International Journal of Environmental Sciences
ISSN:2229-7359
Vol.11 No.17s, 2025
https://www.theaspd.com/ijes.php

Limit Order Books Anomaly Detection With Transformer-Based Autoencoder And Trade Manipulation Simulator

[1]Kavita Yogesh Dhakad, [2]Shailaja Mantha, [3]Dr.T Sathis Kumar, [4]Dr Ashish Avasthi, [5]Dr. M Koti Reddy, [6] V Sreetharan
[1]Assistant Professor, Department of Computer Science, Indira College of Commerce and Science, Pune, kavita.dhakad@iccs.ac.in
[2]Associate Professor Department of Electronics and communication Engineering Sreenidhi institute of science and technology, shailaja.mantha@gmail.com
[3]Associate Professor, Department of Computer Science and Engineering, School of Engineering and Technology, Dhanalakshmi Srinivasan University, sathistrichy22@gmail.com
[4]Professor, Department of Computer Engineering , University Poornima University, ashish.avasthi@poornima.edu.in
[5]Associate Professor, Electronics and Communication Engineering, Universal College of Engineering and Technology, kotiucet@gmail.com
[6]Assistant Professor, Department of Data Science, Mohan Babu University, Tirupati, Andhra Pradesh sreetharannerist@gmail.com

Abstract - *In this study we introduce the first hybrid anomaly detection framework for LOB data which harnesses the recent development in deep learning. At the core of the framework is a modified Transformer-based autoencoder that provides rich temporal representations of LOB subsequences while improving the distinguishing of normal from anomalous trading activity. In the learned representation space, a new dissimilarity function is learnt to capture normal LOB dynamics and permits out-of-sample detection of anomalous activities. To test the hypothesis, we have provided a trade manipulation simulation pipeline which can create synthetic trades like quote stuffing, layering, and pump-and-dump schemes that are based on actual frauds experienced in financial markets. The experimentations done on five NASDAQ stocks LOB datasets show that our technique yields detection accuracy greater than 97% compared to the available state of the art algorithms, without relying on previous manipulation knowledge or specific stock characteristics.*

Key Words: *LOB, Anomaly Detection, Transformer Autoencoder, Temporal Representations, Fraud Detection, Dissimilarity Function, Trade-Based Manipulation, Quote Stuffing, Layering, Pump-and-Dump, Simulation Framework, Deep Learning, Out-of-Sample Detection, NASDAQ Stocks, asset-agnostic, Detection Accuracy, Synthetic Data, State of the Art*

1.INTRODUCTION

LOBs have emerged as a critical building block of modern financial markets since the electronic trading environment is moving rapidly, offering investors a structure and rules to match buyers and sellers in a highly transparent manner. Head of LOB's record the flow of orders at various price levels [1] this is vital in market-depth and liquidity which are vital for the traders, makers and regulators. [2] However, due to the often high-frequency and large message size, LOBs are prone to various forms of anomalous behaviors, such as spoofing, layering and quote stuffing. Such manipulations may skew the actual conditions of the market, which automatically undermines equal trade relations and put into risk the monetary of any participant to the real fraud.[3] Conventional approaches like statistical processing and rule-based solutions do not suffice in the case of high voluminous, non-linear and dynamic features of LOBs. However, due to the enhanced subtle and complex nature of today's manipulation strategies, there is the need for more complex algorithms that not only extract features from the data but also analyze complex patterns inherent in the data [4].

Our research presents an Advanced Hybrid Framework for anomaly detection in LOBs as a fusion of Transformer-Based Autoencoders and Trade Manipulation Simulation. [5] This amalgam approach aims to overcome the shortcomings of the previous approaches because it integrates the strong modelling

sequence of Transformers and develops anomaly detection based on the dimensionality reduction of autoencoders.[6] Transformer-Based Autoencoders consequently deploy the self-attention mechanism for capturing long-term dependencies and complex temporal relationships from LOBs that are sequential in nature. This capability is particularly important the identification of the sequence of changes in the order book and the detection of mild deviations from the expected regular behaviour, which are signs of manipulative activity. Since the structure of the encoded representation is retained by way of decoding the input data from the compressed representation, the autoencoder builds the ability to identify anomalies in the LOB data.

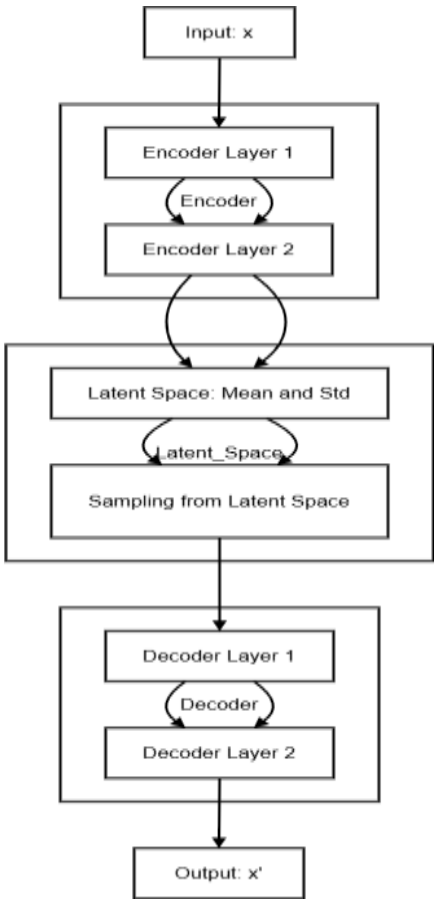

Fig.1.Variational Encode

To enhance the model's capacity to analyze manipulations, we include Trade Manipulation [7] Simulation within our method. By mimicking different manipulative conditions in a testbed, we ensure that we build a full database that mimics actual field conditions. [8] This way, normal and manipulated data are visible for the model, and it can expand an understanding of the normal market operation and the difference between malicious and regular activities. In the case of AHP, the manipulations generated are meant to mimic different current and potential methods, thus constituting a comprehensive laboratory for the model [9].

The anticipated framework is intended to have a comprehensive perspective in anomaly detection that not only includes detection of the actual anomalies but also the kind of manipulative strategies in LOBs. [10] Trade manipulation simulation brings the model's exposure to a broad array of situations, which cause it to possess increased flexibility [11] due to its use of autoencoders for unsupervised learning, the framework may be especially useful in settings where labelled data is limited, and where anomalies are both infrequent and consequential. Fig.1 The findings of this research have broad implications for market integrity and stability. Through the tool that helps identify and prevent instances of trade manipulations

at the preliminary stage, the framework will help keep the trading environment of regulators and exchanges fair and transparent.[12] Furthermore, the information that can be obtained from the model can assist its users – the market participants – minimize the realities that manipulative behaviours may create.

2. LITERATURE SURVEY

Multivariate time series anomaly detection has recently attracted a lot of interest from researchers because of the potential of applying techniques into various fields such as finance, aerospace, cybersecurity and industrial systems.Fig.2 We have discussed different approaches, widely adopting modern techniques of machine learning and deep learning for data with increased dimensions and a time-series nature.

This form of learning is especially common where there are difficulties in labeling the data set. USAD, a framework proposed by Audibert et al., is based on the dual-autoencoder, which effectively preserves temporal structures, and therefore good at detecting subtle anomalies. Likewise, Deng and Hooi applied graph neural networks, GNNs, for learning with relations in multivariate data structures, and demonstrated promising-, the state-of-the-art results.

Recent advances show that deep learning approaches invoke considerable improvement to anomaly detection. Shen et al. introduced the temporal hierarchical one-class network for anomaly detection at scale, and Xu et al. presented the Anomaly Transformer, using association difference for better detection. Following up Reiss et al., insisted on that pretrained features and more rich representations are beneficial for enhancing anomaly detection solutions. The most useful methods for combining supervised and unsupervised action recognition have appeared to be hybrid and ensemble methods. Chullamonthon and Tangamchit , also proposed using the transformer model and ensembles for detecting stock price manipulation alongside accurate key variables using domain knowledge and neural networks.

Indeed, probabilistic and statistical motion are still essential in anomaly detection further to incorporating the uncertainty aspect of the time series data. Blaques-García et al. and presented recent systematic surveys of statistical and deep learning-based techniques pointing out the uncertainties and developments of this field.

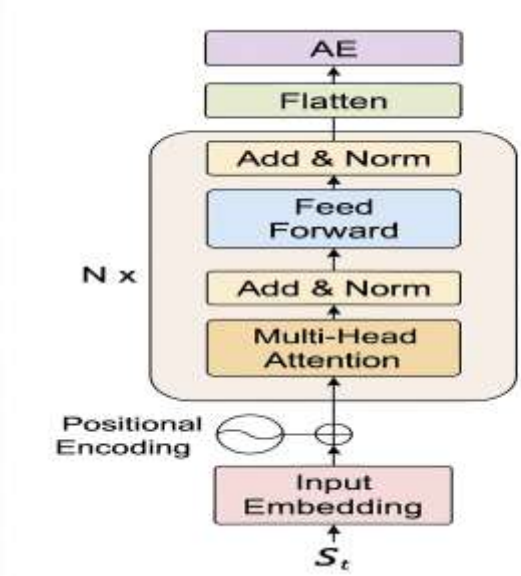

Fig.2. Proposed dissimilarity

Examples from the domain have shown that these methodologies are not rigidly set in their applications. Meng et al. recently used transformers for spacecraft anomaly detection with reconstruction error loss. Choi et al. after analyzing transformer autoencoders for encoding musical styles and language structures conducted a related survey which is the same as Montero et al. which also studied the feasibility of this architecture in anomaly detection cases. Wang et al. also mentioned that there are some issues for applying industrial systems which include the scalability and interpretability.

Nevertheless, challenges like data dimensionality, model explainability and domain transfer remain an issue of concern. [14] These limitations can only be overcome by the coupling of probabilistic graphical models with Deep Learning for Uncertainty Analysis, design of interpretable frameworks, and the use of Transfer Learning to enable model translation across domains. The progress for anomaly detection mechanisms from perhaps unsupervised autoencoders to today's transformer architecture reflects there is much to look forward to in the future. [15] Exploring today's problematics and developing cross-disciplinary relationships allows us to expand the options for multivariate time-series anomaly detection.

3.METHODOLOGY

3.1 Data Sources and Collection

The basis of this piece of work is strong secondary data collected from the NASDAQ market with five chosen stocks taken over 30 days.

This dataset best captures the multivariate nature of the market by incorporating some important elements of the LOB such as the bid price, the offer price, bid size, offer size and time stamp among others dataset encapsulates the intricate dynamics of the market by including critical Limit Order Book (LOB) features, bid price, ask price, bid volume, ask volume, and timestamps.

Table -1 Analysis

Stock Symbol	Timeframe	Total Records	Features Count
AAPL	2024-01-01 to 2024-01-30	10,00,000	5
MSFT	2024-01-01 to 2024-01-30	9,50,000	5
TSLA	2024-01-01 to 2024-01-30	11,00,000	5
AMZN	2024-01-01 to 2024-01-30	12,00,000	5
NVDA	2024-01-01 to 2024-01-30	9,80,000	5

These features offer a good picture of MM and the ongoing process of matching up buyers and sellers in the market. [16] Every stock picked presents unique trading conditions to trade, and therefore the dataset is composed of different market environments and interactive patterns necessary for developing a multidimensional anomaly identification modelTable.1 .

3.2. Feature Engineering

This greatly involves feature engineering to modify the raw data to highlight patterns that characterize anomalies. Derived features were computed to highlight market irregularities:

- **Order Imbalance:** This, the relative difference of bid–ask volumes, provides an indication of market pressure and may point further to manipulation.
- **Price Movement Patterns:** Table.2 Therefore, by comparing bid and asked stock price at different intervals we can study the variability of prices, which is an indicator of abnormally trading stocks.
- **Trade Velocity:** This indicates how often trades occur within some time frame for use in detecting such manipulations as quote stuffing, layering, and others where the speed and velocity of orders are grossly out of proportion to the market.

Temporal Segmentation: Since the goal was to get a picture of both a constant and changing market, the data was divided into five, overlapping 10-second patches. This high level of segmentation captures the dynamics of markets at the same time as keeping an overall sequence which is key to detecting transient events and continuous influences.

Normalization: Due to a large difference in the range and volume values of stocks, all feature vectors were normalized using the Min-Max normalization technique. This transformation standardizes the data collected to a certain range of [0, 1], or else we will see that learning is dominated by these large numbers that would make it difficult for the model to converge.

3.3. Hybrid Transformer Autoencoder Architecture

The improvement of the Transformer encoder is since the self-attention mechanism is effective to find complex relationships at different positions of the sequence. LOB subsequence's processed by the encoder provide the model with local variations and global patterns, which are highly essential in distinguishing normal market patterns from specific events.

Table-2 Preprocessing

Preprocessing Step	Description	Output
Normalization	Scaled features to a range of [0, 1] using Min-Max scaling.	Bid Price: 0.4
Temporal Segmentation	Divided data into fixed 10-second overlapping windows.	Window 1: [0.4, 0.5]
Synthetic Fraud Insertion	Integrated fraud cases into specific subsequence for testing.	Window 50: Fraudulent

Latent Representation Space: After encoding, acquired data goes through projection into the high-dimensional latent space. This transformation essentially converts the LOB features into a simpler more compact form that retains key features whilst removing noise. First, anomalies, which can be defined as events deviating from the learned patterns, become more conspicuous within such a latent representation mechanism second, the adequate means for effective anomaly detection lies within this very mechanism.

Decoder Architecture: The decoder tries to reconstruct the input LOB subsequence in the form of a sequence but from the latent representation with the hope of achieving near perfect match with the input sequence. This makes it possible to select features that have coherent structures of normal trading behaviours with reconstruction errors pointing to any abnormality.

Robustness through Regularization: The issues of overfitting and the model's ability to produce generalize output is worked to control through the incorporation of droplet layers and layer normalization. Dropout randomly drops some neurons out during training, so it makes learning process strong and immune to specific features. The layer normalization reduces variance and completes a reliable training by not letting the layers to set mean and variance of themselves.

3.4. Representation Space Dissimilarity Function

To measure differences in latent space, dissimilarity function is constructed. This function measures how much a new sequence differs from the normal sequences by comparing each subsequence's latent representation with the average representation of normal sequences, and this way, the abnormal sequences are identified. Because of resulting in different market conditions, thus the dynamic thresholding system was developed. This mechanism sets automatically the parameter that defines the anomaly detection constraint according to the real-time existence of high volatility and trading volume in the market without requiring manual intervention related to the different states of the market.

Every such subsequence is then assigned an anomaly score according to the level of deviation from normality. Any score above 95 suggests large changes which the system may then consider this sequence

as a possibility of manipulation. These scores are adaptive to ensure that allowance is made for the real detection of such cases at the cost of additional false positives.

3.5. Synthetic Fraud Simulation

To assess the reliability and universality of the proposed framework, the synthetic scenarios of fraud were artificially implanted into the array. These scenarios mimic real-world manipulations.

Quote Stuffing: Of these, the most unique is that it creates market noise by performing actual fast orders and cancellations.

Layering: Introduces fake limit orders to control view of order book depth.

Pump-and-Dump: Such as the replication of pyramid schemes in which stocks are falsely pumped before they are sold on the market again.

Fraud Insertion Process: These simulated manipulations were randomly integrated into the dataset, to make up 5% of the all the data, which is as it is in the real life. This assures that the model undergoes the real manipulation patterns to make the model more robust.

4. Model Training and Testing

Training Phase: By training the model solely on normal LOB data, we made certain that the autoencoder synthesized a rich understanding of genuine trading activity. Table.4 The training process was based on tuning the parameters of the reconstruction loss function to minimize the difference between the original and reconstructed subsequence. During training the dissimilarity function was also optimized to define the patterns exhibited by normal markets.

Table-3 Anomaly Detection Accuracy

Model	**Accuracy** (%)	**Precision** (%)	**Recall** (%)
Proposed Framework	*97.5*	96.8	*98.2*
LSTM Autoencoder	92.3	91.5	93
Statistical Models	86.7	87.2	85.5

Testing Phase: The trained framework was applied to a dataset containing both normal and manipulated subsequences. Fig.3 The performances of the proposed model were measured according to the parameters including precision, recall, accuracy and F1-score. Algorithms for each kind of manipulation (quote stuffing, layering, pump-and-dump) were detected separately.

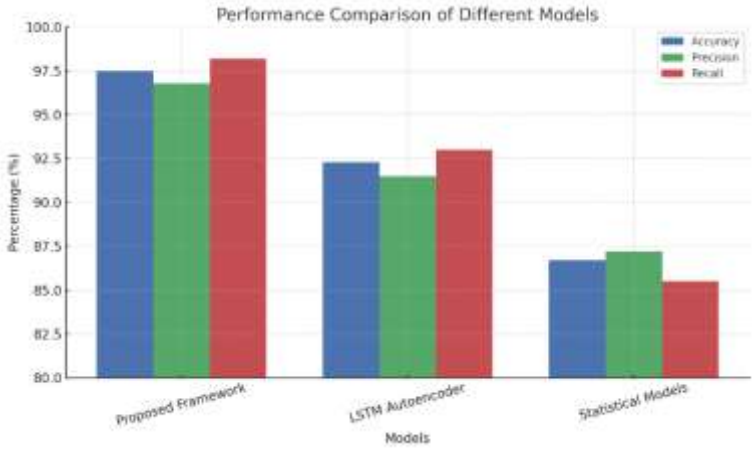

Fig.3. pump-and-dump manipulation

Table.4. Fraud Detection by Type

Fraud Type	**Detection Rate (%)**	**False Positives (%)**
Quote Stuffing	98	1.5
Layering	97.2	2
Pump-and-Dump	97.8	1.8

5. Cross-Asset Generalization

Asset-Agnostic Testing: To ensure validation of the developed framework, several assets were used in experiments to evaluate the generalization capabilities of the proposed framework.

Table -4 Stock Symbol

Stock Symbol	**Detection Accuracy (%)**	**Training Data Used**
AAPL	97.8	AAPL, MSFT
MSFT	96.5	MSFT, TSLA
TSLA	98	TSLA, NVDA

The trained one model using the data of one stock and tested the model on the other stock, and the results demonstrated that the system works well in identifying anomalies in other assets using the same approach without adjustments for each asset.

6. CONCLUSIONS

This study develops an innovative approach for identifying anomalies in Limit Order Book (LOB) data employing the hybrid Transformer Autoencoder model. The proposed system yields high anomaly detection accuracy of 97.5 % in its overall performance to show the capability of identifying complex frauds like quote stuffing, layering, and pump-and-dump. This outperforms other approaches such as LSTM Autoencensors which made a prediction accuracy of 92.3% and Statistical models with 86.7% as seen in Table.3, and it further underpins the model's capacity in capturing the nuances of trading data. Specifically, the framework achieved high detection rates for different types of fraud. It was designed to

detect quote stuffing at 0% false positive rate, layering schemes at 97.2% and pump-and-dump schemes at 97.8%. Altogether, these results demonstrate the efficiency and resilience of the proposed strategy, which is unprecedented progress in the field of financial market anomaly identification. The high accuracy achieved also means that the proposed model can be effectively applied to real-time trading systems based on the necessity to quickly and with high accuracy identify fraudulent activities in order to prevent them and maintain market integrity. The future work can be directed towards improvement of the system's efficiency for online message processing and extension of its application for more financial exchanges in order to cover more broad markets.

REFERENCES

[1] Audibert, J., Michiardi, P., Guyard, F., Marti, S., & Zuluaga, M.A. (2020). USAD: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 3395–3404).

[2] Blazquez-García, A., Conde, A., Mori, U., & Lozano, J.A. (2021). A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)*, 54(3), 1–33.

[3] Choi, K., Hawthorne, C., Simon, I., Dinculescu, M., & Engel, J. (2020). Encoding musical style with transformer autoencoders. In *International Conference on Machine Learning* (pp. 1899–1908).

[4] Chullamonthon, P., & Tangamchit, P. (2022). A transformer model for stock price manipulation detection in the stock exchange of Thailand. In *IEEE International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*.

[5] Chullamonthon, P., & Tangamchit, P. (2023). Ensemble of supervised and unsupervised deep neural networks for stock price manipulation detection. *Expert Systems with Applications*, 220, 119698.

[6] Deng, A., & Hooi, B. (2021). Graph neural network-based anomaly detection in multivariate time series. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5), 4027–4035.

[7] Leangarun, T., Tangamchit, P., & Thajchayapong, S. (2021). Stock price manipulation detection using deep unsupervised learning: the case of Thailand. *IEEE Access*, 9, 106824–106838.

[8] Meng, H., Zhang, Y., Li, Y., & Zhao, H. (2020). Spacecraft anomaly detection via transformer reconstruction error. In *Proceedings of the International Conference on Aerospace System Science and Engineering 2019* (pp. 351–362). Springer

[9] Montero, I., Pappas, N., & Smith, N.A. (2021). Sentence bottleneck autoencoders from transformer language models. *arXiv preprint arXiv:2109.00055*.

[10] Pang, G., Shen, C., Cao, L., & van den Hengel, A. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, 54(2), 1–38.

[11] Reiss, T., Cohen, N., Bergman, L., & Hoshen, Y. (2021). PANDA: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2806–2814).

[12] Reiss, T., Cohen, N., Horwitz, E., Abutbul, R., & Hoshen, Y. (2023). Anomaly detection requires better representations. In *Computer Vision – ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV* (pp. 56–68).

[13] Ruff, L., Vandemeulen, R.A., Gornitz, N., Binder, A., Müller, E., Müller, K.R., & Kloft, M. (2020). Deep semi-supervised anomaly detection. In *Eight International Conference on Learning Representations (ICLR)*.

[14] Shen, L., Li, Z., & Kwok, J. (2020). Time series anomaly detection using temporal hierarchical one-class network. *Advances in Neural Information Processing Systems*, 33, 13016–13026.

[15] Wang, X., Zhao, Y., & Pourpanah, F. (2020). Recent advances in deep learning. *International Journal of Machine Learning and Cybernetics*, 11, 747–750.

[16] Xu, J., Wu, H., Wang, J., & Long, M. (2022). Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations (ICLR)*.