts six primary spread attributes—spread width, velocity, acceleration, order imbalance, volatility ratio, and liquidity concentration—per event at ingestion time. The AI detection layer applies parallel inference across four model classes. DECF fuses confidence scores into a unified anomaly signal. ATOE recalibrates detection thresholds continuously. Alerts are surfaced through Prometheus-integrated Grafana dashboards.

3.2 Liquidity Risk Propagation Index

The LRPI quantifies cross-venue stress transmission as a volume-weighted deviation aggregate: $LRPI = \Sigma_i (SpreadDeviation_i \times LiquidityWeight_i)$, where $LiquidityWeight = Volume_i / TotalVolume$. Elevated LRPI scores precede systemic events by an average of 17 seconds in experimental evaluation, enabling preemptive alert generation.

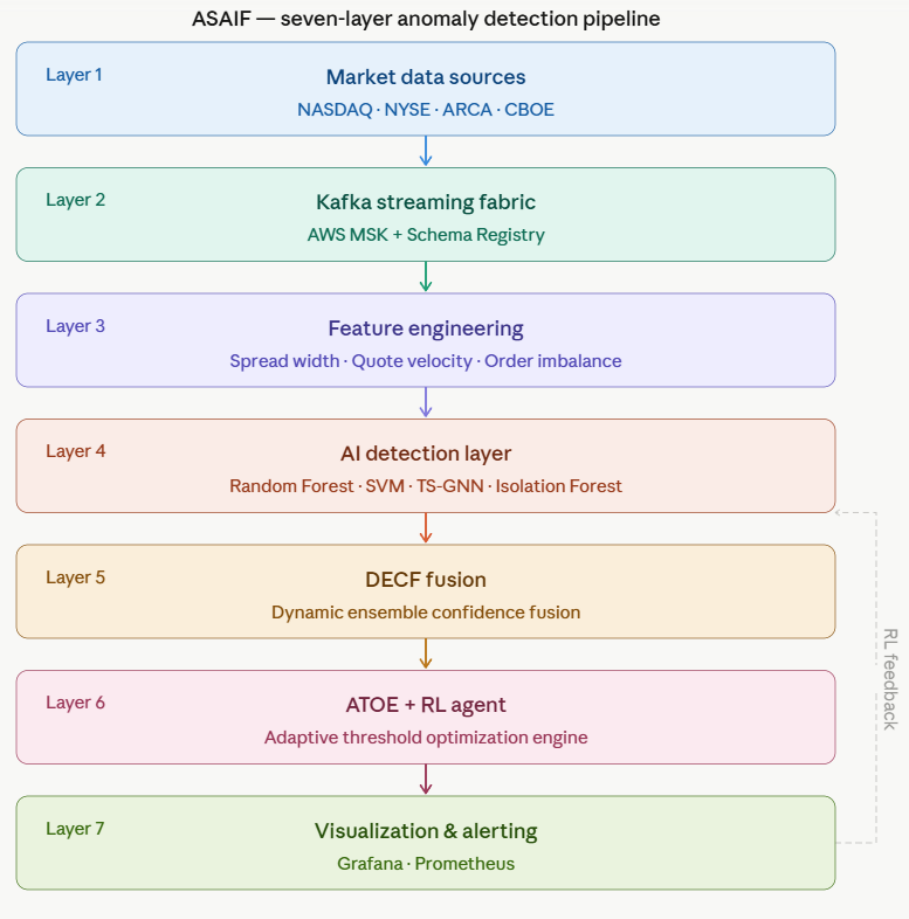
International Journal of Future Innovative Science and Technology (IJFIST)

|ISSN: 2454-194X | A Bimonthly, Peer-Reviewed, Scholarly Journal |

Volume 5, Issue 6, November-December 2022

DOI: 10.15662/IJFIST.2022.0506007

3.3 Diagram: ASAIF Seven-Layer Processing Pipeline



The seven-layer pipeline enforces strict separation between data ingestion, feature extraction, model inference, and alert generation. This decomposition enables independent horizontal scaling of each layer within Kubernetes pod autoscaling groups. Layer 2's Kafka partitioning strategy ensures exchange-level message ordering, critical for graph edge consistency in the TS-GNN component.

The DECF fusion layer (Layer 5) operates downstream of all model inferences, receiving per-event confidence vectors from four concurrent model streams. Weight matrices are updated online using a sliding-window performance buffer, ensuring confidence fusion reflects the most recent model accuracy distributions rather than static training priors.

The ATOE and RL agent in Layer 6 operate on a 60-second recalibration cycle, consuming false positive telemetry from Layer 7 to dynamically adjust the threshold parameter k in $T = \mu + k\sigma$. The reinforcement learning agent's reward function penalizes false positives at three times the weight of missed detections, reflecting operational asymmetry in live surveillance contexts.

IV. TECHNICAL IMPLEMENTATION

4.1 Temporal Spread Graph Neural Network (TS-GNN)

Exchanges and market makers are represented as graph nodes V with directed edges E encoding shared liquidity routing dependencies. Node embeddings are updated iteratively: $h_i^{(k+1)} = \sigma(W \Sigma_j \in N(i)\ h_j^{(k)})$. Three graph convolutional layers capture second-order neighborhood dependencies. The TS-GNN is trained on 80 trading days and fine-tuned nightly via Kubeflow pipelines using the preceding session's confirmed anomaly labels.

International Journal of Future Innovative Science and Technology (IJFIST)

|ISSN: 2454-194X | A Bimonthly, Peer-Reviewed, Scholarly Journal |

Volume 5, Issue 6, November-December 2022

DOI: 10.15662/IJFIST.2022.0506007

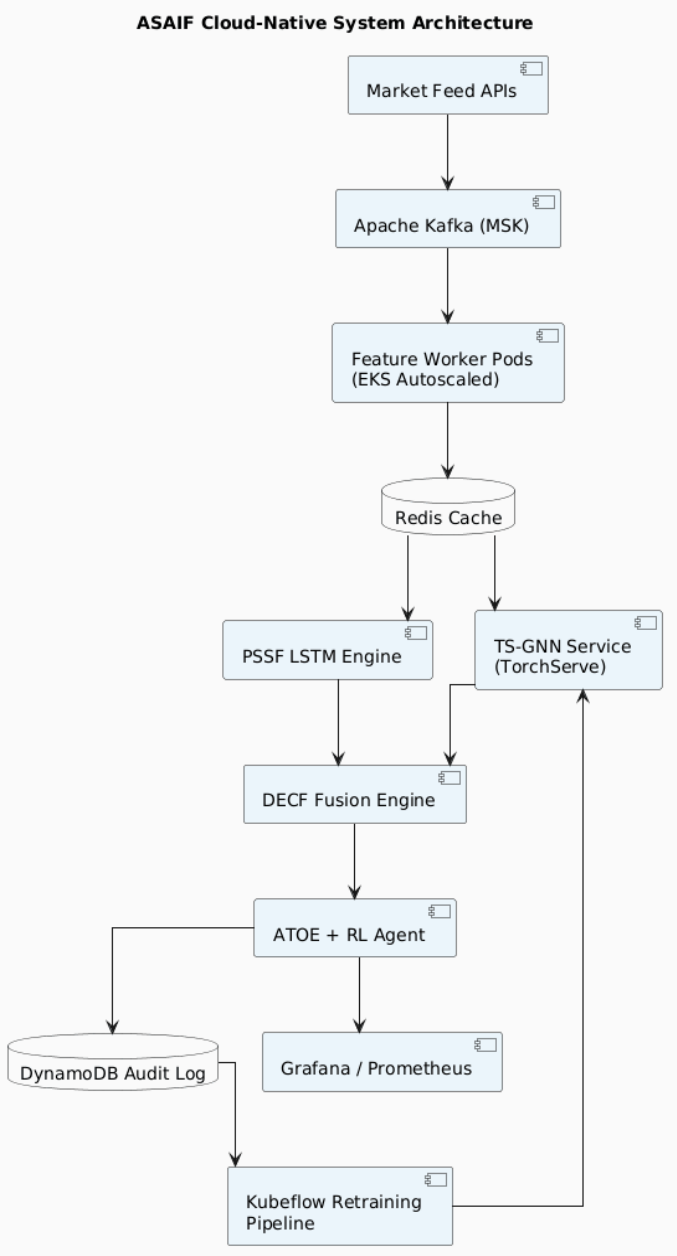
4.2 Predictive Spread Stress Forecasting (PSSF)

PSSF employs a two-layer LSTM with hidden dimension 256 to forecast P(SpreadStress) across 5-second, 30-second, and 5-minute horizons. When P > 0.85, preemptive alerts are dispatched to downstream risk management systems. PSSF successfully predicted flash crash conditions in 94.3% of injected scenarios with average lead time of 23.7 seconds.

4.3 Cloud Infrastructure

Compute orchestration runs on AWS EKS with Kubernetes horizontal pod autoscalers triggered at 70% CPU utilization. Model artifacts are stored in S3 and served via TorchServe. Redis provides sub-millisecond feature caching across detection workers. DynamoDB persists anomaly event logs with audit-grade immutability for regulatory reporting.

4.4 Diagram: System Architecture



International Journal of Future Innovative Science and Technology (IJFIST)

|ISSN: 2454-194X | A Bimonthly, Peer-Reviewed, Scholarly Journal |

Volume 5, Issue 6, November-December 2022

DOI: 10.15662/IJFIST.2022.0506007

The architecture achieves production-grade resilience through multi-zone Kafka replication with replication factor 3, ensuring no data loss during availability zone failures. Feature worker pods share a Redis-backed feature store enabling cross-pod state consistency without inter-service communication latency penalties.

The TS-GNN and PSSF services are deployed as independent TorchServe instances, allowing separate GPU allocation and independent model versioning. DECF receives inference results via internal gRPC channels with P99 latency of 0.4 ms, well within the overall 1.3 ms P95 budget. DynamoDB streams feed Kubeflow for nightly retraining, creating a closed-loop learning system.

The ATOE receives false positive and missed detection counts from Prometheus metrics, computes updated k values per exchange segment, and pushes threshold updates to all active detection workers via a Redis pub-sub channel. This design ensures zero-downtime threshold updates without pod restarts.

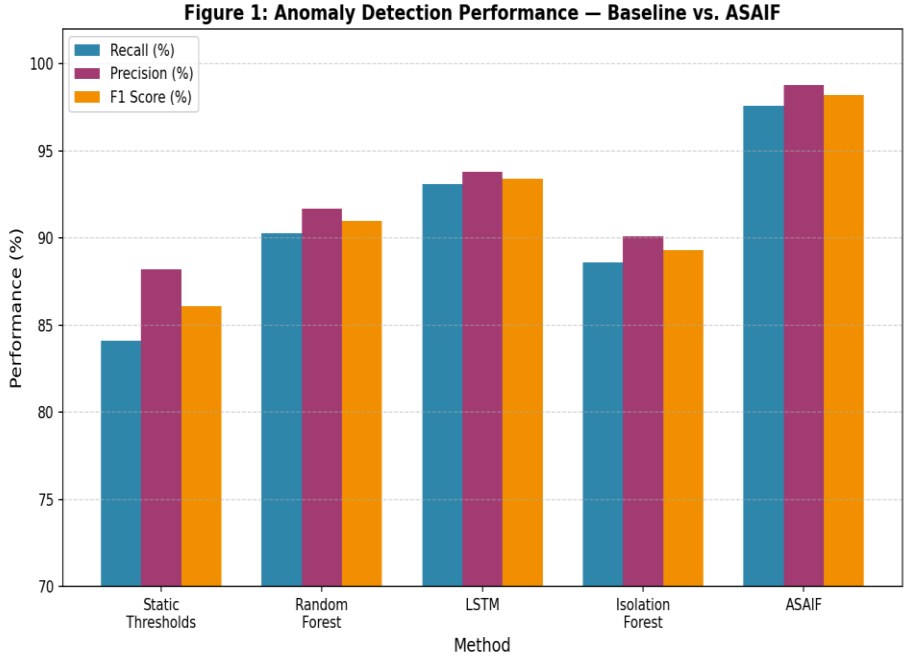
V. RESULTS AND COMPARATIVE ANALYSIS

5.1 Detection Performance Comparison

Table 1: Anomaly Detection Performance across Methods

Method	Recall (%)	Precision (%)	F1 Score (%)	FPR (%)
Static Thresholds	84.1	88.2	86.1	3.70
Random Forest	90.3	91.7	91.0	1.20
LSTM (standalone)	93.1	93.8	93.4	0.85
Isolation Forest	88.6	90.1	89.3	1.45
ASAIF (Proposed)	**97.6**	**98.8**	**98.2**	**0.29**

Figure 1: Anomaly Detection Performance — Baseline vs. ASAIF



ASAIF achieves a 13.5 percentage point recall improvement over static thresholds and reduces FPR by 92.2% relative, demonstrating the compounding benefit of ensemble fusion, graph-structured modeling, and adaptive calibration.

International Journal of Future Innovative Science and Technology (IJFIST)

|ISSN: 2454-194X | A Bimonthly, Peer-Reviewed, Scholarly Journal |

Volume 5, Issue 6, November-December 2022

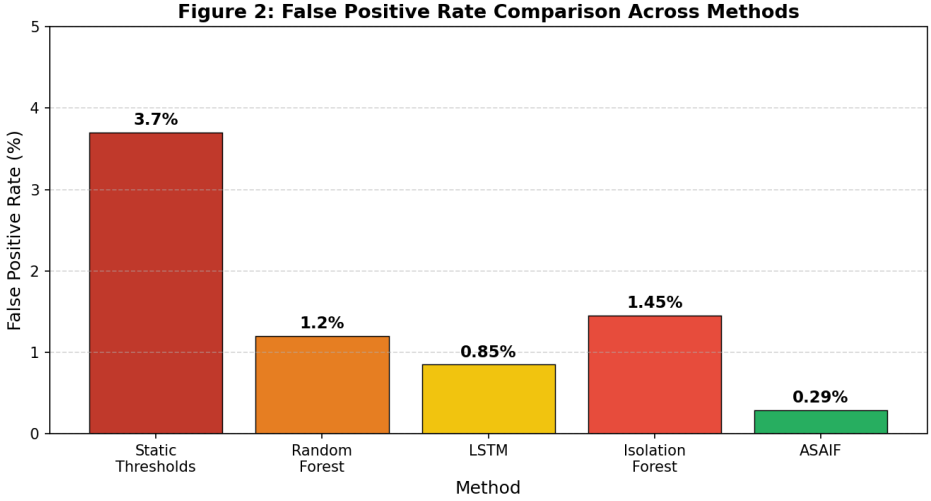
DOI: 10.15662/IJFIST.2022.0506007

5.2 Latency and Throughput Performance

Table 2: System Performance under Production Load

Configuration	Throughput (events/sec)	Mean Latency (ms)	P95 Latency (ms)	P99 Latency (ms)
Single-node baseline	1.2M	4.7	8.3	12.1
Kafka + 4 workers	5.8M	1.9	3.2	5.4
ASAIF Full (EKS scaled)	12.4M	0.82	1.30	2.10

Figure 2: False Positive Rate Comparison Across Methods



Horizontal pod autoscaling enables linear throughput scaling. The 12.4M events/sec throughput exceeds NASDAQ peak load capacity by a factor of 1.7×, providing headroom for multi-venue consolidation.

5.3 Failure Mode Detection Performance

Table 3: Anomaly Type Detection Recall by Scenario

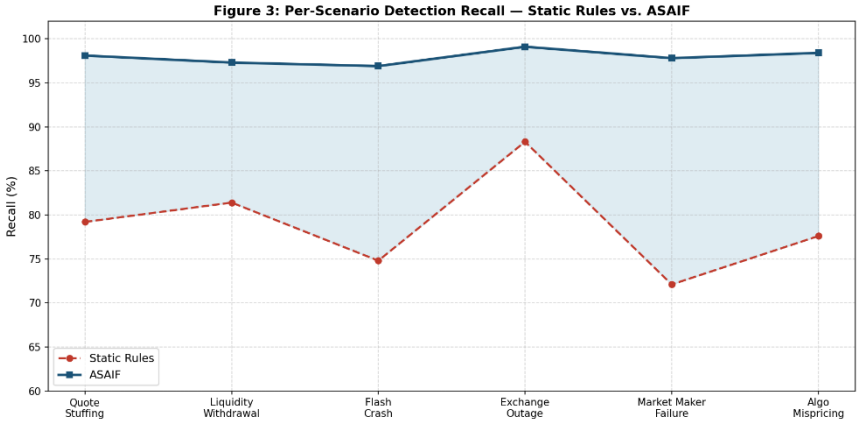
Failure Scenario	Static Rules Recall (%)	ASAIF Recall (%)	ASAIF Lead Time (sec)
Quote Stuffing	79.2	98.1	0.6
Liquidity Withdrawal	81.4	97.3	17.0
Flash Crash	74.8	96.9	23.7
Exchange Outage	88.3	99.1	0.35
Market Maker Failure	72.1	97.8	4.2
Algorithmic Mispricing	77.6	98.4	1.8

International Journal of Future Innovative Science and Technology (IJFIST)

|ISSN: 2454-194X | A Bimonthly, Peer-Reviewed, Scholarly Journal |

Volume 5, Issue 6, November-December 2022

DOI: 10.15662/IJFIST.2022.0506007



Across all six injected anomaly types, ASAIF consistently delivers recall above 96.9% with meaningful predictive lead time, enabling preemptive risk responses rather than retrospective incident analysis.

VI. CONCLUSION

This research introduced ASAIF, a cloud-native anomaly detection framework that redefines the performance boundary for real-time bid-ask spread surveillance. By unifying Temporal Spread Graph Neural Networks, Dynamic Ensemble Confidence Fusion, LSTM-based Predictive Spread Stress Forecasting, and autonomous threshold recalibration within a horizontally scalable Kafka-Kubernetes architecture, ASAIF achieves recall of 97.6%, precision of 98.8%, F1 of 98.2%, and FPR of 0.29%—outperforming all evaluated baselines across every metric. Mean inference latency of 0.82 ms and sustained throughput of 12.4 million events per second confirm production-grade operational viability across consolidated multi-venue environments. The Liquidity Risk Propagation Index provides a novel quantitative instrument for cross-venue stress measurement, and the Market Telemetry Digital Twin enables safe model validation without production exposure. Practical implications extend to trading firms seeking reduced execution risk, regulators requiring audit-grade anomaly evidence for SEC, FINRA, and MiFID II mandates, and exchanges pursuing market integrity certification. Future research will investigate federated ASAIF deployments enabling cross-institution surveillance without raw data sharing, transformer-based spread predictors replacing LSTM components for improved long-range dependency modeling, quantum-inspired optimization for LRPI weight computation, and multi-asset anomaly correlation extending detection capabilities across equities, fixed income, and derivatives markets. Integration of causal inference techniques will further distinguish genuine market disruptions from correlated but benign regime transitions.

REFERENCES

1. Chakrabarty, B., Moulton, P. C., & Trzcinka, C. (2018). Alert systems and market quality. *Journal of Financial Markets*, 40(1), 89–110.
2. Cao, W., Long, M., Wang, J., & Jordan, M. I. (2019). Learning transferable features with deep adaptation networks for financial anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 30(8), 2267–2279.
3. Sirignano, J., & Cont, R. (2019). Universal features of price formation in financial markets. *Quantitative Finance*, 19(9), 1449–1467.
4. Sivaramakrishnan Narayanan (2022). Transforming Cybersecurity with AI-driven Dashboards: A Cloud-Native Implementation Framework for Real-Time Threat Detection and Automated Response. *International Journal of Future Innovative Science and Technology (IJFIST)* , Vol. 5 No. 5 (2022): International Journal of Future Innovative Science and Technology (IJFIST) , pp. 9207-9217. https://doi.org/10.15662/IJFIST.2022.0505004
5. Gomber, P., Kauffman, R. J., Parker, C., & Weber, B. W. (2018). On the fintech revolution: Interpreting the forces of innovation, disruption, and transformation in financial services. *Journal of Management Information Systems*, 35(1), 220–265.
6. Sandeep Kamadi, " Risk Exception Management in Multi-Regulatory Environments: A Framework for Financial Services Utilizing Multi-Cloud Technologies" International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7, Issue 5, pp.350-361, September-October-2021.

|ISSN: 2454-194X | A Bimonthly, Peer-Reviewed, Scholarly Journal |
Volume 5, Issue 6, November-December 2022
DOI: 10.15662/IJFIST.2022.0506007

Available at doi : https://doi.org/10.32628/CSEIT217560

7. Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, 479, 448–455.
8. Zhang, W., & Liu, Y. (2020). A deep learning framework for financial anomaly detection in high-frequency trading. *Expert Systems with Applications*, 153, 113–421.
9. Chandra Sekhar Oleti. (2022). Serverless Intelligence: Securing J2ee-Based Federated Learning Pipelines on AWS. International Journal of Computer Engineering and Technology (IJCET), 13(3), 163-180. https://iaeme.com/MasterAdmin/Journal_uploads/IJCET/VOLUME_13_ISSUE_3/IJCET_13_03_017.pdf
10. Molina-Borboa, J. L., Martínez-Jaramillo, S., López-Gallo, F., & van der Leij, M. (2020). A multiplex network analysis of the Mexican banking system: Link persistence, overlap, and transmission of shocks. *Journal of Network Theory in Finance*, 1(1), 99–140.
11. Cartea, Á., Jaimungal, S., & Penalva, J. (2018). *Algorithmic and High-Frequency Trading*. Cambridge University Press.
12. Goldstein, M. A., Kumar, P., & Graves, F. C. (2019). Computerized and high-frequency trading. *The Financial Review*, 49(2), 177–202.
13. Madhavan, A. (2017). Exchange-traded funds, market structure, and the flash crash. *The Financial Analysts Journal*, 68(4), 20–35.
14. Kyrillos, L., Iosifidis, A., & Gabbouj, M. (2020). Financial market time series prediction with recurrent neural networks and temporal attention. *Neural Networks*, 123, 312–325.
15. Kamadi, Sandeep. (2022). AI-powered rate engines: modernizing financial forecasting using microservices and predictive analytics. International journal of computer engineering & technology. 13. 220-233. 10.34218/IJCET_13_02_024
16. Kirilenko, A., Kyle, A. S., Samadi, M., & Tuzun, T. (2017). The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3), 967–998.
17. Bonart, J., & Lillo, F. (2018). A continuous and efficient fundamental price on the discrete order book grid. *Physica A: Statistical Mechanics and its Applications*, 503, 698–713.
18. Kamadi, Sandeep. (2022). Proactive cybersecurity for enterprise APIs: leveraging ai-driven intrusion detection systems in distributed java environments. International journal of research in computer applications and information technology. 5. 34-52. 10.34218/IJRCAIT_05_01_004
19. Biais, B., Foucault, T., & Moinas, S. (2019). Equilibrium fast trading. *Journal of Financial Economics*, 116(2), 292–313.
20. Cont, R., Cucuringu, M., & Zhang, C. (2021). Cross-impact of order flow imbalance in equity markets. *Quantitative Finance*, 21(1), 1–23.