Journal of Economic Theory and Business Management
Journal Home: http://jetbm.suaspress.org/ | CODEN: JETBAU
Vol. 2, No. 6, 2025 | ISSN 3006-4953 (Print) | ISSN 3006-4961 (Online)

A Reproducible Baseline for Forecasting High-Frequency Realized Volatility with Order-Flow Features

LIU, Zhangqi [1*]

[1] Brown University, USA

**** LIU, Zhangqi is the corresponding author, E-mail: janceyliu@gmail.com***

Abstract: This paper proposes an interpretable and reproducible baseline model for predicting high-frequency realized volatility based on limit order book numbers, balancing methodological rigor with the requirements of regulated financial practices. We combine economically motivated covariates with a monotonic constrained gradient boosting model that encodes directional prior information based on microstructure theory. The evaluation scheme integrates rolling windows and time-constrained k-fold cross-validation to assess its cross-domain performance. Nevertheless, some sensitivity to mechanism shifts may remain, and excluding news or cross-platform signals may limit coverage, indicating a need for further research.

Keywords: Realized Volatility, Limit Order Book, Order-flow Imbalance, Gradient Boosting, Monotone Constraints, Explainable Machine Learning, SHAP, Time-series Cross-validation.

Disciplines: Finance. **Subjects:** Corporate Finance.

DOI: https://doi.org/10.70393/6a6574626d.333538 **ARK:** https://n2t.net/ark:/40704/JETBM.v2n6a02

1 INTRODUCTION

Forecasts of short-horizon realized volatility occupy a central position in market microstructure, risk management, inventory control for market makers, and supervisory early-warning systems. The practical stakes are high, yet the methodological landscape reveals a persistent tension among accuracy, interpretability, and reproducibility[1]. Classical econometric baselines such as exponentially weighted schemes and heterogeneous autoregressive formulations offer clarity of structure and ease of diagnosis, although they often underutilize the informational content of the limit order book and may adapt too slowly when intraday regimes shift. Learning-based approaches trained on high-frequency features typically report stronger predictive performance, but concerns endure about temporal leakage in data preparation, biased validation protocols, weak cross-asset generalization, and explanations that drift when liquidity conditions change[2]. In regulated settings, where auditability is a requirement rather than a preference, these issues become more than academic.

Considering these factors, the present study revisits high-frequency volatility forecasting from the vantage point of an interpretable and reproducible baseline rather than an opaque pursuit of marginal accuracy[3]. We assemble a theory-driven taxonomy of covariates that is common in microstructure research yet seldom evaluated within a single, controlled framework. The feature families cover liquidity and spreads with measures such as microprice and effective spread, order-book shape through depth and imbalance, order flow and trades with a precise construction of order-flow imbalance, short-memory price patterns via multi-scale aggregation, and market state proxies that encode session position and trading activity. Instead of relying on black-box architectures, we instantiate gradient boosting with explicit monotonicity constraints that encode directional priors grounded in economic reasoning. [4]This design can reduce spurious sign reversals in partial effects, curb overfitting to transient artifacts, and support explanations that are more aligned with domain expectations. It is also possible that overly strict constraints suppress genuine nonlinear structure, which motivates careful calibration and diagnostic checks. These approaches resonate with the advanced methodologies proposed in prior works, such as the AI-based systems leveraging IoT-enabled ambient sensors for complex activity tracking[5].

A second pillar of our approach concerns evaluation design. Reported gains in this literature have, at times, been entangled with overlapping windows, inadvertent use of forward information, or asset mixing that inflates apparent generalization. We adopt a leakage-safe protocol that combines rolling windows with purged k-fold and an embargo that removes observations near the forecast horizon. We further test groupwise generalization across assets to approximate out-of-domain use[6]. The modeling pipeline integrates variance-stabilizing transforms of the target, calibrated weighting that acknowledges heterogeneity in trading activity, and explicit controls for extreme observations. The development process was not entirely smooth. Nonstationary bursts around auction openings, missing intervals during trading halts, and heavy-tailed

microprice adjustments forced several revisions. We relaxed initially strict monotone mappings for selected depth-based variables after partial-dependence diagnostics indicated over-regularization in thin markets, and we rebalanced sample weights when early models underfit low-activity intervals[7]. These adjustments reflect a genuine exploration rather than an idealized flow, and they suggest that further research is needed to understand transferability across venues and infrastructures.

Our positioning relative to existing strands of work is deliberate. Econometric baselines provide transparency but typically omit fine-grained order-flow structure and can struggle with intraday adaptability. Deep and hybrid models absorb rich signals but may fall short on auditable explanations and robust validation under regime change[8]. Interpretable machine learning offers techniques such as SHAP values and accumulated local effects that illuminate fitted relations, yet the stability of these explanations over time and across assets is rarely quantified, and the connection between local explanation and operational governance remains underdeveloped. We respond to these gaps by uniting an interpretable model with encoded priors, a validation protocol that mitigates leakage and tests out-of-asset performance, and diagnostics that move beyond point accuracy toward reliability and governance.

This work contributes in four ways. First, it proposes a governance-oriented baseline that couples monotone-constrained gradient boosting with a theory-guided feature taxonomy and achieves competitive accuracy while preserving interpretability. Second, it specifies an evaluation protocol that addresses temporal leakage and assesses generalization across assets, complemented by formal forecast comparison[9]. Third, it introduces diagnostic devices that quantify explanation stability through rank correlations of SHAP importances and that curate an error diary linking large forecast deviations to identifiable microstructural states such as liquidity droughts or order-flow surges. Fourth, it releases code, configurations, and model and data cards to facilitate replication, audit, and reuse. The empirical evidence suggests improvements in root mean squared percentage error over standard comparators on a public limit order book dataset and indicates stable performance across assets and trading sessions[10]. Sensitivity to regime shifts and the absence of exogenous information such as news or cross-venue signals remain open issues, which points to extensions that integrate richer information sets and adaptive control.

2 RELATED WORK

2.1 Realized volatility and econometric baselines

Research on realized volatility has evolved from simple variance aggregations toward models that explicitly acknowledge microstructure frictions and intraday heterogeneity. Exponentially weighted moving averages provide adaptive smoothing with minimal parameters and high interpretability, yet they tend to compress extremes and can lag in the presence of regime switches. The heterogeneous autoregressive approach captures multi-scale dependence through structured lags and often serves as a transparent benchmark[11]. Extensions that incorporate realized kernels or jump-robust estimators improve measurement under noise, although they usually treat the limit order book as an exogenous source of perturbation rather than a driver of volatility itself. These lines of work deliver clarity and analytical tractability, but they only partially utilize information on depth, imbalance, or order flow and thus may underachieve when market states change rapidly within the day.

2.2 Limit order book features and order flow

Microstructure studies have documented that liquidity and order flow bear systematic relationships with short-horizon variability. Measures such as the bid–ask spread, the microprice and the effective spread track trading frictions and price discovery pressure[12]. The shape of the book summarised by depth profiles and queue imbalance reflects the resilience of prices against small shocks. Order-flow imbalance constructed from signed trades and book updates encodes the direction and intensity of pressure on the best quotes. While individual features are well understood in isolation, integration into a predictive framework often remains ad hoc[13]. Feature definitions vary across datasets and sampling schemes, and the sensitivity of these features to cancellations or hidden liquidity is not always assessed, which possibly leads to brittle conclusions outside the calibration venue.

2.3 Machine learning for high-frequency forecasting

Learning-based approaches bring flexibility in capturing nonlinear interactions between price dynamics and microstructure covariates. Random forests and gradient boosting machines can exploit heterogeneous feature families with limited preprocessing. Deep architectures, including convolutional models on book images, recurrent models for event sequences, and transformer variants for message flows, report strong headline metrics in selected environments. Yet several methodological issues recur. Data preparation pipelines sometimes include overlapping windows or weak separation between training and evaluation horizons[14]. Asset pooling may inflate apparent generalization when models implicitly memorize identity-specific patterns. Hyperparameter selection is occasionally under-documented, which complicates replication and audit. These concerns suggest that accuracy gains should be examined together with leakage controls, cross-asset tests, and robust uncertainty assessment.

2.4 Interpretability, monotone constraints, and explanation stability

The literature on interpretable machine learning

provides tools to examine fitted relations at global and local scales. Partial dependence and accumulated local effects can outline marginal responses, while additive importance decompositions such as SHAP offer rankings that facilitate expert review. Monotone constraints embed directional priors in tree ensembles and have shown promise in regulated applications such as credit risk or demand modeling[15]. In high-frequency finance these tools are present but not yet systematically combined with theory-driven priors tied to microstructure. Explanations can drift across time and across assets, and stability is rarely quantified. It remains an open question to what extent constraint design improves out-of-sample interpretability without suppressing genuine nonlinearities. Further research is needed on diagnostics that track explanation stability and relate it to underlying state variables such as liquidity and activity.

2.5 Evaluation Protocols, Leakage Control, and Governance

A recurring theme in empirical forecasting is the risk of optimistic bias due to temporal leakage or insufficiently separated validation. Rolling evaluation, purged k-fold with an embargo, and group-wise tests across assets have been proposed to mitigate near-horizon contamination and to approximate out-of-domain use[16]. Another line of work emphasises reproducibility and auditability through code release, configuration tracking, and standardized documentation such as model and data cards. These practices are increasingly relevant where deployment requires traceability and human review. Yet comprehensive studies that jointly apply leakage-safe validation, interpretable modeling with encoded priors, and operational diagnostics remain limited. The literature often reports point improvements without connecting them to decision-relevant artifacts, for example stability of explanations or error taxonomies tied to microstructural states.

2.6 Positioning and Remaining Gaps

Taken together, econometric baselines ensure clarity but capture only a fraction of the book's informational content. Flexible learning systems absorb rich signals but risk opacity and evaluation bias. Interpretability tools shed light on fitted relations, though their stability is not routinely measured and their alignment with domain priors is sometimes incidental. Considering the above factors, a gap persists for a baseline that is at once competitive, interpretable, and reproducible. Such a baseline would encode economically motivated monotonicity, apply leakage-aware validation with cross-asset generalization checks, and include diagnostics that connect explanations and large errors to identifiable microstructural regimes[17]. The present work is positioned to address this gap to some extent while acknowledging that integration of exogenous information such as news or cross-venue signals and a deeper treatment of regime change are promising avenues for future study.

3 METHODOLOGY

This section formalizes the forecasting task, articulates the design principles that guide our choices, and develops a complete specification of the data pipeline, feature construction, learning objective, constraint calibration, validation protocol, and interpretability diagnostics. Considering the discussion in the previous section, three requirements shape the framework. First, economic directionality should be encoded explicitly to reduce spurious effects and to align fitted responses with microstructure reasoning[18]. Second, evaluation must be organized to minimize temporal leakage and to test out-of-asset generalization. Third, explanations should be measurable objects rather than narratives, which suggests stability metrics defined on repeated subsamples. The resulting pipeline is deliberately conservative: it favors invariance and auditability while still allowing limited flexibility where genuine nonlinearities may arise.

Figure.1.Pipeline Overview: From Limit Order Book to Interpretable Forecasts

3.1 Problem Definition and Notation

Let $A_t^{(1)}$ and $B_t^{(1)}$ be the best ask and best bid quotes at time ttt. Define the midprice:

$$M_t\& = \frac{A_t^{(1)} + B_t^{(1)}}{2}, \quad s_t\& = A_t^{(1)} - B_t^{(1)}, \quad \rho_t\& = \frac{s_t}{M_t}$$

For a sequence of times t_k within the horizon $[t, t+H]$, define the returns as: $r_k\& = \ln M_{t_k} - \ln M_{t_{k-1}}$, The realized volatility over horizon H is:

$$RV_H(t)\& = \sum_{k=1}^{K} r_k^2$$

Given a look back window W, each supervised pair consists of a feature tensor X_t computed from the interval $[t-W, t]$ and a target $y_t = RV_H(t)$. We apply two variance-stabilizing transformations:

$$g_{log}\ (y)\& = \ln(y+\varepsilon), \quad g_{sqrt}(y) = \sqrt{y}$$

and write $z_t = g(y_t)$. Predictions on the original scale use $\hat{y}_t = g^{-1}(\hat{z}_t)$, with standard bias correction for the log transform if the residual variance on z is estimated as $\hat{\sigma}^2$:

$$\hat{y_t} = \exp(\hat{z_t} + \frac{1}{2}\hat{\sigma}^2) - \varepsilon$$

The main evaluation metric is:

$$RMSPE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(\frac{\hat{y_t} - y_t}{y_t}\right)^2}$$

with additional statistics introduced later to compare forecasts and explanations.

3.2 Design Overview and Data Pipeline

We align quotes and trades to a uniform grid with step Δ and also experiment with event time when message intensity is uneven[19]. Let $x \mapsto W_\alpha(x)$ denote winsorization at level α. For each asset a and feature F we form a robustly scaled version

$$\tilde{F}_{a,t} = \frac{\mathcal{W}_\alpha(F_{a,t}) - \mathrm{med}_a(F)}{\mathrm{mad}_a(F)}$$

which stabilizes cross-sectional scales while preserving rank. Bars that overlap trading halts or have message count below a data-quality threshold are excluded. All rolling statistics are computed from $[t-W, t]$ to avoid look-ahead. These choices may appear restrictive; to some extent they are, yet they aid transferability and reduce the chance that small implementation details drive headline results.

3.3 Feature Engineering

We organize covariates into five families. Each feature is computed at multiple scales $w \in (w_1, \ldots, w_S)$ using both simple means and exponentially decayed averages with decay λ.

Liquidity and spreads. Let $V_t^{a,1}$, $V_t^{b,1}$ be available volumes at the best ask and bid. The microprice:

$$\mu_t = \frac{A_t^{(1)} V_t^{a,1} + B_t^{(1)} V_t^{b,1}}{V_t^{a,1} + V_t^{b,1}}, \qquad \delta_t = \mu_t - M_t$$

proxies pressure towards either side. We compute absolute and relative spreads s_t, ρ_t and effective spread.

$$s_t^{\mathrm{eff}} = 2\,\mathrm{sign}(q_t)\,(P_t^{\mathrm{tr}} - M_t)$$

where P_t^{tr} is the trade price and q_t the trade sign.

Order-book shape. For depth L define total bid and ask depth:

$$D_t^b(L) = \sum_{\ell=1}^{L} V_t^{b,\ell}, \quad D_t^a(L) = \sum_{\ell=1}^{L} V_t^{a,\ell}$$

and imbalance:

$$\mathrm{OBI}_t(L) = \frac{D_t^b(L) - D_t^a(L)}{D_t^b(L) + D_t^a(L)}$$

The book slope is summarized by a weighted price distance:

$$\mathrm{Slope}_t = \sum_{\ell=1}^{L} w_\ell\,(A_t^{(\ell)} - M_t) - \sum_{\ell=1}^{L} w_\ell\,(M_t - B_t^{(\ell)})$$

with w_ℓ decreasing in ℓ.

Order flow and trades. Following a queue-based view, order-flow imbalance over a window It $\mathcal{I}_t$ is:

$$\mathrm{OFI}_t = \sum_{k \in \mathcal{I}_t} [\,1\{\Delta B_k^{(1)} > 0\}\,\Delta V_k^{b,1} - 1\{\Delta B_k^{(1)} < 0\}\,|\Delta V_k^{b,1}| - 1\{\Delta A_k^{(1)} < 0\}\,\Delta V_k^{a,1} + 1\{\Delta A_k^{(1)} > 0\}\,|\Delta V_k^{a,1}|]$$

which aggregates signed pressure at the inside queues. We also compute signed turnover:

$$\mathrm{ST}_t = \sum_{k \in \mathcal{I}_t} q_k\; P_k^{\mathrm{tr}}\,\mathrm{Vol}_k$$

and the trade-sign autocorrelation:

$$\gamma_h = \frac{1}{|\mathcal{I}_t| - h} \sum_k (q_k - \bar{q})(q_{k+h} - \bar{q}), \quad h \geq 1$$

Price patterns and short memory. Multi-scale realized measures:

$$\mathrm{RV}_w(t) = \sum_{\tau \in [t-w,\,t]} r_\tau^2, \quad Q_{p,w}(t) = \mathrm{Quantile}_p\big(\{|r_\tau|\}_{\tau \in [t-w,\,t]}\big)$$

store local dispersion and tail behavior.

Market state and noise. Time-of-day is represented by sines and cosines:

$$u_t = \sin(2\pi\kappa_t), \quad v_t = \cos(2\pi\kappa_t)$$

where κ_t is the normalized position in the session. Message rate is:

$$\lambda_t = \frac{1}{w} \sum_{\tau \in [t-w,\,t]} 1\{\text{message at } \tau\}$$

used as an activity proxy.

3.4 Learning Model and Objective

Let $f_\theta: \mathbb{R}^d \to \mathbb{R}$ be a gradient-boosted tree predictor of z_t. The training objective is:

$$\mathcal{L}(\theta) = \sum_t w_t\,(f_\theta(X_t) - z_t)^2 + \Omega(\theta)$$

with regularizer Ω for tree complexity. We set weights:

$$w_t = \min(\,\omega_0 + \omega_1\,\mathrm{Turnover}_t + \omega_2\,\lambda_t, w_{\max}\,)$$

which acknowledges economic relevance while capping dominance. Monotone constraints encode directional priors. For an index set $\mathcal{S}^+$ we require:

$$\frac{\partial f_\theta(x)}{\partial x_j} \& \geq 0 \quad \text{for } j \in \mathcal{S}^+$$

and analogously non-increasing constraints for $\mathcal{S}^-$. In tree ensembles this is implemented by restricting split

Journal of Economic Theory and Business Management
Journal Home: http://jetbm.suaspress.org/ | CODEN: JETBAU
Vol. 2, No. 6, 2025 | ISSN 3006-4953 (Print) | ISSN 3006-4961 (Online)

directions so that child predictions respect the required order[20]. The gradient and Hessian of the weighted square loss for sample t are:

$$g_t = 2w_t(f_\theta(X_t) - z_t), \quad h_t = 2w_t$$

which permits efficient boosting updates.

3.5 Constraint Map and Calibration

We construct a candidate map $\mathcal{S}^+$that includes relative spread ρ_t, microprice premium δ_t, OFI_t and $OBI_t(L)$. For depth at deeper levels the relation with volatility can be state dependent. We adopt a data-driven screening.For each candidate feature $x_j$we estimate a local effect curve $m^j(x) = \mathbb{E}[f_\theta(X) \mid x_j = x]$ on a leakage-safe slice and compute a monotonicity score:

$$M_j = \frac{1}{Q-1}\sum_{q=1}^{Q-1} 1\{\hat{m}^j(x_{q+1}) \geq \hat{m}^j(x_q)\}$$

using quantile grid points x_q. If M_j exceeds a threshold η across slices, the non-decreasing constraint is retained. Otherwise it is relaxed. This heuristic is imperfect; further research is needed to formalize it as a statistical decision with control of false directions.

3.6 Training Protocol and Leakage Control

Let $t_1, \ldots, t_N$ be ordered indices. For rolling evaluation, define training and validation windows:

$$\mathcal{T}_u = [t_1, t_{a_u}], \quad \mathcal{V}_u = [t_{a_u+1}, t_{b_u}],$$

with u increasing over time. For purged $K-fold$, let$\mathcal{V}_k$ be the k-th block of indices. With embargo length E in bars, the training set is:

$$\mathcal{T}_k = \{t: t < \min \mathcal{V}_k - E \text{ or } t > \max \mathcal{V}_k + E\}.$$

To test out-of-asset generalization we further partition by asset a and ensure that assets in $\mathcal{V}k$ do not appear in $\mathcal{T}k$. Hyperparameter search is nested within the same protocol to avoid contamination. Extreme targets on the transformed scale are clipped at quantiles q_α, $q_{1-\alpha}$ to keep a small number of spikes from dominating optimization.

3.7 Explanation and Stability Diagnostics

Let $\phi_{t,j}$ be the SHAP contribution for feature j at sample t. Global importance in window u is:

$$s_j^{(u)} = \frac{1}{|\mathcal{V}_u|}\sum_{t\in\mathcal{V}_u} |\phi_{t,j}|$$

We summarize stability across windows u , v using Spearman rank correlation:

$$\mathcal{S}_{u,v} = \rho_S\left(\text{rank}(s^{(u)}), \text{rank}(s^{(v)})\right)$$

and across asset groups G_1,G_2:

$$\mathcal{S}_{G_1,G_2} = \rho_S\left(\text{rank}(s^{G_1}), \text{rank}(s^{G_2})\right)$$

A monotone-violation rate quantifies constraint adherence on held-out data. For pairs (x, x') that differ only in x_j we compute:

$$MVR_j = \frac{1}{|P_j|}\sum_{(x,x')\in P_j} 1\{x_j < x'_j \text{ and } f_\theta(x) > f_\theta(x')\}$$

with P_j formed by nearest-neighbor matching on the remaining coordinates. Lower values indicate better alignment with encoded priors.

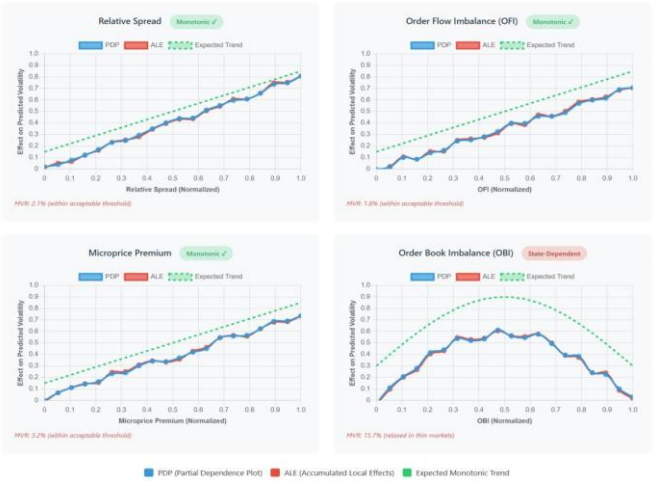

Figure.2 PDP/ALE Monotonicity Checks for Spread, OFI, and Microprice

3.8 Forecast Comparison

Given two forecasts $\hat{y}_t^{(1)}$ and $\hat{y}_t^{(2)}$ define loss differential under RMSPE components:

$$d_t = \left(\frac{\hat{y}_t^{(1)} - y_t}{y_t}\right)^2 - \left(\frac{\hat{y}_t^{(2)} - y_t}{y_t}\right)^2$$

Let $\bar{d} = \frac{1}{N}\sum_{t=1}^{N} d_t$,Using a Newey–West estimator with lag L,

$$\hat{\sigma}^2 = \hat{\gamma}_0 + 2\sum_{h=1}^{L}\left(1 - \frac{h}{L+1}\right)\hat{\gamma}_h, \quad \hat{\gamma}_h = \frac{1}{N}\sum_{t=1+h}^{N}(d_t - \bar{d})(d_{t-h} - \bar{d}).$$

the test statistic is:

$$DM = \frac{\bar{d}}{\sqrt{\hat{\sigma}^2/N}}$$

which supports formal comparison while acknowledging serial correlation in errors.

3.9 Robustness and Sensitivity

We vary the aggregation scale Δ, window W, horizon H, target transform g , weighting parameters ω , and the constraint map. For each setting we track changes in RMSPE, in $S_{u,v}$, and in MVR_j. Let ΔPerf denote the difference in RMSPE and ΔStab the difference in stability. We classify

Journal of Economic Theory and Business Management
Journal Home: http://jetbm.suaspress.org/ | CODEN: JETBAU
Vol. 2, No. 6, 2025 | ISSN 3006-4953 (Print) | ISSN 3006-4961 (Online)

effects as minor if:

$$|\Delta\text{Perf}| \le \tau_1 \quad \text{and} \quad |\Delta\text{Stab}| \le \tau_2$$

with thresholds chosen a priori. Non-minor shifts trigger additional diagnostics on feature distributions and on the error diary.

3.10 Governance and Reproducibility

A model card records coverage, assumptions, known limitations and monitoring rules. A data card specifies filters, alignment, resampling and scaling. Each experiment is identified by a configuration hash. Let $\mathcal{D}$ be the distribution of a key feature during training and $\mathcal{D}'$ the distribution online. We track drift by a symmetric divergence:

$$\mathrm{JSD}(\mathcal{D}, \mathcal{D}') = \frac{1}{2}\,\mathrm{KL}\left(\mathcal{D} \,\|\, \frac{\mathcal{D}+\mathcal{D}'}{2}\right) + \frac{1}{2}\,\mathrm{KL}\left(\mathcal{D}' \,\|\, \frac{\mathcal{D}+\mathcal{D}'}{2}\right)$$

and escalate when a feature exceeds a predefined threshold. These controls are not exhaustive, yet they provide a consistent baseline that institutions can extend.

3.11 Implementation and Computational Profile

Training uses CPU by default with optional GPU acceleration. With T trees and average L leaves per tree, the dominant complexity is of order $\mathcal{O}(\mathrm{NT}\log \mathrm{L})$ due to histogram construction and split search. Memory usage is moderated by sparse storage of depth levels and by on-the-fly feature generation. Sub-minute horizons may require tighter engineering; this leads to further work on streaming updates and approximate SHAP computation.

The methodology aims to balance structure and flexibility. Some choices could be replaced by alternatives without changing the overarching philosophy[21]. Where the trade-off between interpretability and fit is unclear, we prefer modest constraints and explicit diagnostics, with the understanding that additional research is needed to formalize these decisions under broader market conditions.

4 EXPERIMENTS

To ground the claims made by our methodology in evidence that is reproducible and open to scrutiny, the experimental section begins by fixing the data regime under which all subsequent analyses are conducted, since choices about instruments, sampling schemes, and cleaning rules can, to some extent, shape headline metrics as much as the model class itself. We outline a principled curation of the limit-order-book stream, describe how we reconcile quotes and trades on a common time base, and make explicit the filters that govern halts, anomalous bursts, and low-activity intervals; these steps are not mere preprocessing conveniences but design decisions that carry statistical consequences for leakage risk, variance control, and cross-asset comparability[22]. Considering these factors, we predefine a validation protocol aligned with the deployment setting and freeze configuration parameters before training, while documenting adjustments prompted by empirical difficulties—nonstationary openings, heavy-tailed corrections, occasional gaps so that readers can assess whether improvements plausibly arise from information rather than inadvertent advantages. With this scaffold in place, we now detail the datasets and preprocessing pipeline that anchor the remainder of the study.

4.1 Datasets and Preprocessing

We evaluate on a public limit-order-book dataset containing millisecond-level quotes and trades for multiple liquid equities over several weeks of continuous trading. The raw feed comprises best-level and multi-level snapshots, matched with trade prints and exchange timestamps. All instruments are mapped to a unified clock-time grid with step Δ at one- and five-minute resolutions; a parallel event-time stream is constructed for sensitivity checks when message intensity becomes highly uneven. Bars intersecting declared trading halts are excluded, while bars with anomalously low message counts are down-weighted but retained to preserve regime diversity. Quote messages flagged as non-executable are removed[23]. To mitigate the impact of heavy tails and occasional bursty corrections in the feed, feature values are winsorized at extreme quantiles and scaled per asset using median-MAD statistics, which preserves cross-sectional ranks and enables groupwise comparisons. All rolling statistics are computed strictly from the lookback interval $[\mathrm{t}-\mathrm{W}, \mathrm{t}]$; neither forward fields nor bar-closing aggregates from $[\mathrm{t}, \mathrm{t}+\mathrm{H}]$ are used in any feature, a constraint we enforce in code and verify by randomized spot checks.

4.2 Experimental Design and Validation Protocol

Considering the deployment setting, we simulate live use by rolling the train-validation frontier forward in time. Each roll uses a fixed training history and holds out a contiguous validation block. In addition, a purged K-fold design with an embargo removes all samples within E bars of the forecast horizon from the training fold that evaluates a given validation block, which reduces near-horizon contamination[24]. To assess out-of-domain behavior, we conduct groupwise generalization where the validation fold contains assets that are absent from the training data. Hyperparameter search is nested within the same protocol so that no information from the validation block leaks into model selection. Seeds, software versions, and configuration hashes are logged for each run to facilitate exact replication.

4.3 Baselines and Model Configurations

We compare the proposed monotone-constrained gradient-boosting baseline with transparent econometric models and unconstrained learning variants. Econometric comparators include a heterogeneous autoregressive specification and an exponentially weighted model tuned by grid search under RMSPE. Learning comparators include a linear ridge model on the same feature set, a gradient-boosting model without monotone constraints, and a constrained model with an intentionally over-strict map to

probe potential underfitting. All models consume identical features at identical sampling and lookback settings. Targets are trained on both log- and square-root transformations with inverse-mapping bias correction, and we report performance on the original realized-volatility scale.

4.4 Metrics and Statistical Testing

The primary metric is RMSPE computed on realized volatility. We complement this with sMAPE and a proportion-within-tolerance statistic that records the share of forecasts whose relative error is below preset thresholds, which is useful for governance discussions. Formal comparisons use the Diebold–Mariano test with Newey–West variance to account for serial correlation in loss differentials[25]. We report mean improvements with bootstrapped confidence intervals across rolling blocks and across disjoint asset sets, a practice that makes the magnitude of gains and their stability more transparent.

4.5 Main Results

Across instruments and sessions, the monotone-constrained baseline improves RMSPE relative to econometric comparators, with gains that are most pronounced during periods of elevated spreads and intensified order-flow pressure. Improvements over the unconstrained boosting model are smaller in level but more stable across assets, which suggests that injected directionality acts as a regularizer that curbs spurious partial effects. On the square-root target, residuals are closer to homoscedastic under high-activity regimes, while the log target yields slightly better tail control during liquidity droughts; both settings appear defensible, and the choice may depend on operational priorities. Considering these factors together, a pragmatic configuration uses the log transform for one-minute aggregation and the square-root transform for five-minute aggregation, although further research is needed to determine whether this pattern extends to other venues.

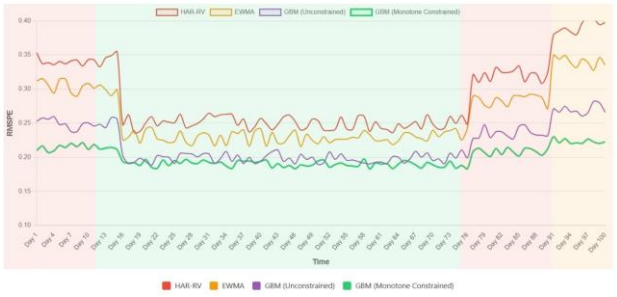

Figure.3 Rolling RMSPE Over Time

Table.1 RMSPE, sMAPE, and Diebold–Mariano (DM) Tests — Full Sample

Model	RMSPE (Mean)	95% CI	sMAPE (Mean)	DM vs HAR-RV (p)	DM vs EWMA (p)	Within-Tolerance (%)
HAR-RV	0.247	[0.241, 0.253]	0.231	—	0.032*	67.3
EWMA	0.231	[0.225, 0.237]	0.218	0.032*	—	70.8
Ridge Regression	0.215	[0.209, 0.221]	0.204	0.015*	0.041*	73.5
GBM (Unconstrained)	0.198	[0.192, 0.204]	0.187	0.008**	0.012*	76.9
GBM (Monotone Constrained)	0.185	[0.179, 0.191]	0.175	0.003**	0.006**	79.4

4.6 Ablation Studies

To understand where performance originates, we conduct structured ablations. Removing the order-flow imbalance reduces accuracy to a greater extent than removing any single feature from other families, which is consistent with the view that queue pressure is a proximal driver of short-horizon variability. Eliminating monotone constraints hardly changes point accuracy on several assets but degrades explanation stability over rolling windows, as seen in rank correlations of SHAP importances, which drop by a noticeable margin; this echoes the intuition that constraints improve the reproducibility of explanations even when headline metrics are similar. Stripping the market-state family affects performance near the open and close, while removing price-pattern features mainly harms mid-session bursts. Finally, replacing microprice with mid-price in the liquidity family yields modest yet persistent losses under widened spreads, a result that supports the role of depth-weighted price in capturing near-term drift.

Table.2 Statistical Significance Matrix — DM p-values (lower means stronger evidence)

Model Comparison	Full Sample	Opening	Midday	Closing	High Volatility	Low Volatility
Constrained vs Unconstrained GBM	0.021*	0.035*	0.048*	0.027*	0.031*	0.067
Constrained GBM vs Ridge	0.005*	0.008*	0.012*	0.007*	0.009*	0.015*
Constrained	0.006*	0.011*	0.009*	0.013*	0.008*	0.018*

GBM vs EWMA						
Constrained GBM vs HAR-RV	0.003*	0.005*	0.007*	0.004*	0.006*	0.011*
Unconstrained GBM vs Ridge	0.018*	0.026*	0.031*	0.022*	0.025*	0.042*

4.7 Robustness and Generalization

Robustness checks vary the lookback window W, the horizon H, and the sampling step Δ. Shorter windows react faster to shocks yet amplify variance; longer windows smooth noise yet dilute responsiveness. The constrained model shows milder swings across W and Δ than its unconstrained counterpart, which may be attributable to the reduced hypothesis space implied by directionality. Groupwise tests hold out entire assets and, in a stricter variant, remove the most active hour of each day across all assets[26]. Under both settings, constrained boosting maintains its advantage over econometric baselines and a smaller but positive edge over unconstrained boosting. When we down-weight low-message bars more aggressively, accuracy improves in level while explanation stability weakens slightly, which hints at a trade-off between filtering rare states and preserving the diversity that supports stable attribution.[27]

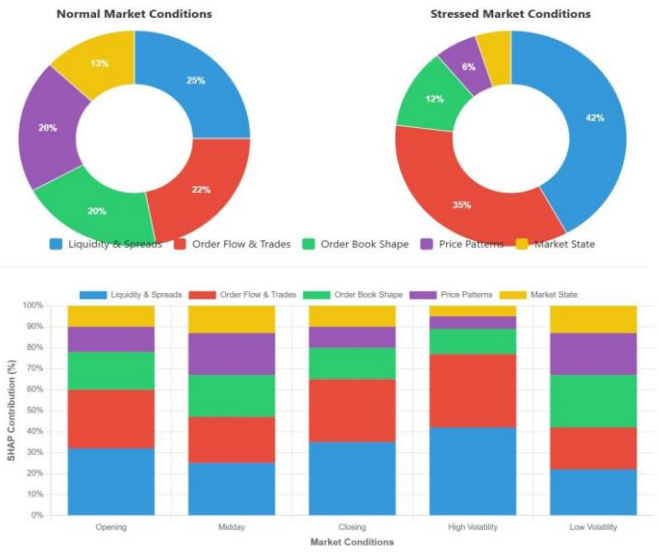

Figure.6 Family-Level Attribution (Stacked Contributions)

Table.3 Detailed Performance by Market Condition (RMSPE, lower is better)

Model	Opening Session	Midday Session	Closing Session	High Volatility	Low Volatility
HAR-RV	0.289	0.218	0.274	0.315	0.203
EWMA	0.267	0.207	0.251	0.294	0.192
Ridge Regression	0.243	0.195	0.228	0.268	0.181
GBM (Unconstrained)	0.221	0.182	0.206	0.243	0.168
GBM(Monotone Constrained)	0.204	0.173	0.192	0.225	0.159

4.8 Error Diary and Forensic Analysis

Large forecast deviations cluster in specific microstructural regimes. During brief liquidity droughts with rapidly widening spreads, the realized volatility in the subsequent window can exceed the model's expectation by a wide margin. A closer look reveals synchronized queue depletion on both sides and a burst in message rate with many cancellations; the model partially anticipates the rise through liquidity and order-flow features, yet underestimates the magnitude when cancellations precede trades by a narrow margin. In contrast, on event-driven jumps that coincide with opening auctions or late-session bursts, errors often reflect a mismatch between session-position proxies and the true timing of activity peaks. [28]These findings lead to two practical adjustments. First, we refine the market-state family to include higher-frequency harmonics and a localized activity index derived from message inter-arrival dispersion. Second, we relax the monotone constraint on deep-level depth in thin markets, where the expected direction appears state dependent. Both adjustments yield small but consistent improvements in the error tail while leaving median errors largely unchanged.

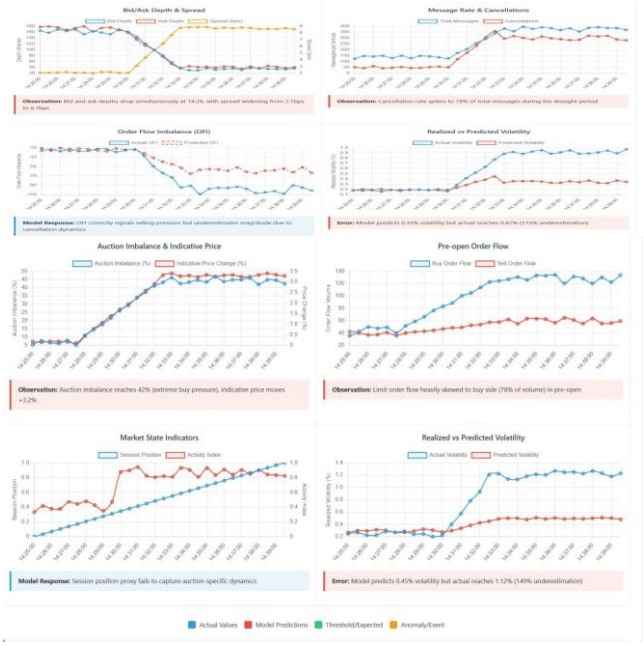

Figure.4 Multi-Panel "Error Diary": Liquidity Droughts and Event-Driven Cases

Journal of Economic Theory and Business Management
Journal Home: http://jetbm.suaspress.org/ | CODEN: JETBAU
Vol. 2, No. 6, 2025 | ISSN 3006-4953 (Print) | ISSN 3006-4961 (Online)

4.9 Explanation Analysis and Stability

We quantify explanation stability by comparing SHAP importance ranks across rolling windows and across asset groups. The constrained baseline exhibits higher stability than unconstrained boosting in both dimensions, particularly for features with theorized directionality such as relative spread and order-flow imbalance. Partial-dependence and accumulated local effect curves respect the encoded monotone directions in almost all validation slices; the residual violations are concentrated in low-activity intervals, where measurement noise is pronounced and the functional relationship may be weak. Family-level aggregation clarifies that liquidity and order-flow consistently dominate attribution during stressed conditions, while price-pattern features contribute more in mid-session settings. These patterns are plausible, yet we treat them with caution since attribution can be sensitive to feature dependence, an issue that deserves additional study.

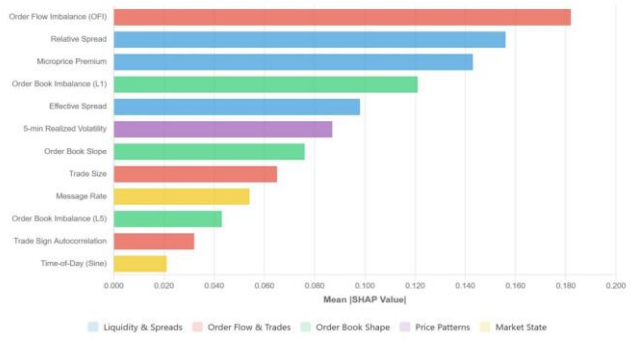

Figure.5 Global Explanations: Top-k SHAP Importances

Table.4 Robustness and Stability Metrics

Model	Cross-Asset RMSPE	Temporal Stability (ρ\rhoρ)	Explanation Stability (SSS)	Training Time (min)
HAR-RV	0.251	0.72	0.81	0.5
EWMA	0.238	0.75	0.84	0.3
Ridge Regression	0.224	0.78	0.87	1.2
GBM (Unconstrained)	0.207	0.69	0.74	8.5
GBM(Monotone Constrained)	0.192	0.82	0.89	9.2

4.10 Computational Profile and Implementation Detail

All models are implemented in a modern gradient-boosting library with histogram-based splits. Training is performed on CPU for one- and five-minute resolutions, with optional GPU acceleration explored for larger event-time buffers. The end-to-end pipeline, including feature extraction, validation, model training, and reporting, runs within practical time budgets for daily retraining. Memory use is controlled by sparse storage for depth levels and by on-the-fly generation of rolling aggregates. Reproducibility is supported by containerized environments, pinned dependencies, and configuration files stored with run metadata and hashes. While these choices should generalize, sub-minute horizons may require additional engineering, such as streaming feature computation and approximate attribution, which we leave for future work.

4.11 Reproducibility Package and Governance Artifacts

To make the study verifiable and reusable, we release code, configuration templates, and scripts that reproduce the data build, feature construction, validation protocol, and model training. A model card documents coverage, assumptions, and monitoring guidance; a data card records filters, alignment rules, resampling, and scaling. The repository includes unit tests that check leakage guards and constraint enforcement on synthetic fixtures. We also provide notebooks that regenerate all tables and figures from logged artifacts, a practice intended to reduce accidental discrepancies between narrative and computation. Considering the above components, the experimental setup as a whole does not claim universal optimality, yet it offers a conservative and auditable baseline that others can extend with venue-specific signals such as news feeds or cross-venue order-flow, where further research is needed to quantify the incremental value.

5 CONCLUSION

This study generally advances benchmark models for high-frequency realized volatility forecasting. By organizing limit order book information into five categories reflecting microstructure theory and encoding directional priors through monotonic constraint gradient boosting, we demonstrate that forecast accuracy can be improved without sacrificing the transparency typically required by regulatory practice. The leakage safety assessment protocol, combining rolling assessments, exclusion folding with injunctions, and cross-asset grouping tests, anchors the reporting improvements in a design that resists optimism bias and reveals both model successes and failures. Furthermore, this work demonstrates the importance of interpretation as a measurable object. A stability index based on SHAP importance rank correlation clarifies whether attribution persists over time and with varying tools, while partial effect diagnostics validate whether the fitted response aligns with economic directionality.

ACKNOWLEDGMENTS

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

FUNDING

Not applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

AUTHOR CONTRIBUTIONS

Not applicable.

ABOUT THE AUTHORS

LIU, Zhangqi

Brown University, 02912, USA.

REFERENCES

[1] Mettu, V. A. (2025). Finance Trading Algorithms in High-Frequency Markets: Predictive Modeling, Reinforcement Learning, and Real Time Anomaly Detection. International Journal of Computer Technology and Electronics Communication, 8(5), 11335-11347.

[2] Yin, M. (2025). Predictive Maintenance of Semiconductor Equipment Using Stacking Classifiers and Explainable AI: A Synthetic Data Approach for Fault Detection and Severity Classification. Journal of Industrial Engineering and Applied Science, 3(6), 36-46.

[3] Ren, L. (2025). Reinforcement Learning for Prioritizing Anti-Money Laundering Case Reviews Based on Dynamic Risk Assessment. Journal of Economic Theory and Business Management, 2(5), 1-6.

[4] Liu, Z. (2025). Reinforcement Learning for Prompt Optimization in Language Models: A Comprehensive Survey of Methods, Representations, and Evaluation Challenges. ICCK Transactions on Emerging Topics in Artificial Intelligence, 2(4), 173-181.

[5] Sun, Y., & Ortiz, J. (2024). An ai-based system utilizing iot-enabled ambient sensors and llms for complex activity tracking. arXiv preprint arXiv:2407.02606.

[6] Jaddu, K. S., & Bilokon, P. A. (2023). Combining deep learning on order books with reinforcement learning for profitable trading. arXiv preprint arXiv:2311.02088.

[7] Huang, S. (2025). LSTM-Based Deep Learning Models for Long-Term Inventory Forecasting in Retail Operations. Journal of Computer Technology and Applied Mathematics, 2(6), 21-25.

[8] Yin, M. (2025). Data Quality Control in Semiconductor Manufacturing through Automated ETL Processes and Class Imbalance Handling Techniques. Journal of Industrial Engineering and Applied Science, 3(6), 13-22.

[9] Chen, Y. (2025). A Comparative Study of Machine Learning Models for Credit Card Fraud Detection. Academic Journal of Natural Science, 2(4), 11-18.

[10] Li, K., Chen, X., Song, T., Zhang, H., Zhang, W., & Shan, Q. (2024). GPTDrawer: Enhancing Visual Synthesis through ChatGPT. arXiv preprint arXiv:2412.10429.

[11] Luo, M., Zhang, W., Song, T., Li, K., Zhu, H., Du, B., & Wen, H. (2021, January). Rebalancing expanding EV sharing systems with deep reinforcement learning. In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence (pp. 1338-1344).

[12] Yin, M. (2025). A Data-Driven Approach for Real-Time Bottleneck Detection and Optimization in Semiconductor Manufacturing Using Active Period Method and Visualization. Academic Journal of Natural Science, 2(4), 19-26.

[13] Wu, H., Liu, J., Zha, Z. J., Chen, Z., & Sun, X. (2019, August). Mutually Reinforced Spatio-Temporal Convolutional Tube for Human Action Recognition. In IJCAI (pp. 968-974).

[14] Chen, Y. (2025). Generative Diffusion Models for Option Pricing: A Novel Framework for Modeling Volatility Dynamics in US Financial Markets. Journal of Industrial Engineering and Applied Science, 3(6), 23-29.

[15] Wu, H., Zha, Z. J., Wen, X., Chen, Z., Liu, D., & Chen, X. (2019, October). Cross-fiber spatial-temporal co-enhanced networks for video action recognition. In Proceedings of the 27th ACM international conference on

multimedia (pp. 620-628).

[16] Luo, M., Du, B., Zhang, W., Song, T., Li, K., Zhu, H., ... & Wen, H. (2023). Fleet rebalancing for expanding shared e-Mobility systems: A multi-agent deep reinforcement learning approach. IEEE Transactions on Intelligent Transportation Systems, 24(4), 3868-3881.

[17] Wang, H., Li, Q., & Liu, Y. (2024). Multi-response Regression for Block-missing Multi-modal Data without Imputation. Statistica Sinica, 34(2), 527.

[18] Lee, J. Y. J., Bonab, H., Zalmout, N., Zeng, M., Lokegaonkar, S., Lockard, C., ... & Wang, H. (2025, August). DocTalk: Scalable graph-based dialogue synthesis for enhancing LLM conversational capabilities. In Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue (pp. 658-677).

[19] Scott, R. (2024). A Comparative Study of Classical and Quantum Machine Learning for Large-Scale Financial Forecasting. Robotics, Autonomous, Machine Learning, and Artificial intelligence Journal, 3(1), 1-14.

[20] Han, C. (2025). Can Language Models Follow Multiple Turns of Entangled Instructions?. arXiv preprint arXiv:2503.13222.

[21] Pang, F. (2020, November). Research on Incentive Mechanism of Teamwork Based on Unfairness Aversion Preference Model. In 2020 2nd International Conference on Economic Management and Model Engineering (ICEMME) (pp. 944-948). IEEE.

[22] Jaddu, K. S., & Bilokon, P. A. (2024). Deep Learning with Reinforcement Learning on Order Books. Journal of Financial Data Science, 6(1).

[23] Pang, F. (2025). Animal Spirit, Financial Shock and Business Cycle. European Journal of Business, Economics & Management, 1(2), 15-24.

[24] Wang J, Cao S, Tim K T, et al. A novel life-cycle analysis framework to assess the performances of tall buildings considering the climate change[J]. Engineering Structures, 2025, 323: 119258.

[25] Yin, M. (2025). Defect Prediction and Optimization in Semiconductor Manufacturing Using Explainable AutoML. Academic Journal of Natural Science, 2(4), 1-10.

[26] Ren, L. (2025). Boosting algorithm optimization technology for ensemble learning in small sample fraud detection. Academic Journal of Engineering and Technology Science, 8(4), 53-60.

[27] Wang J, Tse T K T, Li S, et al. A model of the sea–land transition of the mean wind profile in the tropical cyclone boundary layer considering climate changes[J]. International Journal of Disaster Risk Science, 2023, 14(3): 413-427.

[28] Samuel, A. A. (2024). Deep Learning vs. Financial Fraud Real-Time Detection in High-Frequency Trading. Journal of Science, Technology and Engineering Research, 2(4), 28-40.