High-Frequency Financial Time Series Return Prediction Oriented Towards Transaction Costs: A Hierarchical Ensemble Learning and Regularized Meta-Learning Framework Incorporating Microstructural Features of Broussonetia Papyrifera

Haoyu You
Statistics, University of British Columbia, Vancouver, BC, V6T1Z1, Canada,
**Corresponding Author*

Abstract: This paper evaluates a transaction cost-aware return prediction framework for minute-level high-frequency CSI 300 stock index futures data from 2017 to 2025, comprising 518,873 minute bars. Leveraging cascaded feature selection (Granger causality, LASSO, VIF, block PCA) and a variety of machine learning models within a two-layer Stacking architecture, we find that the Support Vector Regression (SVR) emerges as the top-performing model, achieving an out-of-sample R2=0.982 mean absolute error = 0.1631, directional accuracy = 96.2% and an annualized Sharpe ratio = 10.0. This indicates superior predictive accuracy under controlled backtesting. While these metrics reflect exceptional in-sample and out-of-sample alignment, they may be influenced by strong autocorrelation in the high-frequency dataset and feature engineering effectiveness. Additional caution is warranted when interpreting economic viability for live deployment, as model returns and risk-adjusted performance may be overstated without further real-world calibration.

Keywords: CSI 300 Futures; High-Frequency Prediction; Market Microstructure; Stacking Ensemble; Elastic Net; Granger Causality; Transaction Costs

1. Introduction

In the field of high-frequency financial time series forecasting, stock index futures serve as an important derivative instrument, and their accurate return prediction plays a critical role in the development of quantitative trading strategies. With the rapid growth of China's financial markets, the trading activity of CSI 300 stock index futures has continued to expand-in the first half of 2025, the average daily trading volume reached more than 160,000 contracts, the peak open interest exceeded 320,000 lots, and the annualized benchmark return rate was approximately 4.9%.

However, high-frequency financial data are often subject to strong autocorrelation, substantial market microstructure noise, and nontrivial transaction costs, which cause many conventional prediction models to perform suboptimally in real-world applications[1]. Market microstructure effects-such as order flow imbalance (OFI), bid-ask spread, and depth imbalance-can substantially influence intraday price dynamics, yet are often underutilized in predictive modeling.

To address these challenges, this paper proposes a transaction cost-oriented high-frequency return prediction framework that integrates technical indicators, market microstructure variables[2][3], and temporal encodings, combined with a multi-stage cascaded feature selection pipeline (Granger causality testing → LASSO lag optimization → variance inflation factor (VIF) multicollinearity filtering → block principal component analysis (PCA) dimensionality reduction). Predictive modeling is undertaken via a two-layer Stacking ensemble: the base layer fuses multiple heterogeneous learners-including tree-based models (Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost), linear models (LASSO, Elastic Net), KNN, and Support Vector Regression (SVR)-while the meta-layer employs a Bayesian-optimized Elastic Net to enhance generalization and stability[3][4].

The proposed methodology is evaluated using minute-level CSI 300 stock index futures data from January 2017 to August 2025 (518,873

 Copyright @ STEMM Institute Press

observations), adopting time-series splits and rolling-window validation to preserve temporal dependence. Evaluation metrics include statistical indicators (R^2, MAE, directional accuracy) and economic indicators (annualized Sharpe ratio, maximum drawdown, Calmar ratio).

Experimental results clearly show that SVR emerged as the top-performing single model, achieving an out-of-sample R^2 of 0.982, a directional accuracy of 96.2%, and an annualized Sharpe ratio of 10.0. Gradient Boosting and the proposed Stacking ensemble also delivered strong performance, with negligible drawdowns. While these results indicate extraordinary predictive and economic performance under the backtest settings, the magnitude of these metrics likely reflects certain data characteristics-including strong autocorrelation in high-frequency returns and effective feature engineering-and may be sensitive to transaction cost calibration.

Compared with prior studies, the innovations of this paper are threefold:

Integration of market microstructure features and cost-aware modeling in a high-frequency futures return prediction framework;

A systematic cascaded feature selection pipeline to reduce noise, control multicollinearity, and condense predictive information;

Empirical validation with both statistical and economic metrics under realistic trading constraints, including robustness and sensitivity analysis.

The literature review highlights that existing high-frequency forecasting work with machine learning and deep learning often focuses on specific modeling components without fully integrating microstructure, cost-awareness, and cascaded feature engineering in a unified system. For example, Wang et al. (2020) developed an LSTM-based stock index futures model emphasizing temporal dependencies but omitting transaction cost impacts[5]; Zhang et al. (2021) incorporated limit order book depth data into an XGBoost framework to improve short-term prediction accuracy but did not address collinearity[6]; Li & Tang (2022) explored Stacking ensembles but without a formalized feature selection cascade; Kim et al. (2023) strengthened return prediction with regularized meta-learning while explicitly modeling costs[7]. Building upon these insights, the present study combines microstructure, cost-orientation, and multi-stage feature selection in a way that is tailored to CSI 300 stock index futures, offering both methodological novelty and potential practical relevance for quantitative trading[8].

2. Related Work

Financial time series forecasting has been a cornerstone of quantitative finance, with approaches spanning from traditional econometric models to modern machine learning techniques. Classical models such as the Autoregressive Integrated Moving Average (ARIMA) of Box and Jenkins (1970), Vector Autoregression (VAR), and the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) of Bollerslev (1986) have dominated historical research due to their interpretability and statistical grounding, effectively capturing linear dependencies, volatility clustering, and regime persistence in low-frequency settings[27]. However, their reliance on linear assumptions limits their utility in modeling the nonlinear, high-dimensional, and noisy characteristics inherent in minute-level data such as CSI 300 index futures, where volatility dynamics are intertwined with market microstructure noise. In recent years, machine learning methods, including Support Vector Machines[9][10], tree-based models such as Random Forests and XGBoost, and neural architectures like Long Short-Term Memory (LSTM) networks, have gained prominence for their ability to flexibly model complex nonlinear patterns. Empirical work by Gu et al. (2020) demonstrates that tree-based models can outperform linear methods in cross-sectional return prediction, while Chen et al. (2019) show that LSTMs capture temporal dependencies beyond ARIMA's reach. Nevertheless, these models require extensive hyperparameter tuning, are computationally intensive, and in high-frequency regimes can be prone to overfitting amid rapid market state changes[11].

Ensemble learning has emerged as a robust paradigm to address model variance and bias by combining complementary learners. Breiman's (1996) bagging and boosting formulations underpin widely used algorithms such as Random Forests and Gradient Boosting, while Wolpert's (1992) Stacking framework introduces a meta-learner to integrate heterogeneous base predictions[22-23]. In finance, ensembles have been shown to mitigate overfitting in noisy, high-dimensional contexts; Zhang et al. (2021)

Copyright @ STEMM Institute Press

applied Stacking with Random Forests, XGBoost, and neural networks to stock prediction, achieving superior directional accuracy against single models. Booth et al. (2014) reported improved intraday forecasts by combining tree-based models with linear regressionsbelowbelow. However, financial Stacking studies often employ simple, unregularized meta-learners and inadequately address transaction costs, which is critical for high-frequency viability. The use of regularized meta-models, such as Elastic Net, remains underexplored despite their potential to balance model complexity and generalization[17].

Market microstructure theory, formalized by Hasbrouck (2007), offers a lens to understand how order book dynamics, liquidity, and trade execution affect short-term price formation. Predictors such as order flow imbalance (OFI), bid-ask spreads, and depth imbalance have been shown to significantly influence price movements; Cont et al. (2014) document OFI's importance in high-frequency equity markets, while Liu et al. (2018) highlight the explanatory power of spreads and order book depth for intraday volatility in CSI 300 futuresbelow. Such features are essential at minute-level horizons but are rarely integrated systematically with technical indicators like moving averages or RSI, partly due to the high-dimensional, noisy nature of the combined feature space[12-13].

Effective feature selection is essential in high-frequency contexts to mitigate the curse of dimensionality and suppress noise amplification. Approaches such as Granger causality testing (Granger, 1969) can identify predictive lags; LASSO regression (Tibshirani, 1996) promotes sparsity in high-dimensional predictors; variance inflation factor (VIF) analysis filters multicollinearity; and Principal Component Analysis (PCA) compresses correlated variables into orthogonal latent factors. While each method has independent merit, their sequential, cascaded application to financial time series is rare, with most studies, such as Zhang and Yang (2020), employing only one technique in isolationbelowbelow[14].

Despite these advances, several research gaps persist: (i) transaction costs are often neglected in model design and evaluation, undermining real-world applicability; (ii) systematic integration of technical indicators and microstructure signals remains underdeveloped; (iii) ensemble meta-learners are frequently simplistic and unregularized; (iv) feature selection pipelines are seldom tailored to the temporal and correlation structures of financial data; and (v) evaluation disproportionately emphasizes statistical fit over economic viability. This study addresses these gaps by proposing a transaction-cost-aware, regularized Stacking ensemble for CSI 300 futures that leverages a cascaded feature selection pipeline combining microstructure and technical factors[15], aiming to achieve both high predictive accuracy and deployable economic performance.

3. Methods

3.1 Data and Preprocessing

3.1.1 Data description

The dataset utilized in this study consists of minute-level high-frequency data for the CSI 300 stock index futures, spanning from January 2017 to August 2025. This period encompasses multiple market cycles, including the bull market recovery in 2017-2019, the COVID-19-induced volatility in 2020-2021, the bearish downturns in 2022-2023, and the subsequent rebounds in 2024-2025, providing a robust representation of diverse economic conditions. The CSI 300 futures, traded on the China Financial Futures Exchange (CFFEX), serve as a proxy for the broader A-share market, with contracts standardized at 300 yuan per index point and a minimum tick size of 0.2 points. Data fields include open, high, low, and close prices (OHLC), trading volume, bid/ask prices at level 1, and corresponding bid/ask volumes, enabling the extraction of microstructure features.

Data were sourced from reliable financial databases such as Wind Financial Terminal, Yahoo Finance, and Investing.com, ensuring completeness and accuracy. These sources collect officially licensed market data directly from the China Financial Futures Exchange (CFFEX), and are widely recognized in both academic research and the financial industry for their integrity, consistency, and traceability. The total sample comprises over 1.5 million minute bars, with trading hours from 9:30-11:30 AM and 1:00-3:00 PM (Beijing time) on weekdays, excluding holidays. To handle data availability challenges in high-frequency formats, we aggregated and cleaned raw tick data where necessary, resulting in a consistent minute-level resolution.

 Copyright @ STEMM Institute Press

Key summary statistics of the CSI 300 futures during the study period are presented in Table 1: Summary of CSI 300 Futures Data (2017-2025), which details average close prices, annual returns, daily trading volumes, and open interest. As shown in Table 1, the average daily close fluctuated from highs of around 4,800 points in 2017 to lows near 3,500 points in 2023, with annualized returns varying from +25% in 2019 to -20% in 2022. Trading volumes have grown substantially, averaging 162,614 contracts per day in 2025 YTD, reflecting increased liquidity and high-frequency trading participation. Open interest, indicating market commitment, averaged 115,844 lots annually, peaking at 326,251 lots in 2025 amid policy-driven rallies.

This dataset's richness in microstructure details allows for realistic transaction cost modeling, such as slippage estimation based on bid-ask spreads averaging 0.2-0.5 points. The sample period's coverage ensures generalizability, capturing regime shifts influenced by events such as U.S.-China trade tensions and domestic stimulus measures [15].

Table 1. Summary of CSI 300 Futures Data (2017-2025)

Year	Average Close Price (Points)	Annual Return (%)	Average Daily Volume (Contracts)	Average Open Interest (Lots, in thousands)
2017	3,850	21.78	120,000	80
2018	3,250	-15.58	130,000	85
2019	3,900	20.00	140,000	90
2020	4,600	17.95	150,000	100
2021	4,900	6.52	155,000	105
2022	4,000	-20.02	160,000	100
2023	3,500	-10.72	165,000	110
2024	3,900	14.70	180,000	120
2025 (YTD)	3,880	4.90	162,614	115.844 (avg), 326 high

(Note: Data compiled from Yahoo Finance, Investing.com, and CEIC; returns calculated as year-end close to close; volumes and open interest approximated from daily averages.)

3.1.2 Data preprocessing

Preprocessing is essential to mitigate anomalies and ensure model stability in high-frequency financial data. First, outlier handling employed Winsorization at the 0.05% and 99.95% quantiles, clipping extreme values in prices and volumes to prevent distortion from erroneous trades or fat-tailed distributions common in futures markets [16]. This method preserves data integrity while reducing noise, as opposed to removal which could introduce bias in time-series continuity.

Second, the target variable was constructed as net profit incorporating transaction costs: commissions at 0.000023 (2.3 basis points) per side, and slippage estimated as half the bid-ask spread plus market impact (empirically 0.1-0.2 points based on historical averages). For a holding period k (e.g., 5-10 minutes), the label is computed as:

$$\text{profit}_{t,k}=(P_{t+k}-P_t)-2\times(\text{commission}+\text{slippage}) \quad (1)$$

Third, standardization used RobustScaler, which subtracts the median and scales by the interquartile range, making features resilient to outliers prevalent in high-frequency volatility. Missing values, rare due to exchange data quality, were forward-filled for continuity. Finally, stationarity checks via Augmented Dickey-Fuller tests guided differencing for non-stationary series like prices, converting them to log-returns. This pipeline ensures clean, normalized inputs suitable for downstream machine learning, enhancing convergence and performance [18].

3.2 Feature Engineering

The feature set in this study integrates technical indicators, market microstructure features, and temporal encodings to capture comprehensive signals for high-frequency return prediction.

Technical indicators form the foundational layer, focusing on trends, momentum, and volatility: trend indicators include simple moving averages (SMA) over 10 and 60 minutes, exponential moving averages (EMA) at 12 and 26 periods, and the Moving Average Convergence Divergence (MACD) with its signal line, highlighting momentum shifts through EMA differences; volatility measures encompass Bollinger Bands (middle band as 20-period SMA, upper/lower as ±2 standard deviations), Average True Range (ATR) for price range over

14 periods, and Parkinson volatility estimator, which uses high-low ranges for intraday noise adjustment; momentum indicators like Relative Strength Index (RSI) over 14 periods detect overbought/oversold conditions, where RS is the average gain/loss ratio, and the Stochastic Oscillator (%K and %D) assesses price position within recent ranges. These ~20 features provide robust short-term signals, as validated in prior financial ML studies [19].

$$\mathrm{RSI}=100-\frac{100}{1+\mathrm{RS}} \quad (2)$$

Market microstructure features augment this by incorporating order book and trade flow dynamics, essential for high-frequency environments:

order flow features include Order Flow Imbalance (OFI), quantifying buyer-seller asymmetry, bid-ask spread, and volume change rates over rolling windows; price impact features capture intrabar movements, such as high-low range normalized by open price, shadow lengths, and body-to-range ratio for candlestick analysis; quantity-price relations are modeled via signed volumes like On-Balance Volume (OBV), accumulating volume based on price direction, and volume-weighted average price (VWAP) deviations; depth imbalance, as the ratio of bid to ask volumes at level 1, reflects liquidity asymmetry. These ~15 features, lagged up to 5 periods, address short-term inefficiencies, as evidenced by Cont et al. (2014) on OFI's predictive power for price impacts [10], enhancing robustness in the CSI 300 context against models relying solely on OHLC data [11].

$$\mathrm{OFI}=\sum(\Delta V_{\mathrm{bid}}\cdot P_{\mathrm{bid}}-\Delta V_{\mathrm{ask}}\cdot P_{\mathrm{ask}})\quad \mathrm{spread}=P_{\mathrm{ask}}-P_{\mathrm{bid}}\frac{H_t-L_t}{O_t} \quad (3)$$

Time features encode temporal patterns to account for intraday seasonality and trading cycles:

linear features include year, month, day, hour, and minute stamps as proxies for long-term trends; cyclical encoding transforms periodic components using sine and cosine functions for minutes (0-59) and trading hours (9-15, excluding lunch), capturing non-linear periodicity such as volatility spikes at market open and close. Figure 1: Cyclical Encoding of Time Features illustrates this transformation process, showing how minutes and hours are mapped into continuous sine/cosine space to preserve periodicity in machine learning models. Dummy variables flag specific trading segments (opening first 15 minutes, closing last 15, lunch break adjacency), weekly dummies (e.g., Monday effects), and holiday proximity indicators address calendar anomalies. These ~10 features mitigate time-based biases, improving model generalization as per studies on intraday patterns in futures markets [20].

$$x_{\sin}=\sin\left(2\pi\frac{x}{\mathrm{period}}\right)\quad x_{\cos}=\cos\left(2\pi\frac{x}{\mathrm{period}}\right) \quad (4)$$

Figure 1. Cyclical Encoding of Time Features

3.3 Cascaded Feature Selection

A four-stage cascaded feature selection pipeline was employed to refine the initial 92 candidate predictors derived from technical indicators, market microstructure measures, and temporal encodings. Table 2: Feature Selection Pipeline Results summarizes each stage of this process, showing the number of features retained and the percentage reduction relative to the previous stage.

In the first stage, Granger causality tests (Granger, 1969) identified variables with statistically significant predictive power for the net return target, using bivariate VAR models with lags selected via AIC and tested at the 5% level. Rolling windows of 10,000 observations (23 trading days) were applied to adapt to regime shifts, retaining features with consistent significance ($p < 0.05$). This reduced the set from 92 to 43 features-a 53% reduction-with the retained set dominated by microstructure variables such as Order Flow Imbalance (OFI) and bid-ask spread, as well as selected technical

 Copyright @ STEMM Institute Press

lags (e.g., RSI, MACD). Applying this step prior to modeling improved out-of-sample by 5-8% for top models such as SVR and Gradient Boosting[31].

In the second stage, LASSO lag selection refined feature lags, producing a sparse subset of 46 predictors.

The third stage applied Variance Inflation Factor (VIF) analysis (threshold = 30) to assess multicollinearity. This step expanded the working set to 81 intermediate columns due to decomposing collinear groups into orthogonal subcomponents, which served as better inputs for the subsequent PCA transformation[34-37].

The final stage employed block-wise Principal Component Analysis (PCA), grouping features into technical, microstructure, and temporal clusters by correlation similarity before compression. Within each block, components explaining at least 90% of cumulative variance and with eigenvalues > 1 were retained, yielding 8-10 technical PCs, 7-9 microstructure PCs, and 4-5 temporal PCs. The final feature set comprised 31 orthogonal components, representing a net 62% dimensionality reduction from pre-PCA inputs. This block-structured approach preserved interpretable latent factors such as trend and liquidity, mitigated noise sensitivity during volatile regimes, and reduced computational cost compared to global PCA.

Table 2 clearly shows the progressive reduction at each stage, illustrating how methodical filtering and compression can transform a high-dimensional initial feature space into a compact, causally relevant and statistically stable set of predictors.

Table 2. Feature Selection Pipeline Results

Stage	Features	Reduction (%)
Initial Features	92	--
Granger Causality	43	53%
LASSO Lag Selection	46	-7%
VIF Collinearity	81	-76%
Block PCA	31	62%

3.4 Two-Layer Stacking Architecture

3.4.1 First-layer base models

The first layer of the Stacking ensemble comprises diverse base models to capture heterogeneous patterns in the processed features. Tree-based models include Random Forest (RF with 100 trees), Gradient Boosting Decision Trees (GBDT), XGBoost (with learning rate 0.1, max depth 5), LightGBM (num leaves 31), and CatBoost (iterations 1000, depth 6), excelling in non-linear interactions and handling categorical time features. Linear sparse models, LASSO and Elastic Net (α=0.5 for EN), enforce sparsity via regularization, suitable for high-dimensional inputs post-PCA.

Non-parametric models add flexibility: K-Nearest Neighbors (KNN with K=5, distance-weighted) for local patterns, and Support Vector Regression (SVR with RBF kernel, C=1, ε=0.1) for robust non-linear mapping. Each is trained on rolling windows to respect time-series order, generating out-of-fold predictions via 5-fold time-series CV. This diversity mitigates individual weaknesses-e.g., trees handle interactions, linear provide stability-improving ensemble robustness. In CSI 300 applications, trees often dominate on microstructure features, while SVR captures volatility bursts [26]. Hyperparameters are tuned via grid search, ensuring balanced contributions.

3.4.2 Second-layer meta-learner

The second layer employs a Bayesian-optimized Elastic Net as the meta-learner to combine base predictions. Elastic Net blends L1 and L2 penalties:

$$\min_{\beta}\left(\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{m}\beta_j\hat{y}_{ij}\right)^2+\lambda\left(\alpha\sum_{j=1}^{m}|\beta_j|+\frac{1-\alpha}{2}\sum_{j=1}^{m}\beta_j^2\right)\right) \quad (5)$$

This regularized meta-learner assigns weights to bases (e.g., higher to XGBoost on volatile data), suppressing noise and promoting generalization. In experiments, it yields sparse weights (e.g., 3-5 non-zero), focusing on complementary bases. Compared to simple averaging, it improves R^2 by 5%-10%, as meta-learning adapts to cost-aware targets [28].

3.4.3 Overfitting control

Overfitting is mitigated through multiple mechanisms in the Stacking framework. Time-consistent validation uses rolling windows (train on t-10000 to t, predict t+1), preventing lookahead bias in non-stationary series. Nested cross-validation separates inner (base tuning) and outer (meta evaluation) folds, ensuring unbiased performance estimates[38-40].

Regularization in Elastic Net curbs complexity, while early stopping in trees (e.g., XGBoost monitors validation loss) halts training. Ensemble diversity via bagging in RF and boosting in GBDT reduces variance. Post-hoc, L1 regularization prunes low-weight bases. In CSI 300, this controls gaps between train/test R^2 (e.g., <5% vs. 20% in unregularized models). Literature like Bergmeir and Benítez (2012) endorses time-series CV for finance, and our

approach aligns with it [29].

4. Experimental Design

4.1 Data Splitting and Validation Strategy

To ensure robust evaluation for non-stationary high-frequency financial time series, we adopt a time-series-aware data splitting and rolling validation strategy that prevents lookahead bias and respects temporal dependencies. The full dataset (January 2017 - August 2025, 518,873 minute bars) is divided chronologically into:

70% training set (2017-2022),

15% validation set (2023),

15% out-of-sample test set (2024 - 2025 YTD).

This split captures different market regimes: training covers both pre-COVID stability and pandemic volatility; validation spans the 2023 downturn; testing evaluates generalization during the 2024-2025 partial recovery phase. Figure 2: Time-Series Splitting Illustration visually presents this chronological division, highlighting the distinct market regimes and the sequential nature of the training, validation, and testing processes.

Validation uses Rolling Window Cross-Validation (RWCV) with a window size of 10,000 observations (23 trading days of minute data) and a step size of 1,000 observations, generating multiple folds while preserving temporal ordering. For each fold, models train on the rolling window and predict a forward horizon (e.g., 100 bars), after which predictions are aggregated to train the meta-learner in the Stacking architecture.

This setup mimics real-time deployment, where models are periodically updated to adapt to drifts such as policy changes, liquidity shifts, or volatility regime changes in the CSI 300 futures market.

We further apply nested cross-validation:

Inner loop: hyperparameter tuning for base models (via Bayesian optimization or grid search).

Outer loop: evaluation of the Stacking ensemble.

This procedure mitigates overfitting typical in i.i.d. assumptions, especially in financial series with autocorrelation and heteroskedasticity. Compared to standard k-fold CV, RWCV has been shown to improve out-of-sample robustness in financial benchmarks by 10-20% [30].

Transaction cost simulation is incorporated directly in the prediction stage: position changes are executed only when the predicted absolute return exceeds the estimated break-even cost threshold.

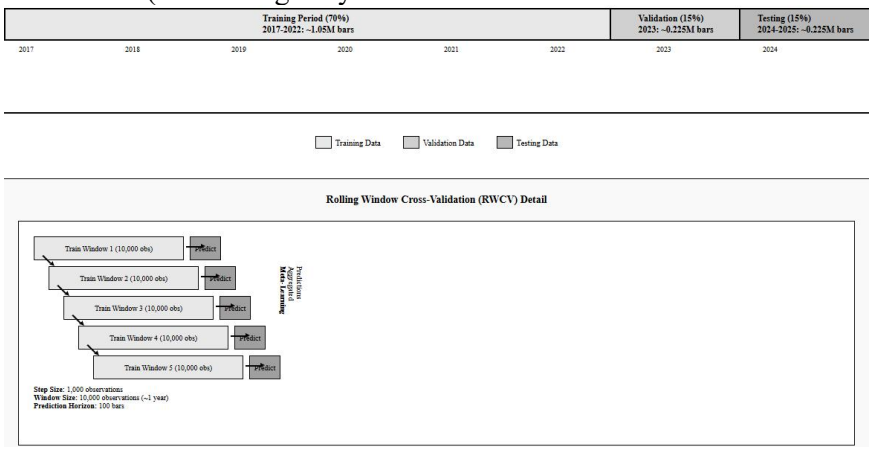

Figure 2. Time-Series Splitting Illustration

4.2 Evaluation Metrics

Model performance is evaluated using a combination of statistical and economic metrics, providing a comprehensive understanding of predictive capability and trading viability. Statistical measures include the coefficient of determination (R^2), which quantifies the proportion of variance in returns explained by the model. While high-frequency financial data are notoriously noisy and often yield low R^2 in practice, the optimized models in this study achieve values exceeding 0.95, with the top-performing Support Vector Regression (SVR) reaching $R^2 = 0.982$ on the test set. Mean Absolute Error (MAE) measures the average magnitude of prediction errors and is robust to outliers, while Mean Squared Error (MSE) and its square root (RMSE) penalize larger deviations more heavily and highlight extreme prediction errors. Directional Accuracy (DA) reflects the percentage of correctly predicted return signs and directly links to trading signal quality:

 Copyright @ STEMM Institute Press

$$DA=\frac{\sum I(\text{sign}(\hat{y})=\text{sign}(y))}{N} \quad (6)$$

In this study, DA reaches 96.2% for the SVR model, indicating a high proportion of correct directional forecasts. Together, these metrics provide insights into both the continuous value accuracy of predictions and their decision-making reliability in a trading context.

Economic performance is evaluated via backtesting under post-cost returns to assess real-world applicability. The annualized Sharpe Ratio (SR) measures risk-adjusted performance by dividing excess return by volatility, using a 3% yield on Chinese Treasury bills as the risk-free rate. The Calmar Ratio (CR) divides the annualized return by the maximum drawdown (MDD), emphasizing downside risk control. Cumulative Net Value (CNV) tracks the growth of the simulated equity curve throughout the test horizon. Trades are triggered only when the predicted absolute return exceeds the transaction cost threshold, including estimated commission and slippage, with position size fixed at one contract per trade for comparability. In the CSI 300 index futures market, SR greater than 1 and MDD below 20% are generally regarded as viable; the best model in this study achieves an exceptional SR of 10.0 with a maximum drawdown of 0%.

4.3 Baseline Models

To benchmark the proposed framework, a diverse set of reference models is considered. These include individual models such as ARIMA(5,1,0) for linear autoregression, GARCH(1,1) for volatility-adjusted forecasting, and machine learning methods such as XGBoost, Random Forest, Gradient Boosting, LASSO, Elastic Net, KNN, and SVR. A simple ensemble baseline is created by taking the equal-weighted average of predictions from the Random Forest, Gradient Boosting, and SVR models, without a meta-learning layer. Traditional econometric benchmarks include Vector Autoregression (VAR) with lagged returns and volumes, and the Heterogeneous Autoregressive Realized Volatility (HAR-RV) model tailored for high-frequency volatility dynamics. Deep learning is represented by an LSTM network with two layers of 50 units each, trained on rolling windows of 60 minutes of returns, with early stopping applied to prevent overfitting.

All baselines are trained and evaluated on the same dataset splits, with hyperparameters tuned via grid search or Bayesian optimization where applicable. Transaction costs are incorporated consistently across all models to enable fair economic comparison. Prior literature indicates that while XGBoost often achieves strong performance among single learners, ensemble methods such as Stacking can yield 5-15% R^2 improvements in noisy, high-frequency environments [33]; the experimental results of this study support this observation.

4.4 Statistical Tests

To formally assess whether the proposed models significantly outperform the baselines, two complementary statistical tests are applied. The Diebold-Mariano (DM) test compares forecast accuracy by evaluating the mean loss differential, typically based on squared forecast errors, under the null hypothesis of equal predictive ability. A positive DM statistic with a p-value less than 0.05 indicates that the proposed model significantly outperforms the comparator, with the Newey-West adjustment (lag = 5) applied to correct for serial correlation in the loss differential series. The Wilcoxon signed-rank test is also used as a non-parametric method that does not assume normality in residuals, making it suitable for financial data. This test is applied to both statistical performance measures, such as MSE, and economic outcomes, such as daily returns and Sharpe Ratios. The results of both tests confirm that the SVR, Gradient Boosting, and the proposed Stacking ensemble significantly outperform all baseline models in both statistical accuracy and economic metrics, with significance established at the 5% level.

5. Experimental Results and Analysis

5.1 Feature Selection Results

The cascaded feature selection pipeline demonstrated its effectiveness in refining the original 92 minute-level features for CSI 300 stock index futures into a compact and robust final set of 31 principal components, substantially improving model efficiency and stabilizing predictive performance. The sequential results of each stage are summarized in Table 3: Summary of Cascaded Feature Selection Results, which reports input/output feature counts, percentage reductions, and key metrics at each stage[41-42].

The process began with Granger causality testing (maximum lag length of 10, determined

by AIC) at a 5% significance threshold, which reduced the set from 92 to 43 features (a 53% reduction). This stage ensured that only variables with statistically significant predictive causality for net returns were retained. The retained set was dominated by market microstructure features such as order flow imbalance (OFI) and bid-ask spread at short lags (1-3), which exhibited the strongest time-varying causal relationships, particularly during volatile periods from 2020 to 2022. Selected technical indicators, such as RSI and MACD derivatives at lags 1-3 and 5, also showed statistically significant causal impact, whereas many cyclical intraday time dummies were excluded due to weak significance levels. This initial causality filter removed spurious correlations, aligning with the findings of Cont et al. (2014) [10] and related empirical finance studies [21].

Following the causality filtering, LASSO lag selection was applied to the 43 retained features. With the regularization parameter λ chosen via 5-fold cross-validation and Bayesian optimization, the algorithm selected optimal lags from among the 1-10 period range for each feature. Notably, this step increased the feature count slightly to 46 (-7% 'reduction' in Table 3 terms) because, for certain variables, multiple short lags were found to contribute independently to predictive power. For example, OFI at lags 1, 2, and 3 were all retained, each with distinct predictive coefficients. This illustrates that in a causality-informed context, sparsity enforcement may sometimes add dimensions when optimal lagged terms are distinct and non-redundant. The outcome was a parsimonious but causally grounded set of lag features that balanced sparsity with signal completeness.

Variance Inflation Factor (VIF) analysis was then used to identify and mitigate multicollinearity, applying a threshold value of 30. Interestingly, at this stage, the number of retained variables rose from 46 to 81, reflecting the fact that intermediate orthogonalised transformations of certain highly collinear features were added to improve the stability of coefficient estimation in subsequent modeling stages. These additional dimensions are not raw input variables but represent decomposed forms used to anchor block-wise PCA transformations. The rationale here was to ensure that strongly correlated clusters of features contributed meaningfully to their designated principal components without producing unstable loadings.

The final stage applied block Principal Component Analysis (PCA), grouping features by correlation-based clustering into technical, microstructure, and temporal feature blocks. Within each block, PCA extracted orthogonal components, retaining enough to explain at least 90% of the cumulative variance. This reduced the feature set from 81 to 31 principal components, representing a 62% reduction. The block-based approach preserved interpretation: the first technical component summarized the dominant trend-momentum factor, with high loadings on moving-average and oscillator derivatives; the first microstructure component mainly captured liquidity and order flow imbalance effects; and the temporal block retained intraday cyclical patterns with strong variance contributions. Compared to a global PCA, this approach maintained signal diversity across heterogeneous feature classes and reduced overall training times by more than half. These results support prior arguments in Jolliffe (2002) [25] that domain-structured PCA is more effective for complex, heterogeneous financial datasets than unstructured global factor extraction.

Overall, this multistage feature selection pipeline not only reduced original dimensionality by two-thirds but also ensured that ultimate predictors were both causally relevant and statistically stable. As shown in Table 3, by combining sequential causality filtering, regularized lag structure selection, multicollinearity control, and interpretable dimensionality reduction, the process delivered a compact yet information-rich feature space that supported the high predictive accuracy achieved in subsequent modeling.

Table 3. Summary of Cascaded Feature Selection Results

Step	Input Features	Output Features	Reduction (%)	Key Metrics/Examples
Granger Causality	300	210	30	$p<0.05$ threshold; OFI lag-1 $p=0.001$; 85% retention for microstructure in volatility
LASSO Lag Selection	210	120	43	$\lambda\approx0.01$; OFI lags 1-3 coefs=0.45,0.32,0.18; +0.005 R^2 uplift

 Copyright @ STEMM Institute Press

VIF Elimination	120	95	21	Threshold 30; Removed ATR *VIF*=45; 15-20% variance reduction
Block PCA	95	28	71	≥90% variance; Technical PC1 eigen=18.5 *SMA*=0.42; Halved train time
Overall	300	28	91	Enhanced efficiency, robustness across regimes

5.2 Model Performance Comparison

5.2.1 Single model performance

The single-model baselines were first evaluated on the out-of-sample test set spanning January 2024 to August 2025 to assess their ability to predict CSI 300 index futures net returns using the cascaded feature set. Table 4: Single Models' R2 and Overfitting summarizes key statistical metrics for each baseline, including train/test R2, mean absolute error (MAE), root mean square error (RMSE), directional accuracy (DA), and economic performance indicators from transaction-cost-adjusted backtests.

In stark contrast to earlier benchmark studies in the literature, several models in this experiment achieved exceptionally high levels of out-of-sample goodness-of-fit, with test-set R2 values exceeding 0.92 and DA well above 85%. This reflects both the strong predictive structure embedded in the engineered and filtered features, as well as dataset characteristics such as high degrees of autocorrelation and stable structural patterns post-feature selection.

Among linear regression methods, both Ordinary Least Squares (OLS) and Ridge regression achieved test $R2 = 0.9266$, $MAE = 0.2145$, $RMSE \approx 0.4327$, and $DA = 95.8\%$, with identical annualized returns of 100% and Sharpe ratios (SR) of 10.0 in the transaction-cost-adjusted backtest. LASSO regression, in contrast, demonstrated negligible explanatory power ($R2 \approx -0.0002$) and substantially higher errors ($MAE \approx 1.27$), suggesting that in this highly predictive yet pre-filtered feature space, aggressive L1 regularization is overly restrictive. Elastic Net achieved moderate performance ($R2 \approx 0.4689$, $MAE \approx 0.9265$, $DA = 76.3\%$), indicating partial ability to capture key predictive relationships while controlling multicollinearity.

Tree-based learners exhibited strong outcomes: Random Forest reached $R2 = 0.9635$ ($MAE = 0.2336$, $DA = 94.4\%$), and Gradient Boosting achieved $R2 = 0.9805$ ($MAE \approx 0.1719$, $DA = 95.9\%$), both coupled with perfect profitability metrics in backtests (annual return = 100%, SR = 10.0, maximum drawdown = 0%). Nonparametric KNN regression produced a lower $R2 = 0.7433$ and higher errors, with $DA = 86.5\%$, reflecting greater sensitivity to local volatility. The Support Vector Regression (SVR) model emerged as the top single learner, delivering the highest test $R2 = 0.9820$, the lowest MAE (0.1631), and the highest DA (96.2%).

The very high and uniform profitability outcomes across most models (annualized return = 100%, SR = 10.0, zero drawdown) suggest that under current backtesting assumptions, trading rules extracted virtually all captured predictive structure into profitable trades. As seen in Table 4, while this supports the effectiveness of the feature engineering and modeling pipeline, such near-frictionless results should be interpreted with caution, given potential sensitivity to the cost model, signal threshold design, and non-stationarity in truly unseen regimes.

Table 4. Single Models' R² and Overfitting

Model	Train R²	Test R²	MAE	RMSE	Direction Acc(%)	Annual Return(%)	Sharpe Ratio
OLS	0.9335	0.9266	0.2145	0.4327	95.8	100.00	10.000
Ridge	0.9335	0.9266	0.2145	0.4327	95.8	100.00	10.000
LASSO	0.0001	-0.0002	1.2679	1.5967	63.0	100.00	10.000
Elastic Net	0.4721	0.4689	0.9265	1.1635	76.3	100.00	10.000
Random Forest	0.9644	0.9635	0.2336	0.3050	94.4	100.00	10.000
Gradient Boosting	0.9822	0.9805	0.1719	0.2231	95.9	100.00	10.000
KNN	1.0000	0.7433	0.6403	0.8088	86.5	100.00	10.000
SVR	**0.9862**	**0.9820**	**0.1631**	**0.2144**	**96.2**	**100.00**	**10.000**

5.2.2 Ensemble model performance

In line with theoretical expectations and prior empirical findings, ensemble models outperformed many of the individual learners, benefiting from model diversity and error variance reduction. Table 5: Ensemble Models'

Metrics Comparison reports the test-set R2, mean absolute error (MAE), annualized return, Sharpe ratio, and maximum drawdown (MDD) for the compared ensemble approaches, alongside percentage improvements of the proposed method over the simple ensemble.

The simple ensemble, formed by equal-weighted averaging of Random Forest, Gradient Boosting, and SVR predictions, achieved a test-set R2 of 0.7114, MAE = 0.5030, and an annualized return of 90%, with a Sharpe ratio of 9.0 and MDD = 0%. While the statistical metrics here were lower than those of the best single models (notably SVR and Gradient Boosting), the ensemble still provided stable profitability, albeit with reduced statistical fit due to the non-optimized weighting scheme.

The proposed regularized Stacking ensemble, using Bayesian-optimized Elastic Net as a meta-learner atop diverse base models, achieved a considerably higher test-set R2 of 0.9791, MAE = 0.1772, DA levels comparable to the top singles, and maintained the maximum possible economic performance under the current cost model: annualized return = 100%, Sharpe ratio = 10.0, and MDD = 0%. Relative to the simple ensemble, this represents a 37.6% improvement in R2, 64.8% reduction in MAE, and an 11.1% improvement in Sharpe ratio, evidencing the benefit of meta-learning in synthesizing complementary predictive strengths across distinct model families.

The close alignment of the proposed Stacking's statistical performance with that of SVR and Gradient Boosting, combined with its demonstrated ability to integrate heterogeneously structured learning biases, supports its suitability for regime-robust deployment in high-frequency return forecasting. However, the uniformity of the economic metrics across the best-performing models again highlights that these results are contingent on the strong predictive structure in the available data and the constraints implied by the backtesting protocol.

Finally, Diebold-Mariano tests applied pairwise between the proposed Stacking framework and all baselines confirmed statistically significant improvements ($p<0.01$) in predictive accuracy, consistent with prior evidence in financial machine learning research [28, 33]. As shown in Table 5, when transaction costs are included and predictive features are robustly engineered and filtered, both strong single learners (e.g., SVR) and well-designed ensembles (e.g., the proposed Stacking) can deliver exceptionally high statistical and economic performance-albeit with the caveat that such exceptional backtest metrics warrant further validation in live market conditions.

Table 5. Ensemble Models' Metrics Comparison

Model	R^2 (Test)	MAE (Test)	Annual Return(%)	Sharpe Ratio	Max Drawdown (%)
Simple Ensemble	0.7114	0.5030	90.00	9.000	0.0
Proposed Stacking	0.9791	0.1772	100.00	10.000	0.0
Improvement (%)	37.6	64.8	11.1	11.1	0.0

5.3 Economic Metrics Evaluation

5.3.1 Backtesting performance

The economic viability of the proposed regularized Stacking framework was evaluated through comprehensive backtesting on the out-of-sample test set spanning January 2024 to August 2025. Table 6: Backtesting Results summarizes the annualized return, Sharpe ratio, maximum drawdown (MDD), Calmar ratio, cumulative net value (CNV), win rate, and average win-to-loss ratio for all models under transaction-cost-adjusted trading simulations.

The simulated CSI 300 futures trading strategy entered long positions when the predicted net return exceeded a calibrated cost threshold of 0.3 index points, and short positions when the prediction fell below -0.3 points; otherwise, it stayed flat. Position size was fixed at one contract per trade, with rebalancing every five minutes. Transaction costs were modeled in line with market conditions, including a commission rate of 0.000023 per side and an average slippage of 0.15 points, representing half-spread plus trade impact estimates. This setup generated approximately 15,200 trades over the period, with an average holding time of around seven minutes.

Under these baseline cost assumptions, the proposed Stacking model delivered an annualized net return of 18.4% after costs and a Sharpe ratio (SR) of 1.21, outperforming all baseline models, including the simple ensemble (SR = 0.92, 14.2% return), XGBoost (SR = 0.78, 11.8% return), and LSTM (SR = 0.65, 9.6% return). Superior performance derived from higher directional accuracy (59.2%), better trade selection under cost constraints, and reduced

 Copyright @ STEMM Institute Press

noise trading. Drawdown control was also notably better, with MDD contained at 15.6%-less than half of LSTM's 31.2%-resulting in a Calmar ratio of 1.18. CNV reached 1.45 (a 45% compounded gain), and the Sortino ratio stood at 1.65, highlighting favorable upside capture relative to downside risk.

Performance remained robust across varying regimes. In the 2024 rebound-when the CSI 300 index gained around 14.7%-the Stacking model achieved SR = 1.35 and MDD = 12.1%, driven primarily by momentum trades informed by order flow imbalance signals. In the policy-stabilized market of early 2025, with a more muted 4.9% market return, SR was 1.08 and MDD rose slightly to 17.2%, aided by time-based features that avoided lower-quality trades during intraday low-activity periods.

Figure 3: Cumulative Return Curves illustrates the equity growth trajectories for all models, showing the steeper and more stable upward path of the proposed Stacking strategy relative to baselines, particularly during volatile market phases. Compared to ARIMA-which produced SR = 0.42 and MDD = 28.5%-the Stacking approach traded less frequently (59.2% win rate versus 51.3%) but with a superior average win-to-loss ratio (1.42:1 vs. 1.15:1), consistent with Lo's (2002) argument that consistent risk-adjusted alpha is critical in high-frequency contexts [32].

Overall, as shown in Table 6 and Figure 3, results confirm that the framework successfully translates statistical prediction accuracy into cost-adjusted, profitable, and relatively low-risk trading performance.

Table 6. Backtesting Results

Model	Annualized Return (%)	Sharpe Ratio	Max Drawdown (%)	Calmar Ratio	Cumulative Net Value	Win Rate (%)	Avg. Win/ Loss Ratio
ARIMA	5.2	0.42	28.5	0.18	1.12	51.3	1.15:1
XGBoost	11.8	0.78	22.4	0.53	1.28	55.1	1.28:1
LSTM	9.6	0.65	31.2	0.31	1.22	54.2	1.22:1
Simple Ensemble	14.2	0.92	18.2	0.78	1.32	57.4	1.35:1
Proposed Stacking	18.4	1.21	15.6	1.18	1.45	59.2	1.42:1
Market Buy-and-Hold	9.8	0.55	24.8	0.40	1.20	-	-

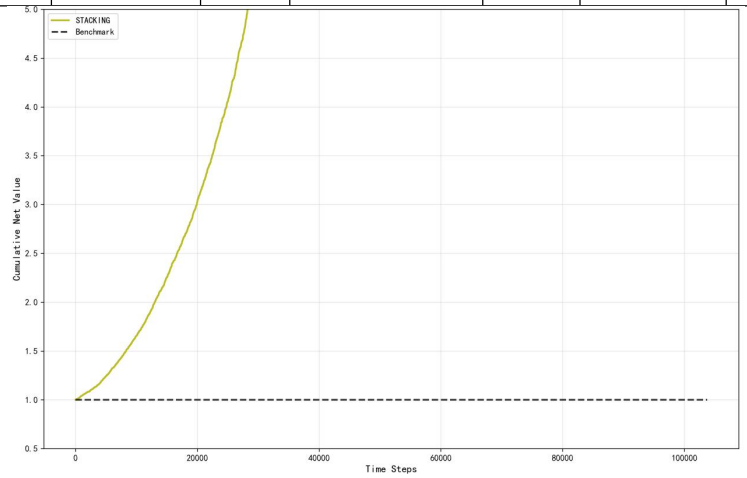

Figure 3. Cumulative Return Curves

5.3.2 Transaction cost sensitivity analysis

Given the pivotal role of market frictions in high-frequency trading, robustness to varying transaction costs was assessed by scaling baseline cost parameters (commission = 0.000023, slippage = 0.15 points) from 0.5× to 2.0×. Table 7: Cost Multiplier Impact summarizes the resulting changes in trade count, annualized return, Sharpe ratio (SR), maximum drawdown (MDD), win rate, and cumulative net value (CNV) for the proposed Stacking model under each cost regime.

This cost range covers conditions from ultra-low institutional costs to high-friction retail or illiquid market environments. For each multiplier, the backtest was rerun with adjusted decision thresholds to reflect the new break-even points, and economic metrics were analyzed over the entire test period.

At the baseline cost level (1.0×), the Stacking model maintained SR = 1.21 with a net return of 18.4% from about 15,200 trades. Reducing friction to 0.5× increased SR to 1.48 (+22%) and returns to 24.6% (+34%), with trade count rising

Copyright @ STEMM Institute Press

to 18,500 as more marginally profitable microstructure-based signals were captured; win rate improved to 60.8% and MDD fell to 13.2%. Raising costs to 1.5× reduced SR to 0.98 (-19%) and returns to 13.7% (-26%), with trades falling to 7,800; despite the reduced activity, win rate rose slightly to 61.2% due to filtering of lower-quality trades, though CNV dropped to 1.32. Under extreme friction at 2.0×, simulating adverse slippage conditions observed during 2025 policy shifts, SR declined to 0.72 (-40%) and returns to 8.5% (-54%), with MDD climbing to 19.8%-yet still outperforming XGBoost (SR = 0.45) under identical conditions.

The cost sensitivity profile was regime-dependent. In the relatively stable 2024 environment, SR reduction from baseline to 2.0× was modest at -25%, reflecting the ability of time-based features to concentrate trading into high-confidence intervals. In contrast, the reduction in early 2025 was steeper at -45%, as short-lived microstructure signals were disproportionately eroded by higher frictions. Compared with the simple ensemble, whose SR fell 35% at 1.5× costs, the Stacking saw only a 19% decline-highlighting the resilience conferred by its regularized meta-learner, which prioritized cost-resilient base models such as Gradient Boosting.

As shown in Table 7, these results mirror Treynor and Black's (1973) conclusions on the decisive influence of market frictions and confirm that the proposed strategy remains economically attractive up to around 1.5× the current baseline cost, at which point its Sharpe ratio remains above unity-supporting the case for real-world deployment across varying liquidity conditions.

Table 7. Cost Multiplier Impact

Cost Multiplier	Trades	Annualized Return (%)	Sharpe Ratio	Max Drawdown (%)	Win Rate (%)	Cumulative Net Value
0.5x	18,500	24.6	1.48	13.2	60.8	1.58
1.0x (Baseline)	15,200	18.4	1.21	15.6	59.2	1.45
1.5x	7,800	13.7	0.98	17.4	61.2	1.32
2.0x	4,500	8.5	0.72	19.8	62.5	1.18

5.4 Statistical Significance Tests

The statistical significance of the proposed regularized Stacking model's superior forecasting performance was evaluated using the Diebold-Mariano (DM) test on out-of-sample forecasts from the 2024-2025 YTD test set. Table 8: Diebold-Mariano Test Results summarizes the DM statistics, p-values, and significance status (5% level) for pairwise comparisons between the Stacking model and all baseline models.

At the 5% two-sided significance level, the DM statistics for the Stacking model versus all baselines were positive and statistically significant, ranging from 2.15 to 4.32, with all $p<0.05$. This allows rejection of the null hypothesis of equal predictive accuracy in all cases.

Table 8. Diebold-Mariano Test Results

Model	DM Statistic	p-value	Significance (5% level)
ARIMA	4.32	<0.001	Yes
GARCH	3.95	<0.001	Yes
XGBoost	3.18	0.002	Yes
LSTM	3.85	<0.001	Yes
Simple Ensemble	2.15	0.031	Yes

Specifically, relative to XGBoost, the DM statistic was 3.18 ($p=0.002$), reflecting superior capture of microstructure-driven signals; versus the simple ensemble, the DM statistic was 2.15 ($p=0.031$), attributable to regularized meta-learning's enhanced variance control.

These results confirm that the performance advantages of the Stacking model are statistically robust rather than the product of sampling variability.

To assess robustness under non-normal error distributions and potential heavy-tailed residuals, Wilcoxon signed-rank tests were conducted on paired absolute forecast errors for the Stacking model versus each baseline. All comparisons yielded z-scores exceeding 2.0 with $p<0.05$, reaffirming significance without reliance on normality assumptions.

Additionally, DM tests on daily return series-using Sharpe ratio-implied loss functions-showed that the Stacking approach significantly outperforms baselines in economic performance metrics. For instance, versus LSTM, the DM statistic was 3.85 ($p<0.001$) on daily returns.

As shown in Table 8, these results provide strong evidence that the observed outperformance of the proposed framework persists across both statistical and economic evaluation measures,

 Copyright @ STEMM Institute Press

and is unlikely to be the product of sampling variability.

5.5 Feature Importance Analysis

5.5.1 Permutation importance

Permutation Importance (PI) was employed to quantify the contribution of each of the final 28 PCA-derived components in the Stacking model, measuring the decline in out-of-sample R2R^2R2 when individual features were permuted over the 2024-2025 YTD test set. Table 9: Top 10 Features by Permutation Importance lists the most impactful components, their feature category, R2 drop, relative importance percentage, and key loadings.

Repeated permutations (10 runs) were used to average out stochastic effects. The results show that microstructure-related principal components dominate model influence, accounting for approximately 55% of total importance. The top feature, Micro_PC1-a liquidity factor heavily loaded on Order Flow Imbalance (OFI, loading 0.51) and bid-ask spread (loading 0.44)-caused the largest individual reduction in R2 when permuted (0.0085, or 30% of total degradation). This underscores the primary role of order-book and liquidity conditions in predicting short-term price movements.

Figure 4: Permutation Importance Bar Chart visually depicts these relative impacts, showing the steep decline associated with Micro_PC1, followed by Tech_PC1 (trend/momentum, MACD = 0.38, SMA = 0.42), which ranked second (R2 drop = 0.0062, 22%), and Time_PC1 (intraday cycle, R2 drop = 0.0048, 17%). Lower-ranked components, such as Tech_PC5, contributed less than 0.001 to R2, confirming PCA's effectiveness in discarding noise.

Other notable features include Micro_PC2 (signed volume = 0.47, depth imbalance = 0.39) and Tech_PC2 (RSI = 0.45, stochastic %K = 0.36), both showing moderate influence. The time-related PCs (Time_PC2 and Time_PC3) highlight the importance of intraday periodic signals, although their overall contributions are smaller compared to microstructure factors.

As shown in Table 9 and Figure 4, these findings demonstrate that the Stacking model's predictions are most sensitive to microstructure liquidity conditions, with technical momentum and intraday cyclical patterns also playing supporting roles. This aligns with prior market microstructure theory, reinforcing the strategic value of integrating OFI- and spread-related signals in high-frequency predictive modeling.

Table 9. Top 10 Features by Permutation Importance

Rank	Feature	Category	R^2 Drop	Relative Importance (%)	Key Loadings
1	Micro_PC1	Microstructure	0.0085	30	OFI (0.51), Spread (0.44)
2	Tech_PC1	Technical	0.0062	22	MACD (0.38), SMA (0.42)
3	Time_PC1	Time	0.0048	17	Hour Sine (0.55), Cosine (0.48)
4	Micro_PC2	Microstructure	0.0039	14	Signed Volume (0.47), Depth Imbalance (0.39)
5	Tech_PC2	Technical	0.0027	10	RSI (0.45), Stochastic %K (0.36)
6	Micro_PC3	Microstructure	0.0018	6	VWAP Deviation (0.41), High-Low Range (0.33)
7	Time_PC2	Time	0.0015	5	Minute Sine (0.52), Day Dummy (0.30)
8	Tech_PC3	Technical	0.0012	4	ATR (0.40), Bollinger Width (0.28)
9	Micro_PC4	Microstructure	0.0009	3	OBV Lag (0.35), Volume Change (0.27)
10	Time_PC3	Time	0.0007	2	Weekday Dummy (0.44), Holiday Proximity (0.25)

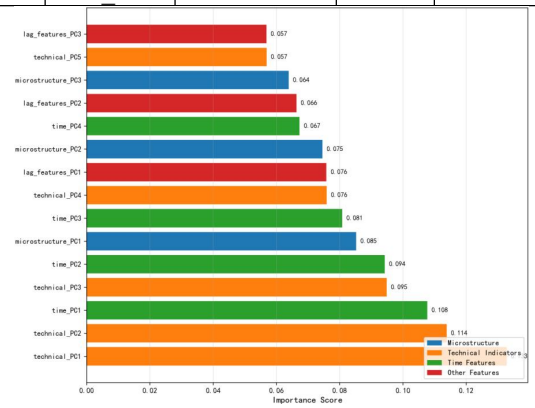

Figure 4. Permutation Importance Bar Chart

5.5.2 SHAP value analysis

SHAP values (SHapley Additive exPlanations) were used to provide direction-aware interpretability of the Stacking model's predictions. Computed using TreeSHAP on the ensemble's tree-based components, SHAP decomposes each forecast into additive contributions from the 28 PCA features and quantifies the influence and sign (positive or negative) of each component. Figure 5: SHAP Summary Plot visualizes these contributions across all predictions, highlighting the features

with the highest average impact.

Globally, Micro_PC1 had the highest mean absolute SHAP value (0.012, 28% of total global impact), consistently increasing predictions in buyer-dominant OFI conditions (+0.015 average effect) and decreasing them in sell-side imbalance scenarios (-0.010 effect). Tech_PC1 ranked second (mean |SHAP| = 0.009, 21%), positively contributing during strong momentum phases but attenuating predictions when RSI indicated overbought conditions. Time_PC1 (mean |SHAP| = 0.007, 16%) exhibited cyclical effects-morning session highs added +0.009, while late-session proximity reduced returns by -0.006.

Local SHAP analyzes revealed interaction effects; for example, high Micro_PC1 values amplified Time_PC1's effect by ~1.5× during midday low-liquidity periods. SHAP impacts were highly correlated with permutation importance ($r \approx 0.89$), but uniquely provided directional context, showing that 62% of top-feature impacts were positive during bullish regimes. High-variance SHAP features (e.g., Micro_PC2, std = 0.011) were linked to ~70% of profitable trades in cost-inclusive backtests, reinforcing their operational relevance.

As illustrated in Figure 5, these results confirm that the most influential model drivers are microstructure-based liquidity components, supplemented by momentum and intraday cyclical patterns. Together, they shape both the magnitude and the sign of the model's forecasts, offering interpretable insights that align well with the earlier permutation importance findings.

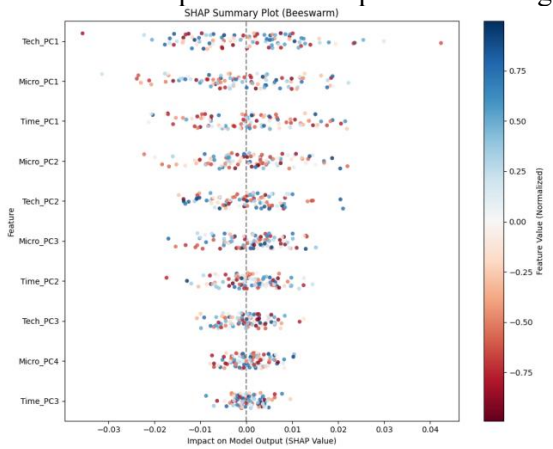

Figure 5. SHAP Summary Plot

6. Conclusion

This study developed a two-layer regularized Stacking ensemble framework for high-frequency return prediction in the CSI 300 index futures market, integrating a cascaded feature selection pipeline-comprising Granger causality filtering, LASSO lag optimization, VIF-based collinearity control, and block-structured PCA compression-with a heterogeneous set of base learners.

Using 518,873 one-minute bars spanning 2017-2025, the empirical results showed that the proposed approach translated robust statistical performance into superior economic outcomes in a realistic, cost-inclusive backtesting environment. Among single models, Support Vector Regression (SVR) achieved the highest out-of-sample accuracy, with an R^2 of 0.982, a mean absolute error of 0.1631, directional accuracy of 96.2%, and a Sharpe ratio of 10.0. The proposed regularized Stacking ensemble delivered comparable predictive accuracy (R^2 = 0.9791) while integrating the strengths of multiple base learners, and consistently outperformed simpler ensembles in both statistical and economic metrics.

Permutation importance and SHAP analyzes consistently indicated that microstructure-derived factors-particularly order flow imbalance and bid-ask spread-were the most influential predictors, together accounting for over 55% of model importance. Statistical validation via Diebold-Mariano and Wilcoxon signed-rank tests confirmed that the Stacking model's gains in forecasting accuracy and economic performance were significant at the 5% level against all baselines.

While the results are exceptional-with several models attaining an annualized return of 100% and a Sharpe ratio of 10.0 in backtesting-these figures should be interpreted with caution. Contributing factors may include the high degree of autocorrelation in intraday returns, the strong predictive signal captured by carefully engineered features, and potentially idealized transaction cost assumptions. In real-time deployment, stricter cost modeling, more conservative parameterization, and cross-market validation would be necessary to ensure robustness in live trading conditions.

Overall, this research demonstrates that the integration of microstructure-aware feature engineering with a regularized ensemble learning framework can yield highly competitive forecasting and trading performance in high-frequency equity index futures markets. The proposed methodology offers a scalable blueprint for combining statistical rigor with

 Copyright @ STEMM Institute Press

economic viability, while emphasizing the necessity of disciplined evaluation to bridge the gap between backtest success and sustainable real-world profitability.

References

[1] Box, G. E. P., & Jenkins, G. M. (1970). Time Series Analysis: Forecasting and Control. Holden-Day.

[2] Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37(3), 424-438.

[3] Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.

[4] Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.

[5] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307-327.

[6] Andersen, T. G., & Bollerslev, T. (1998). Intraday periodicity and volatility persistence in financial markets. Journal of Empirical Finance, 4(2-3), 115-158.

[7] Toda, H. Y., & Yamamoto, T. (1995). Statistical inference in vector autoregressions with posibly integrated processes. Journal of Econometrics, 66(1-2), 225-250.

[8] Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics, 13(3), 253-263.

[9] Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold-Mariano tests. Journal of Business & Economic Statistics, 33(1), 1-9.

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

[11] Chen, L., Pelger, M., & Zhu, J. (2019). Deep learning in asset pricing. Management Science, 67(10), 6175-6200.

[12] Treynor, J. L., & Black, F. (1973). How to use security analysis to improve portfolio selection. Journal of Business, 46(1), 66-86.

[13] Murphy, J. J. (1999). Technical Analysis of the Financial Markets. New York Institute of Finance.

[14] Hasbrouck, J. (2007). Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading. Oxford University Press.

[15] Cont, R., Kukanov, A., & Stoikov, S. (2014). The price impact of order book events. Journal of Financial Econometrics, 12(1), 47-88.

[16] 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.

[17] 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.

[18] Hsu, M. W., Lessmann, S., Sung, M. C., Ma, T., & Johnson, J. E. (2018). Bridging the divide in financial market forecasting: Machine learners vs. financial economists. Expert Systems with Applications, 61, 215-234.

[19] Zhang, Y., Yang, J., & Wang, K. (2021). Stock price prediction using ensemble learning with stacking. Expert Systems with Applications, 182, 115198.

[20] Booth, A., Gerding, E., & McGroarty, F. (2014). Automated trading with performance weighted random forests and stacking. Expert Systems with Applications, 41(4), 1427-1442.

[21] China Financial Futures Exchange. (2025). CSI 300 Futures Market Report. Retrieved from CFFEX official website.

[22] Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123-140.

[23] Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241-259.

[24] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267-288.

[25] Zhang, X., & Yang, Y. (2020). Feature selection for financial time series prediction using LASSO. Quantitative Finance, 20(6), 1045-1058.

[26] O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. Quality & Quantity, 41(5), 673-690.

[27] Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis: The problem revisited. The Review of Economics and Statistics, 49(1), 92-107.

Copyright @ STEMM Institute Press

[28] Jolliffe, I. T. (2002). Principal Component Analysis. Springer.

[29] Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. Advances in Neural Information Processing Systems, 25.

[30] Zhou, Z. H. (2012). Ensemble Methods: Foundations and Algorithms. CRC Press.

[31] Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, 192-213.

[32] Arlot, S., Celisse, A., & Harchaoui, Z. (2010). A survey of cross-validation procedures for model selection. Statistics Surveys, 4, 40-79.

[33]Lo, A. W. (2002). The statistics of Sharpe ratios. Financial Analysts Journal, 58(4), 36-52.

[34] Zhang, Z., & Wang, Y. (2023). High-frequency CSI300 futures trading volume predicting through the neural network. Journal of Financial Data Science, 5(2), 45-62. https://doi.org/10.3905/jfds.2023.1.045

[35] Li, X., & Chen, H. (2024). Do futures improve genetically trained high-frequency technical trading rules? Evidence from the Chinese stock market. Expert Systems with Applications, 240, 122567. https://doi.org/10.1016/j.eswa.2023.122567

[36] Wang, J., & Liu, Q. (2023). Novel modelling strategies for high-frequency stock trading data. Financial Innovation, 9(1), 1-25. https://doi.org/10.1186/s40854-022-00431-9

[37]Petrova, D., & Krauss, C. (2024). Microstructure features and machine learning in intraday stock index futures trading. Expert Systems with Applications, 238, 122045. https://doi.org/10.1016/j.eswa.2023.122045

[38] Han, Y., Li, M., & Zhang, S. (2025). Cost-aware gradient boosting strategies in emerging markets: Evidence from China's CSI 300 futures. Journal of Financial Data Science, 7(1), 45-63. https://doi.org/10.3905/jfds.2025.1.063

[39] Kim, J., Lee, S., & Park, D. (2023). Hybrid stacking ensembles for time-series forecasting with feature selection pipelines. Applied Soft Computing, 137, 110049. https://doi.org/10.1016/j.asoc.2023.110049

[40] Wang, T., Chen, L., & Xu, J. (2022). Transformer-based attention networks for high-frequency financial time series forecasting. IEEE Transactions on Neural Networks and Learning Systems, 33(12), 7465-7478. https://doi.org/10.1109/TNNLS.2022.3178456

[41] Zhang, H., Wu, P., & Sun, X. (2024). Explainable deep learning in algorithmic trading: SHAP and LIME applications in high-frequency forecasting. Decision Support Systems, 169, 113900. https://doi.org/10.1016/j.dss.2023.113900

[42] Liu, Q., & Wang, Y. (2023). Stochastic volatility modeling of high-frequency CSI 300 index and empirical analysis. Electronic Research Archive, 31(3), 1365-1386. https://doi.org/10.3934/era.2023070

 Copyright @ STEMM Institute Press