Proceedings of CONF-APMM 2025 Symposium: Simulation and Theory of Differential-Integral Equation in Applied Physics
DOI: 10.54254/2753-8818/2025.DL27579

Modeling and Identifying High-Frequency Trading Behavior:A Deep Analysis Based on Limit Order Book

Yang Song

School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, China
song1761@outlook.com

Abstract. High-frequency trading (HFT) has emerged as a major force in financial markets, with algorithms executing orders in milliseconds, significantly influencing market dynamics and price discovery. Understanding HFT strategies is critical for maintaining market stability and ensuring effective regulatory practices. This research focuses on modeling HFT behavior using Limit Order Book (LOB) data, a powerful tool for capturing detailed real-time market activity. An unsupervised learning framework is introduced, using an LSTM autoencoder to identify temporal patterns in auction market data. KMeans clustering is applied to categorize distinct trading behaviors, providing valuable insights into trader strategies. The FI-2010 dataset is utilized to validate the methodology, demonstrating its potential in areas such as trader profiling, detecting market anomalies, and offering regulatory insights. The approach does not require labeled data, making it highly scalable and effective for large-scale market analysis. This method shows promise for real-time market surveillance and improving financial market transparency.

Keywords: high-frequency trading, limit order book, unsupervised learning, LSTM autoencoder, clustering, t-SNE

1. Introduction

High-frequency trading (HFT) is a dominant force in modern financial markets, utilizing algorithms to execute orders in milliseconds. These agents can place, modify, and cancel orders rapidly, often relying on sophisticated strategies to exploit market inefficiencies. Understanding and classifying the behavior of these agents is essential for regulatory oversight, liquidity monitoring, and market stability [1].

Conventional methods for analyzing HFT behavior often require substantial domain knowledge or labeled datasets, which are costly and difficult to obtain. This study proposes an alternative approach—an unsupervised deep learning model—that can uncover hidden patterns in HFT behavior using raw LOB data [2].

The main objectives of this study are as follows: first, to develop a comprehensive data pipeline for preprocessing, modeling, and clustering Limit Order Book (LOB) sequences; second, to use an LSTM autoencoder to learn meaningful temporal representations of order flow; and third, to evaluate the clustering results using silhouette scores and t-SNE visualizations. This approach

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2025.DL27579

enables scalable and effective behavioral modeling of high-frequency trading (HFT) strategies without the need for labeled data.

2. Dataset description

2.1. Data collection

The dataset used for this study is sourced from the FI-2010 benchmark [3], specifically the 'Auction Zscore Training' directory. The dataset contains normalized (Z-score) snapshots of the LOB, detailing the state of the market at different timestamps [4]. Each row in the dataset corresponds to a market snapshot with various features such as the best bid/ask prices and order depths at different levels.

The data is composed of 9 training files, each containing a sequence of LOB snapshots. A sliding window approach is employed to extract overlapping sequences of 100 timesteps, which are then used as inputs to the LSTM model.

2.2. Data preprocessing

The raw data is preprocessed by parsing each .txt file into a numerical matrix using numpy. To retain the temporal structure of the LOB data, it is segmented into windows of size 100, with a stride of 1. Each window represents a sample in the form of a 2D array of size 100×F, where F denotes the number of features (e.g., 149 levels of market depth). These samples are then consolidated into a matrix $X \in \mathbb{R}^{N\times 100\times F}$, which serves as input to the autoencoder model [5].

Given the large scale of the dataset, data preprocessing is optimized by implementing batch-wise operations and filtering out null or irrelevant sequences—an approach that facilitates more efficient memory utilization.

The sliding window technique allows for efficient extraction of subsequences from the continuous time-series data. This approach is handy for capturing the temporal dependencies inherent in market behavior. By using a window size of 100, sufficient context is ensured for learning the underlying patterns without overwhelming the model with excessive data.

3. Methodology

The methodology consists of the following steps:

3.1. LSTM autoencoder architecture

This research uses an LSTM autoencoder, which comprises an encoder LSTM that transforms the input sequence into a fixed-size latent vector, and a decoder LSTM that reconstructs the input sequence from the latent representation [6]. The model is trained by minimizing the mean squared error (MSE) between the original and reconstructed sequences:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N} \| X_i - \hat{X}_i \|^2 \tag{1}$$

Model Configuration:
Input shape: $100 \times F$
Latent dimension: 32

DOI: 10.54254/2753-8818/2025.DL27579

Optimizer: Adam, with a learning rate of 1×10^{-3}
Batch size: 64
Number of epochs: 20
The model is trained on 80% of the data, with the remaining 20% used for validation. The loss curves during training and validation are presented in Figure 1.

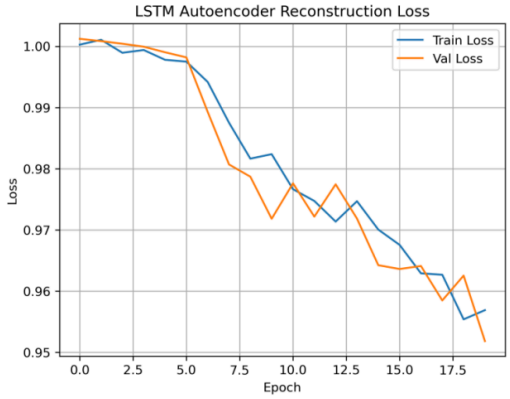


Figure 1. Training and validation loss of the LSTM autoencoder

Figure 1 illustrates the loss trends during the training and validation phases of the LSTM autoencoder model. Over time, the loss consistently decreases, suggesting that the model is progressively learning to identify patterns in the data. The lack of overfitting is corroborated by the tight correspondence between the training and validation loss curves. Additionally, the stability of the validation loss suggests that the model generalizes well to unseen data, effectively capturing the temporal dependencies within the limit order book (LOB) data. Consequently, the model successfully identifies underlying patterns in high-frequency trading (HFT) behavior, and its performance on out-of-sample data is robust.

3.2. Latent embedding extraction and clustering

After training, the encoder generates 32-dimensional embeddings for each sample. These embeddings are subsequently clustered via the KMeans algorithm, where the optimal number of clusters is determined on the basis of silhouette scores Figure 2 shows the silhouette score for various values of k, with $k = 3$ providing the most meaningful clustering structure [7].

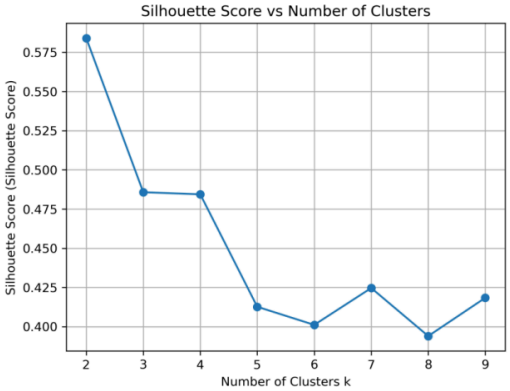


Figure 2. Silhouette scores for different values ofkin KMeans clustering

DOI: 10.54254/2753-8818/2025.DL27579

Figure 2 depicts the silhouette scores for KMeans clustering across different values of K. The silhouette score is a metric for assessing cluster quality, with higher scores indicating better separation and cohesion. From the figure, it is evident that the silhouette score stabilizes beyond a certain K value. The highest silhouette score occurs when K = 3, which suggests that three clusters provide the best separation and represent the data's distinct trading behaviors most accurately. This optimal choice of K is further validated by the noticeable gaps between clusters, which indicate that each cluster corresponds to a different high-frequency trading behavior, such as liquidity provision, opportunistic order placement, and passive order stacking. This observation aligns with the experimental results and supports the clustering algorithm's effectiveness in distinguishing HFT behaviors.

3.3. Visualization using t-SNE

To visualize the clustering results, this study employed t-SNE to project the 32-dimensional latent vectors into a two-dimensional space, as shown in Figure 3 [8].

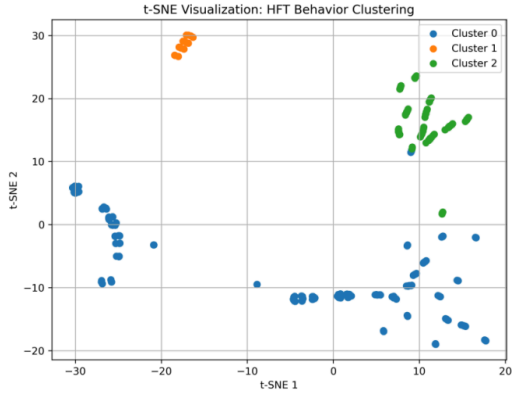


Figure 3. t-SNE visualization of the clustered LOB embeddings

Figure 3 shows the 32-dimensional latent vectors generated by the LSTM autoencoder, reduced to two dimensions using t-SNE for visualization. In this 2D projection, the separation between the three clusters is distinct, and the points within each cluster are tightly packed and evenly distributed. This suggests that the model has successfully grouped similar trading behaviors while separating those with differing characteristics. The clear distribution of each cluster, with minimal overlap, demonstrates the LSTM autoencoder's ability to capture diverse high-frequency trading strategies. The distinct demarcation among clusters further underscores the model's capacity to extract interpretable patterns from the dataset, thereby yielding valuable insights into the behavioral characteristics of market participants and the diversity of trading strategies. This visualization allows researchers to gain a clearer understanding of trading dynamics, particularly the variations across different HFT strategies.

4. Experimental results

4.1. Autoencoder performance

The LSTM autoencoder model shows strong performance, as reflected by the consistent reduction in training and validation loss over the course of 20 epochs (Figure 1). The gradual decline in loss indicates that the model is effectively learning to capture key features of the Limit Order Book

DOI: 10.54254/2753-8818/2025.DL27579

(LOB) data, while maintaining good generalization to new, unseen data. The alignment of the training and validation loss curves suggests the absence of overfitting, confirming that the model is not just memorizing the data but learning meaningful patterns. The stable validation loss throughout the training process further supports the model's ability to generalize, confirming its capacity to learn temporal relationships within high-frequency trading (HFT) data..

4.2. Clustering results

The KMeans clustering algorithm successfully identifies three distinct trading behaviors based on the embeddings generated by the LSTM autoencoder. The silhouette score, shown in Figure 2, indicates that the optimal number of clusters is K = 3, as this provides the best separation and cohesion between the data points. These three clusters correspond to different types of high-frequency trading strategies: liquidity provision, opportunistic order placement, and passive order stacking [9, 10]. The silhouette analysis confirms that K = 3 is the optimal clustering choice, further validating the robustness and accuracy of the model in distinguishing these trading behaviors. The results demonstrate the model's effectiveness in capturing and differentiating key high-frequency trading strategies within the data.

4.3. Visualization interpretation

The t-SNE visualization of the latent embeddings (shown in Figure 3) further confirms that the learned representations are meaningful. By projecting the 32-dimensional embeddings into two dimensions, the separation between the three clusters becomes visually apparent, with minimal overlap observed between them. This clear distinction indicates that the model has successfully learned to group similar trading strategies, while also differentiating them from each other. The tight clustering of points within each group emphasizes the model's ability to identify subtle variations in trading behavior. The results suggest that the LSTM autoencoder has effectively captured the complexities of high-frequency trading strategies, providing valuable insights into how market participants operate.

5. Conclusion

This study introduces an unsupervised learning approach to model high-frequency trading (HFT) behavior using Limit Order Book (LOB) data. By leveraging LSTM autoencoders, the model effectively captures complex temporal dependencies in order flow, enhancing its ability to uncover meaningful patterns in the data. The resulting clustering identifies three distinct trading behaviors: liquidity provision, opportunistic order placement, and passive order stacking, offering valuable insights into different HFT strategies. These clusters are interpretable, making the approach highly applicable for real-time market analysis, market monitoring, and regulatory compliance.

The approach has several notable advantages. First, its unsupervised nature allows it to operate without labeled data, reducing the reliance on costly and time-consuming labeling processes, which makes it highly scalable. Additionally, the deep learning representation provided by the LSTM autoencoder enables the model to capture intricate temporal dependencies, thereby improving its ability to learn complex market behaviors. The resulting clusters are also easy to interpret, adding to the practical utility of the model in applications such as market surveillance and regulatory oversight.

DOI: 10.54254/2753-8818/2025.DL27579

However, there are some limitations to this approach. The model's performance is highly dependent on the quality and diversity of the data used. If the dataset does not adequately represent all possible market conditions, the model's generalization ability may be compromised. Furthermore, training deep learning models on large datasets is computationally intensive, requiring significant resources and time. Future work should focus on optimizing the model's efficiency to reduce training times and improve scalability.

Looking forward, several avenues for enhancement can be explored. Semi-supervised learning techniques could be integrated to augment the model's predictive accuracy, leveraging the incorporation of labeled market events—such as news releases or market shocks. Additionally, exploring graph-based models, like graph neural networks (GNNs), could offer deeper insights into the complex relationships within LOB data. The framework could also be extended for real-time anomaly detection, enabling the identification of irregular trading patterns or potential market manipulation as they happen.

In conclusion, the framework developed in this study provides a robust, scalable, and interpretable solution for modeling HFT behavior using LOB data. Its potential for real-time market analysis and regulatory applications is substantial, laying a strong foundation for future advancements in automated market surveillance and trading strategy analysis.

References

[1] Aldridge, I. (2013). "High-frequency trading: a practical guide to algorithmic strategies and trading systems." John Wiley & Sons.
[2] Gould, M. D., et al. (2013). "Limit order books." Quantitative Finance, 13(11), 1709-1742.
[3] Passalis, N., et al. (2017). "Time-series classification using the FI-2010 dataset." ICASSP.
[4] Zhang, Z. et al. (2019). "DeepLOB: Deep convolutional neural networks for limit order books." IEEE Transactions on Signal Processing.
[5] Sirignano, J., and Cont, R. (2019). "Universal features of price formation in financial markets: perspectives from deep learning." Quantitative Finance, 19(9), 1449–1459.
[6] Hochreiter, S., and Schmidhuber, J. (1997). "Long short-term memory." Neural Computation, 9(8), 1735-1780.
[7] Dixon, M., Halperin, I., and Bilokon, P. (2020). "Machine learning in finance: from theory to practice." Springer.
[8] van der Maaten, L., and Hinton, G. (2008). "Visualizing data using t-SNE." Journal of Machine Learning Research, 9(Nov), 2579–2605.
[9] Kercheval, A. N., and Zhang, Y. (2015). "Modeling high-frequency limit order book dynamics with support vector machines." Quantitative Finance, 15(8), 1315-1329.
[10] Zhang, X., and Zohren, S. (2021). "LOBCAST: Limit Order Book Forecasting using Causal Transformers." arXiv preprint arXiv: 2108.10504.