ESP-JETA
ESP Journal of Engineering & Technology Advancements
ISSN: 2583-2646/ Volume 6 Issue 4 October-December 2026
Paper Id: JETA-V6I4P110/ Page No: 90-97

*Original Article*

# Explainable AI Frameworks for Optimized Trading Decisions in Highly Liquid Market Environments

**Chaitanya Rudra**
*New York University, NY*

Received Date: 06 November 2026 Revised Date: 11 November 2026 Accepted Date: 16 November 2026

***Abstract:*** *Highly liquid financial markets present a challenging environment for explainable artificial intelligence (XAI) because predictive accuracy, decision latency, market impact, nonstationary, and model transparency interact with trading performance. This review examines the latest peer-reviewed journal articles published from 2015 to 2024 on the themes of machine learning in trading, prediction of limit order books, market microstructure modelling and explainable AI techniques for optimal trading decisions. The surveyed literature shows progress on deep neural forecasting, order flow representation, ensemble learning and multi-horizon prediction. However, the extent to which explanation-aware execution improves trading outcomes remains insufficiently established. Post hoc attribution methods provide useful diagnostic information, but their explanations may lack temporal stability and causal interpretability. Moreover, these methods are not always evaluated against the latency, position-sizing, and execution constraints that determine practical trading feasibility. A major research gap is the absence of standardized evaluation frameworks that jointly assess explanation quality, predictive performance, transaction costs, trading risk, regime sensitivity, and computational latency. Explainable trading research thus requires a technique of assessment that focuses on the decisions to be made, and in which interpretability is not only discussed in hindsight about the model, but also becomes a part of model validation, risk management and economically relevant trade selection.*

***Keywords:*** *Algorithmic Trading, Explainable Artificial Intelligence, Limit Order Book, Market Microstructure, Machine Learning, Trading Optimization*

## I. INTRODUCTION

Highly liquid markets are quite a challenge to the AI-powered trading environment because these markets combine narrow bid–ask spreads, rapidly changing order-book states, intense competition, and short intervals between signal generation and execution. Machine learning models are therefore particularly important in financial time series forecasting and algorithmic systems of decision making, particularly when nonlinear interactions and high dimensional observations reduce the usefulness of simpler parametric specification. Although deep learning has been applied to financial forecasting in many scenarios, the translation of out-of-sample predictive performance into realized economic value is hampered by the significant variability in data preparation and validation methods across these applications, as well as in forecasting, measurement, and addressing trading costs. This separation is important in high liquidity markets where it may be eroded by spread and fees, adverse selection, market impact and model delay.

In the field of limit order book research, representations tailored to market microstructure demonstrate potential for short-horizon price prediction. DeepLOB exploits deep architectures to extract predictive structure from the high-frequency order book states by combining the idea of convolutional feature extraction with recurrent modelling [2]. However, the predictive capabilities of these architectures introduce important challenges for model interpretation. The signal that a trading engine gives can be for a direction or for an execution without being clearly determined as the result of the order flow, short-term imbalances, unusual liquidity conditions, correlated price levels or artefacts of the market state. In practice, two forecasts with the same level of accuracy can thus have vastly different risk characteristics.

Explainable artificial intelligence (XAI) provides techniques for explaining results of models back to their inputs, for providing local surrogate explanations, for determining the importance of features at the global level, and for designing intrinsically interpretable architectures. The overall XAI literature distinguishes the different levels of scope, different stages of explanation, dependence on the model, and the target audience for the explanation, and it highlights many unanswered questions about fidelity, robustness, causality, and the usefulness of the explanations for humans [3]. In liquid markets, these concerns become more acute. Order-book variables are correlated, which renders uniqueness difficult to establish; explanations based on past data can be problematic due to regime shifts; the applications occur at the millisecond level, which sets a constraint on computational power; and a technically accurate explanation may nevertheless provide limited information about whether a trade should be executed.

This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/2.0/)

Feature-attribution approaches have been shown to enhance the explainability of complex financial models [4] and have been used in prominent applications (e.g., for credit risk) despite the fact that much of the finance-specific explainability research has been focused outside the realm of latency-sensitive trading. The direct peer-reviewed literature integrating XAI with optimized trading in highly liquid markets remains limited. This review therefore also draws on adjacent journal literature concerning machine-learning-based trading and general XAI methodology. This review, then, critically investigates that intersection, and not mature explainable-trading evidence that lies adjacent to it. The main aim is to identify how existing predictive and explanatory approaches can be structured into a decision-centred framework, what the findings reported mean for optimized trading and what methodological gaps are not conducive to the reliable deployment of explanation-aware approaches.

## II. LITERATURE REVIEW

The study of machine-learning trading revealed that information of economic value can be derived from massive equity universes using nonlinear models. In their study of S&P 500 statistical arbitrage, Krauss, Do, and Huck compared the performance of deep neural networks, gradient-boosted trees, and random forests and found that there are economically significant differences in the architectures, the design of the ensembles, and the realistic back testing [5]. Later, Fischer and Krauss demonstrated that LSTM networks could take advantage of the temporal dependencies in order to predict the equity market and outperform several common baselines for the period they consider [6]. Sirignano and Cont also found the common price-formation structure across instruments using deep learning techniques on LOB data, indicating that certain patterns in market microstructure may generalize across different financial instruments [7]. Together, this cluster enables machine learning to be regarded as a valid input to decision making but also highlights the fact that the above does not guarantee that learned relations remain economically interpretable or stable once markets experience shifts in regime.

***Table 1: Summary of Key Findings***

| Ref | Focus | Key Findings |
|---|---|---|
| [5] | S&P 500 statistical arbitrage | Tree ensembles and neural models produced differentiated economic performance, with model design materially affecting trading results. |
| [6] | LSTM equity prediction | Temporal representation learning improved directional forecasting relative to examined conventional baselines. |
| [7] | Deep learning of price formation | Common microstructure patterns appeared across multiple instruments, supporting cross-asset representation learning. |
| [8] | Limit-order-book benchmark | Forecast quality depended strongly on horizon, class definition, and evaluation protocol. |
| [9] | Deep neural market classification | Nonlinear neural classification captured directional information but retained substantial black-box complexity. |
| [10] | Deep portfolio construction | Representation learning can connect prediction with portfolio choice, although interpretation of latent features remains difficult. |
| [11] | Machine learning in asset pricing | Nonlinear models captured interactions and predictor effects that traditional linear specifications missed. |
| [12] | Multi-horizon order-book forecasting | Horizon-aware deep architectures improved short-horizon modelling and highlighted hardware constraints relevant to deployment. |
| [13] | Deep order-flow imbalance | Order-flow representations contained alpha across horizons, strengthening the case for microstructure-centred predictors. |

Subsequent market-microstructure research has introduced increasingly specialized predictive architectures. Ntakaris et al. created the standardized limit-order-book benchmark for forecasting mid-price and emphasized the influence of forecasting horizon, class structure, and the evaluation methods on the reported results [8]. Dixon, Klabjan and Bang presented classification based market prediction using deep neural networks and provided evidence for the ability of deep neural networks to enhance the directional classification problem in financial environments when considered in a nonlinear fashion [9]. Heaton, Polson and Witte took deep-learning reasoning one step further in portfolio construction and showed that representation learning can support portfolio-allocation decisions in addition to predictive modelling [10]. Gu, Kelly and Xiu identified economically significant interactions between predictors in empirical asset pricing and demonstrated that the gain is significant when using nonlinear machine learning, and that the architecture and regularization matter significantly [11]. Limited by the need for hardware implementations, however, Zhang, Zohren and Roberts showed that multi-horizon limit-order-book forecasting architectures and computationally efficient implementations [12]. However, in more recent work, Kolm, Turiel and Westray demonstrated that order-flow imbalance features play a significant role in forecasting multiple-horizon alpha, confirming their predictive value across the microstructure time frame [13]. Table 1 summarizes representative studies

of prediction in liquid financial markets that provide the predictive basis for explanation-aware trading. The comparison shows that the literature has moved away from relatively rough classifiers of equity, towards more specific representations of order books and multi-horizon forecasting, and that explanation quality is not often a known outcome.

A second research cluster concerns explanation methodology. Adadi and Berrada [14] present a few other dimensions that are complementary to the ones mentioned above for categorizing XAI methods: scope of explanations, relationship between the explanation method and the underlying model used for predictions, and how explanations are produced. Local methods describe individual predictions, and global methods describe broader model behaviour. The techniques can be model-specific, or model-agnostic, with the latter having the ability to be applied to multiple model classes. Explanation fidelity and comprehensibility are evaluation issues and cannot be combined with these classification dimensions [14]. In [15] the authors show how the property of SHAP-style attribution of tree models can relate local predictions with a consistent global understanding of the features, which is of interest for ensemble-based trading strategies. Vilone and Longo highlight that there are no universal standards for the evaluation of explanations and that comprehensibility, correctness, usefulness and robustness are different properties and not synonyms [16]. Similarly, Linardatos, Papastefanopoulos and Kotsiantis have reported a wide variety of interpretability techniques that vary in their suitability for a given model and analytical goal [17].

The literature shows an asymmetry when combined. Predictive trading research provides more and more sophisticated architectures, but not always explanation stability testing, and XAI research provides more and more sophisticated explanation algorithms but not always trading utility in the presence of spread, latency, inventory or market impact constraints. Consequently, a feature-attribution result may be computationally valid without providing information that improves trading decisions or economic outcomes. The central unresolved issue is not only how to explain a forecast, but whether an explanation can enhance the acceptance of a trade, risk evaluation, model oversight, or decision on a particular model, in a market with ephemeral predictive relationships.

## III. METHODOLOGY

There are four methodological families in the evidence. The first are deep temporal and spatial architectures such as LSTM structures, convolutional order book models, and multi-horizon architectures [2], [6], [12]. These models work well under conditions where information about the prediction is spread over price levels, volumes, and event sequences, and where the representations are latent, so direct interpretation is complicated. The second family includes tree ensembles and nonlinear tabular learners [5], [11], for which global importance measures and decomposition using SHAP can be more easily calculated [15]. The third category are explicit representations of microstructure, like order-flow imbalance [7], [13]. These variables can also help to increase the economic interpretability of the model by pre-processing liquidity pressure before model fitting, thereby decreasing the reliance on post hoc explanation. The fourth category are model-agnostic or model-specific post-hoc XAI techniques, the main advantage of which is their ability to work with pre-existing predictive systems, but whose explanations can be affected by correlated inputs, reference distributions and/or perturbation assumptions [14, 16, 17].

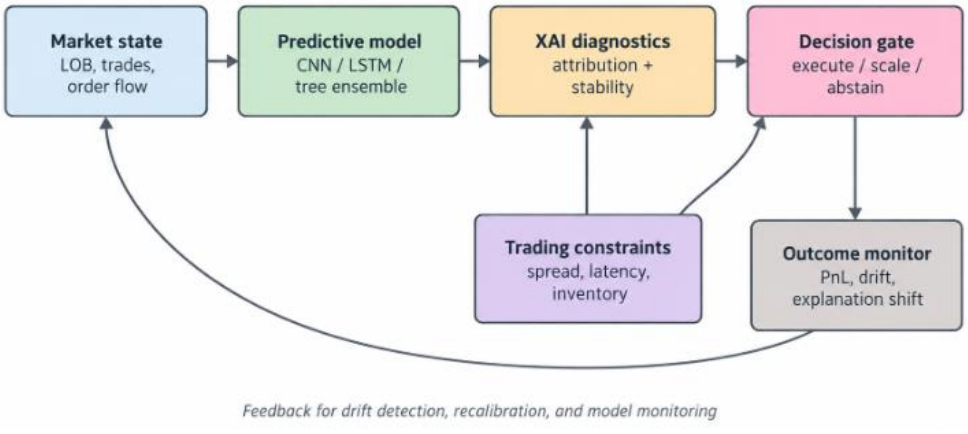



***Figure 1: Explanation-Aware Decision Architecture for Highly Liquid Markets***

An optimized XAI framework should integrate interpretability into the trading decision process rather than treating explanations solely as post-hoc diagnostics. The proposed architecture is illustrated in Figure 1: Microstructure observations are fed into a predictive model, prediction explanations are assessed against relevant liquidity conditions, market regimes, and attribution-stability criteria and the decision gate transforms the collected evidence into actions (execution, abstention or risk scaling). This structure is important because, when explanation is used, it is not only for reporting back on what happened but is connected to the market-state plausibility, as well as the stability of the attribution, influencing an actual decision.

***Table 2: Method Comparison***

| Ref | Method | Strengths | Limitations |
|---|---|---|---|
| [2] | CNN–LSTM order-book model | Captures spatial depth structure and temporal dependence jointly. | Internal representations are difficult to map directly to executable market mechanisms. |
| [5] | Random forests, boosted trees, DNNs | Supports model comparison and nonlinear decision boundaries. | Feature relevance can vary materially across model families and market periods. |
| [6] | LSTM sequence modelling | Represents persistent temporal dependencies without manual lag selection. | Attribution across long dependent sequences can be unstable and difficult to localize. |
| [11] | Nonlinear machine-learning asset pricing | Captures interactions and nonlinear predictor effects. | Economic interpretation can depend on correlated predictors and regularization choices. |
| [13] | Deep order-flow imbalance | Builds modelling around a microstructure quantity with direct economic meaning. | Predictive strength remains horizon- and regime-dependent. |
| [15] | SHAP-based tree explanation | Produces additive local attributions with global aggregation possibilities. | Correlated financial features complicate attribution interpretation. |
| [16] | XAI evaluation framework | Separates explanation correctness, usefulness, and comprehensibility. | Does not supply trading-specific latency or economic-utility criteria. |
| [17] | Comparative interpretability taxonomy | Clarifies suitability of multiple interpretation families. | General evaluation principles require adaptation to event-time market data. |

This figure connects order-book state, predictive modelling, explanation diagnostics, execution constraints and feedback monitoring in a single trading-decision architecture. The value of it is that it makes XAI a risk-sensitive decision gate that can alter or deny the actions of the model before they're exposed to the market. After summarizing the methodological characteristics reported in the various sources in [2] and in [5]–[17], a conceptual framework was created, as shown in Figure 1.

Another distinction that is often confused in financial XAI is methodological, which is clarified by the framework. The importance of a predictor, its explanation quality, and usefulness to decisions are distinct. A faithful attribution may identify the features influencing a model's predictions, but it does not establish whether the resulting trading decision is economically justified and will provide little evidence that the resulting action is robust to perturbations or can be done after costs. In contrast, the order-flow imbalance can have strong domain plausibility while still lacking a demonstrated causal interpretation [13]. Evaluation should therefore be a mixture of predictive metrics, economic outcomes, robustness of explanations and computational feasibility.

## IV. DISCUSSION

There are three general findings based on the evidence reported. First, performance in liquid markets is increasingly dependent on a representation that matches the market microstructure, not just on the number of parameters that are added to the model. DeepLOB demonstrates that using order book structure in conjunction with temporal modelling can extract short-term information [2] and Ntakaris et al.'s benchmarking study demonstrates that results are significantly different in terms of classes and forecast horizons [8]. Complementary evidence that some properties of price-formation are shared across instruments are provided by Sirignano and Cont [7], and Kolm, Turiel and Westray show how order-flow imbalance can be used to support multi-horizon alpha extraction [13]. The findings support a temporally local, market-mechanism-based explainability framework. If one does not map price-level coordinates to liquidity pressure, imbalance, spread configuration, and depth migration, the resulting explanation might be mathematically correct, but not economically informative.

Second, the reported predictive gains do not necessarily equal optimized trading results. Krauss, Do and Huck provide statistical-arbitrage evidence that the model family has an impact on economic outcomes [5] while Fischer and Krauss demonstrate useful prediction benefits from sequence modelling [6]. Gu, Kelly and Xiu also show that nonlinear machine learning can be used to derive economically relevant predictor relationships [11]. Even so, classification accuracy, mean squared error, and information coefficients can be used to rank the models differently from risk-adjusted returns, when turnover and costs are added to the decision problem. This distinction remains important in highly liquid markets, where transaction costs

may be small individually but can accumulate substantially under high trading frequency. In short-horizon trading strategies, the economic consequences of frequent trading may become more pronounced because queue position, adverse selection, transaction costs, and execution delays can materially affect realized performance. It is therefore important for an explainable framework to consider whether the information from the attribution aids in making the action better, rather than simply explaining the reasons for a given score.

Table 3 compares representative outcomes using the evaluation bases reported in the literature. In equity selection, order book forecasting and microstructure modelling, there exists good predictive evidence, while there is little direct evidence between the quality of explanations and trading performance.

***Table 3: Comparison of Reported Evaluation Approaches and Findings***

| Ref | System / Context | Metric / Basis of Comparison | Outcome |
|---|---|---|---|
| [5] | S&P 500 statistical arbitrage | Predictive and economic back test performance | Nonlinear model families produced materially different trading outcomes, supporting architecture-sensitive model selection. |
| [6] | Equity directional prediction | Classification and strategy performance | LSTM modelling captured useful temporal dependencies during the examined sample. |
| [2] | High-frequency limit order books | Multi-class mid-price prediction | CNN–LSTM representation improved extraction of spatial-temporal order-book information. |
| [7] | Multi-instrument price formation | Cross-instrument predictive generalization | Deep models identified recurring microstructure structure across instruments. |
| [8] | Limit-order-book benchmark | Horizon-specific classification | Performance changed strongly with horizon and experimental design. |
| [11] | Cross-sectional asset pricing | Out-of-sample prediction and economic relevance | Nonlinear interactions and flexible regularization improved predictive asset-pricing models. |
| [12] | Multi-horizon LOB forecasting | Forecast performance and computational implementation | Horizon-specific architecture and hardware jointly influenced deployable performance. |
| [13] | Order-flow-based alpha extraction | Predictability across horizons | Deep order-flow imbalance retained information across multiple short horizons. |
| [15] | Tree-based XAI | Local additive attribution and global aggregation | SHAP-style explanation provides structured feature decomposition but not trading-specific economic validation. |

Third, the studies in XAI reveal criteria which are not always measured in predictive trading research. The distinction of explanation types and the importance of the dependence of interpretability on the explanatory goal are emphasized by Adadi and Berrada [14]. One of the most relevant methods for tree-based systems is provided by the authors of the “SHAP values” paper by Lundberg et al., as SHAP-based methods decompose individual model predictions into feature-attribution values, which can also be aggregated to characterize broader patterns of feature importance [15]. In trading applications, these methods may provide trade-level diagnostics; for example, they can help determine whether a buy signal is driven by spread compression, imbalance, lagged returns, volatility state, or cross-sectional characteristics. However, there is a problem of identification because variables in the market are correlated. Order-book level variables tend to be co-moving, and lagged returns tend to be co-moving with order-book variables, which means that attribution to correlated variables may not necessarily be about an economic mechanism, but about a modelling and conditioning choice. Feature-attribution explanations should therefore be interpreted primarily as diagnostic evidence of model behaviour rather than as direct evidence of causal market mechanisms.

Another challenge is temporal stability. Not only is there no universal criterion for XAI quality as demonstrated by Vilone and Longo [16] but there is significant methodological variety among interpretability techniques, as reported by Linardatos, Papastefanopoulos and Kotsiantis [17]. For trading purposes, the stability of the explanation should be tested over periods of time, various market regimes, different levels of liquidity, and different retraining periods. An abrupt change in feature-attribution patterns despite relatively stable predictive outputs may indicate changes in the input distribution, model sensitivity, feature dependence, or parameter stability. However, attribution drift alone does not establish the presence of concept drift. This type of behaviour can be operationally useful even if the explanation isn't straightforward: attribution drift can be a variable in risk-control. An explanation framework should therefore not only document the existing feature

contributions, but also their levels of persistence, variance and divergence from historically proven regimes. The distribution of emphasis in the 17-paper review corpus is shown in Figure 2. Figure 2 illustrates the distribution of research themes within the 17 publications included in the selected evidence corpus. The limited representation of directly evaluated explanation-aware trading systems in this corpus highlights an evidence gap warranting further investigation. The bar chart in Figure 2 categorizes the reviewed journal corpus based on the main contribution to the liquid-market prediction, general XAI methodology, finance-specific XAI, or directly assessed explanation-aware trading. The imbalance needs to be noted because the methodological availability must not be confused with the empirical evidence that XAI is a trading-performance component. The number of articles from the 17 references cited and included in this review is reported in Figure 2 and was coded in this review.

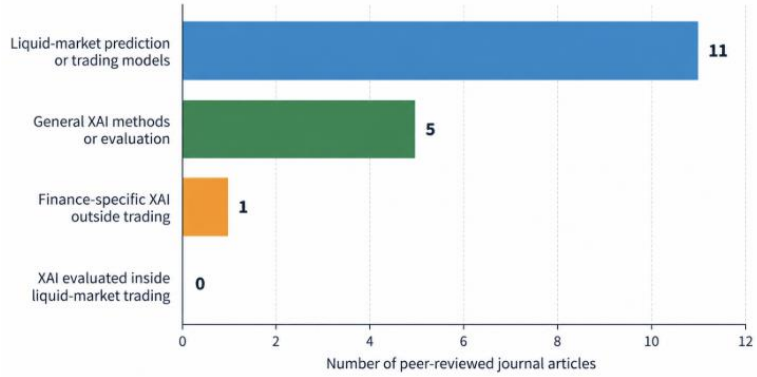



***Figure 2: Distribution of Evidence Across the Reviewed Literature***

Another underdeveloped link is computational feasibility. However, in multi-horizon forecasting, hardware acceleration is an important factor to address, as inference delay is substantial, which can offset the benefit of a high-frequency predictor, as in explicit consideration of hardware acceleration [12] by Zhang, Zohren and Roberts. Various post-hoc techniques are evaluated without having to take such constraints into account. In a very liquid market, therefore, explanation latency ought to be measured as a system property. Computationally lightweight attribution methods may be suitable for real-time trade-gating mechanisms when their measured execution latency satisfies the system's operational constraints. This separation prevents an unwarranted assumption that all explanatory operations have to happen at the same frequency as the orders are put forth.

Figure 3 maps representative predictive-model and explanation-method families according to real-time feasibility and applicability to execution decisions. One can see from the figure that the families of methods vary not only in their interpretability but also in their estimated real-time feasibility and direct applicability to execution decisions, so that no one approach to explanation will meet the needs of all trading functions. Figure 3 shows the visualization of the representative model and explanation families based on the qualitative reviewers' coding of the execution relevance and computational feasibility of the models. Bubble sizes represent qualitative assessments of the interpretability potential of each method family, based on the methodological characteristics identified in the reviewed literature. These assessments are illustrative rather than empirically calibrated measurements. Its importance is that it illustrates the advantages of hybrid architectures: a lightweight set of attributions can be used for real time control, while a more powerful set of explanations can be used for monitoring and validation. Figure 3 is based on ordinal scores based on methodological characteristics described in [2], [5]–[17] and not on reported numerical benchmark values.

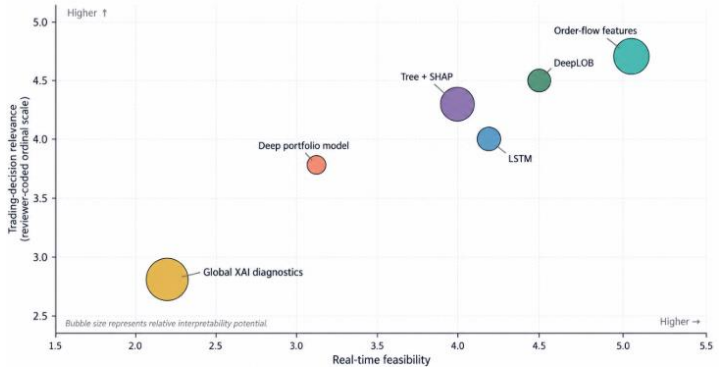



***Figure 3: Method Families Mapped to Real-Time Feasibility and Trading-Decision Relevance***

Joint evidence also suggests the need to balance model complexity and interpretability. Explaining the order book dynamics is difficult for deep architectures [2, 12] due to the distribution of the interactions between the features, the time and space in which they occur. SHAP-style explanations can also be applied to complex predictive models, although computational cost, correlated inputs, and the choice of explanation method may complicate their practical use. Furthermore, the availability of such explanations does not establish an improvement in economic performance [15]. The objective should not be transparency for its own sake, but rather an appropriate level of interpretability that supports a clearly defined trading, risk-management, or model-governance objective. Explanation-derived diagnostics should be evaluated for fidelity, temporal stability, computational latency, and economic relevance. Their usefulness for trading should then be assessed separately through their measurable effects on execution decisions, risk exposure, and out-of-sample performance.

This criterion suggests a number of explicit system-level metrics that were not included in most of the reviewed trading studies. Explanation stability can be assessed using either rank correlation or divergence in the feature-attribution vectors of the adjacent event windows. Inference latency of the model can be reported along with explanation latency. This can be done by comparing a baseline strategy to an explanation-gated strategy where trades are either rejected or scaled back when the attribution pattern is in violation of validated conditions. The behaviour of the explanation can be broken down into the various "regimes" of volatility, spread, depth and order flow, in order to evaluate regime sensitivity. Most crucially, such tests need to be performed in a strictly out-of-sample environment, with transaction costs and temporal separation, or else post-hoc interpretation, data leakage and retuning may result in over-optimistic backtest performance.

Empirical research, then, is the key missing component, rather than conceptual. A number of candidate methods already exist that are high-performing, domain-informed, equipped with attribution tools, and supported by explanation-evaluation criteria [2], [13]–[17]. Rigorous evidence that explanation-aware decision gates achieve better risk-adjusted trading performance, calibration, failure detection of tails or tail-risk control than equally optimized black-box baselines is still missing. Until this is known, assertions of XAI's capacity to optimize trading decisions should be more limited than assertions of XAI's capacity to increase transparency.

## V. FUTURE DIRECTIONS

Explainability should be considered as a trading goal in itself, rather than as a visualization layer. One of the key requirements is the development of benchmark protocols which are able to harmonise predictive performance with net return, turnover, maximum drawdown, execution cost, inference latency, explanation latency and attribution stability. Limit-order-book benchmarks help to provide a useful foundation for horizon-specific prediction [8] and multi-horizon modelling also demonstrates that computational implementation materially affects deployability [12]. An explanation-aware benchmark should preserve chronological separation between training, validation, and test datasets.

The second is related to the stability of the explanation when the inputs to the microstructure are correlated. While structured local feature attributions [15] are available for SHAP methods, future applications would benefit from a clear distinction between feature assumptions of conditional and interventional SHAP. Stability tests should measure the amount of change an explanation undergoes when subjected to economically insignificant perturbations, retraining, or transitioning into a different market regime. Highly unstable explanations may lead to abstention or to a smaller position size or to changing the model, but not just to being shown as diagnostic charts.

The third priority is hybrid interpretability. Quantities such as order-flow imbalance can be intrinsically meaningful intermediate representations [13] and deep models can still have nonlinear forecasting capacity. Post-hoc attribution is then capable of describing model dependence on interpretable intermediate features that are economically relevant instead of thousands of less distinguishable raw order-book coordinates. This architecture may improve the computational efficiency and practical usefulness of explanations, although these benefits require empirical validation.

Lastly, there should be a separation between human and automated consumers of explanations. Traders and risk-management professionals may want to have all kinds of diagnostics at the regime level, while automated execution engines would need machine-readable constraints that can fit into a latency budget. Future studies should confirm if explanation-derived vetoes, confidence scaling, regime filtering, and/or execution modifications will provide statistically and economically significant improvements over black-box baselines. This kind of assessment would turn explainability into the verifiable component of a trading design implementation.

## VI. CONCLUSION

Explainable AI offers potentially valuable mechanisms for improving the transparency and governance of machine-learning-based trading systems in highly liquid markets, although direct evidence of improvements in execution quality and risk-adjusted performance remains limited. Powerful neural, ensemble, and microstructure-based models can be used to extract predictive information from equity characteristics, temporal sequences, states of the limit-order book, and order flow.

Another advantage of XAI methods is that they can uncover feature contributions, model dependence, and model changes in decision logic. Evidence to support the claim that such explanations enhance execution quality or affect risk-adjusted trading results is limited.

The central conceptual direction identified by this review is a transition from explaining predictions retrospectively to incorporating explanation-derived diagnostics into trading decisions. An effective framework should determine the stability, economic feasibility, computational feasibility and consistency of model reasoning with the liquidity condition before a model action is put into the market. These properties cannot be established by prediction accuracy alone, nor can profitability be inferred from interpretability alone.

Some weaknesses that need improvement are the lack of robustness of explanations when the market has correlated characteristics, the need to evaluate the explanations when regime changes, computational delay, causal overinterpretation, and the lack of standardized economic benchmarks. Future progress is likely to depend on architectures combining microstructure-informed representations, high-capacity prediction, lightweight real-time attribution, and explicit execution constraints. The practical value of explanation-aware trading should therefore be evaluated through measurable outcomes, including trade-gating effectiveness, risk-adjusted position sizing, drift-detection performance, computational latency, and out-of-sample economic results.

## VII. REFERENCES


[1] 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.

[2] Zhang, Z., Zohren, S., & Roberts, S. (2019). DeepLOB: Deep convolutional neural networks for limit order books. *IEEE Transactions on Signal Processing, 67*(11), 3001–3012.

[3] Barredo Arrieta, A., Díaz-Rodríguez, 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.

[4] Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics, 57*(1), 203–216.

[5] Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research, 259*(2), 689–702.

[6] 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.

[7] Sirignano, J., & Cont, R. (2019). Universal features of price formation in financial markets: Perspectives from deep learning. *Quantitative Finance, 19*(9), 1449–1459.

[8] Ntakaris, A., Magris, M., Kanniainen, J., Gabbouj, M., & Iosifidis, A. (2018). Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods. *Journal of Forecasting, 37*(8), 852–866.

[9] Dixon, M., Klabjan, D., & Bang, J. H. (2017). Classification-based financial markets prediction using deep neural networks. *Algorithmic Finance, 6*(3–4), 67–77.

[10] Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: Deep portfolios. *Applied Stochastic Models in Business and Industry, 33*(1), 3–12.

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

[12] Zhang, Z., Zohren, S., & Roberts, S. (2021). Multi-horizon forecasting for limit order books: Novel deep learning approaches and hardware acceleration using intelligent processing units. *IEEE Journal of Selected Topics in Signal Processing, 15*(1), 215–224.

[13] Kolm, P. N., Turiel, J., & Westray, N. (2023). Deep order flow imbalance: Extracting alpha at multiple horizons from the limit order book. *Mathematical Finance, 33*(4), 1044–1081.

[14] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160.

[15] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67.

[16] Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion, 76*, 89–106.

[17] Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23*(1), 18.