International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 8, Number 6, Year 2026
DOI: https://doi.org/10.62051/ijcsit.v8n6.08
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Information-Driven Resampling and Market Regime Detection: A Futures Trading Framework Based on GMM and Multi-Model Ensemble Learning

Zhankun Wang *

Hangzhou Dianzi University, Hangzhou, 310018, China
*Corresponding Author: wangzhankun0826@gmail.com

ABSTRACT

Fixed-interval candlestick sampling cannot adequately represent the non-uniform arrival of information in futures markets. This study develops an information-driven market-regime detection and trading framework. One-minute data are resampled along a hybrid information axis constructed from standardized trading-volume intensity, realized volatility, price momentum, and a high-low spread proxy. A GMM-HMM identifies latent regimes, after which XGBoost and CatBoost are trained on 400-dimensional lagged-price features; trades are executed only when both models agree. Using rebar futures data from 2010 to 2025 and reserving 2023-2025 for out-of-sample testing, the complete strategy achieves an annualized return of 11.33%, a maximum drawdown of -7.80%, and a Sharpe ratio of 0.88, outperforming conventional sampling, single-model, and no-regime-filtering alternatives in ablation tests. The results show that combining information-driven resampling with regime-aware model fusion improves trade selectivity and risk-adjusted performance.

KEYWORDS

Ensemble learning; Gaussian-mixture hidden Markov model; Information-driven resampling; Quantitative futures trading

1. INTRODUCTION

China's commodity futures market features active trading and dense intraday data, yet market information does not arrive uniformly. Fixed-time candlesticks can smooth meaningful variation during active periods and accumulate redundant observations during quiet periods, making them ill suited to markets characterized by volatility clustering, nonlinear dynamics, and regime switching.

Research on information clocks provides a basis for addressing this problem. Mandelbrot and Taylor [1] linked price variation to trading activity, while Easley et al. [2] further clarified the relationship between trading activity and price formation. De Prado [3] systematized event-driven bars, and Sokolovsky et al. [4] and Zhang [5] showed that joint representations of price and volume can improve the quality of machine-learning inputs.

Market-regime detection and ensemble learning form another research stream. Hamilton [6] established the regime-switching framework, and Liu [7] demonstrated the value of Markov methods for short-horizon volatility prediction. XGBoost, introduced by Chen and Guestrin [8], is well suited to high-dimensional financial features, while Gupta et al. [9] showed that regime-aware model fusion can improve trading performance. However, few studies jointly integrate event-driven sampling, regime detection, and ensemble decision-making.

To address this gap, this study develops a three-stage framework integrating information-driven resampling, GMM-HMM regime detection, and multi-model consensus decision-making, and evaluates it using rebar futures data from 2010 to 2025. Resampling increases the information density of observations, the GMM-HMM provides macro-regime filtering, and XGBoost and CatBoost use lagged-price features for directional classification. A trade is executed only when the two models agree. The 2023-2025 period is strictly reserved for out-of-sample backtesting.

2. RESEARCH DESIGN

Using the rqdatac API, this study obtains one-minute data for the main continuous rebar futures contract (RB99) from 2010 to 2025. The fields include open, high, low, close, and trading volume, and the sample spans multiple complete market cycles.

2.1. Hybrid-Axis Volatility-Threshold Resampling Method

The resampling method standardizes trading-volume intensity, realized volatility, price momentum, and a high-low spread proxy, maps them onto a hybrid information axis, and creates a new candlestick only when the accumulated information reaches a threshold.

The four standardized indicators on the hybrid information axis are defined as follows:

(1) Trading-volume intensity:

$$VI_t = \frac{Vol_t - \mu_{Vol}}{\sigma_{Vol}} \tag{3-1}$$

Where μVol and σVol denote the five-period rolling mean and standard deviation of trading volume, respectively; σVol is replaced by 10^{-10} when it equals zero.

(2) Volatility: the true range and standardized ATR are defined as:

$$TR_t = \max(TR_1, TR_2, TR_3) \tag{3-2}$$

$$TR_1 = High_t - Low_t$$
$$TR_2 = |High_t - Close_{t-1}|$$
$$TR_3 = |Low_t - Close_{t-1}|$$

Where $High_t$, Low_t, and $Close_{t-1}$ denote the current high, current low, and previous-period close, respectively.

$$ATR_t = \text{rolling}_{\text{mean}}(TR_t, window = 14) \tag{3-3}$$

$$NormalizedATR_t = \frac{ATR_t - \mu_{ATR}}{\sigma_{ATR}}$$

Where μATR and σATR denote the 14-period rolling mean and standard deviation of ATR, respectively.

(3) Momentum:

$$Momentum_t = P_t - P_{t-14} \tag{3-4}$$

$$NormalizedMomentum_t = \frac{Momentum_t - \mu_{Mom}}{\sigma_{Mom}}$$

Where momentum is the difference between the current close and the close 14 periods earlier, standardized by its 14-period rolling mean and standard deviation.

(4) Market microstructure proxy:

$$Spread_t = \frac{High_t - Low_t}{Close_t} \tag{3-5}$$

$$NormalizedSpread_t = \frac{Spread_t - \mu_{Spread}}{\sigma_{Spread}}$$

Where $High_t$, Low_t, and $Close_t$ denote the current high, low, and close, respectively.

The system processes the raw one-minute candlesticks chronologically, stores them in a buffer, and accumulates the absolute values of the four standardized indicators. Once the hybrid information measure reaches the threshold, the buffered interval is aggregated into a new candlestick with a variable time span.

The cumulative information in the buffer is:

$$\begin{aligned} VI_{sum} &= \sum |VI_t| \\ ATR_{sum} &= \sum |NormalizedATR_t| \\ Mom_{sum} &= \sum |NormalizedMomentum_t| \\ Micro_{sum} &= \sum |NormalizedSpread_t| \end{aligned} \tag{3-6}$$

Resampling is triggered when the weighted information measure reaches threshold T:

$$(V_{sum} \cdot w_V + ATR_{sum} \cdot w_{ATR} + Mom_{sum} \cdot w_{Mom} + Micro_{sum} \cdot w_{Micro}) \geq T \tag{3-7}$$

Where the four weights are set to 0.25 by default and may be varied in robustness tests.

Upon triggering, Open is the opening price of the first buffered bar; High and Low are the interval extrema; Close is the closing price of the last buffered bar; and Volume is the sum of interval volume. The buffer is then cleared and accumulation restarts.

2.1.1. Statistical Properties of the Resampled Data

Figure 1 compares kernel density estimates (KDEs) of logarithmic returns from the original time bars and the weighted-threshold resampled bars.

The original returns exhibit pronounced leptokurtosis and fat tails, whereas the resampled distribution is closer to the standard normal distribution. This indicates that information-driven sampling reduces time noise and improves the statistical properties of model inputs.

2.2. Unsupervised Market-Regime Detection with GMM

The GMM-HMM uses current volatility, the L-lag logarithmic return, and the first-order logarithmic return as observations to identify latent market regimes across different time scales.

$$HLdiff = ln(High) - ln(Low) \tag{3-8}$$

Where High and Low denote the high and low of the resampled candlestick, respectively.

Medium-term price momentum is measured by the L-lag logarithmic return:

$$logreturnX = ln(P_t) - ln(P_{t-L}) \tag{3-9}$$

Where P_t and P_{t-L} are the closing prices of the current and L-period-lagged resampled candlesticks.

Short-term price movement is measured by the first-order logarithmic return:

$$logreturn = ln(P_t) - ln(P_{t-1}) \tag{3-10}$$

Where P_{t-1} is the closing price of the preceding resampled candlestick.

The model specifies two hidden states, each represented by three Gaussian-mixture components. Parameters are estimated with the Baum-Welch algorithm, and the most probable state sequence is obtained with the Viterbi algorithm. Based on cumulative returns, the higher-return state is labeled 1 and the lower-return state is labeled 0; no regime boundaries are manually imposed.

Figure 2 shows that the GMM-HMM state sequence corresponds closely to the price trend of the underlying asset.

State 1 clusters in rising intervals, whereas state 0 occurs more often in declining or weak intervals, indicating that the model captures persistent price regimes.

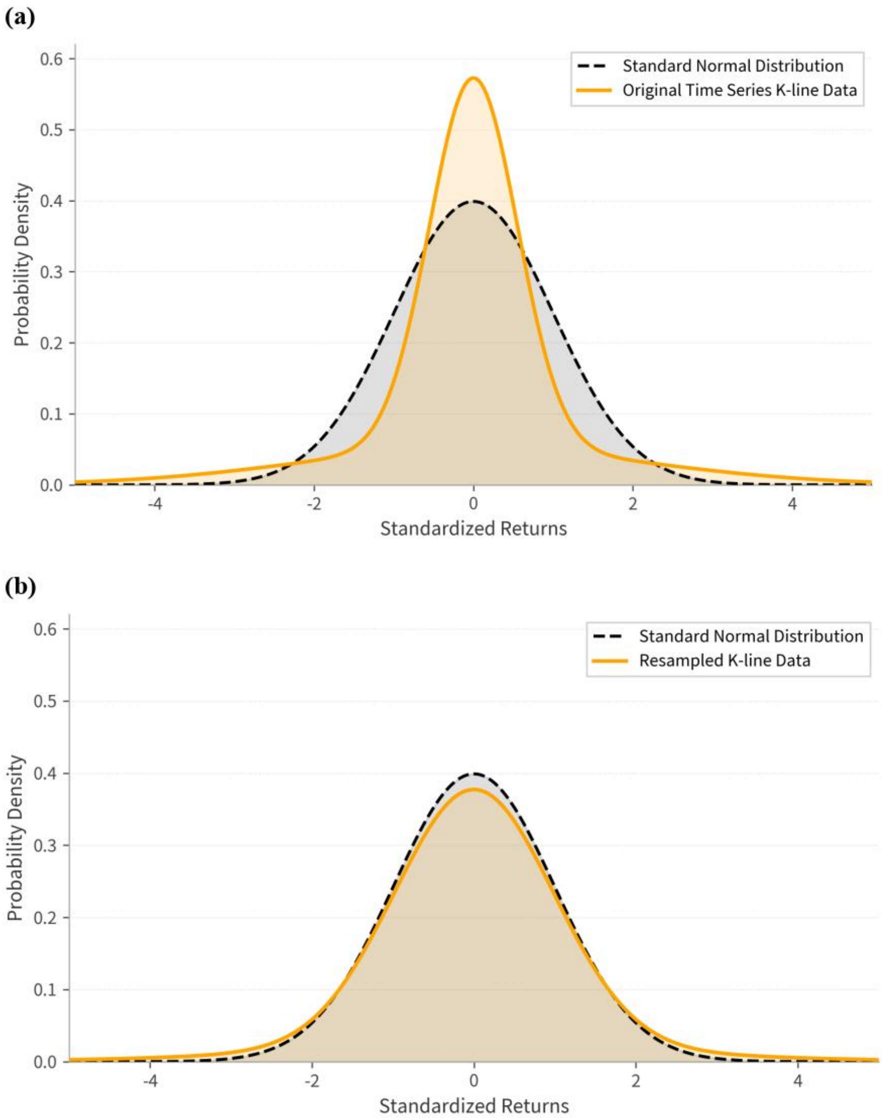

Figure 1. Kernel density estimates of log returns for one-minute rebar data before and after resampling: (a) raw time bars; (b) weighted-threshold resampled bars

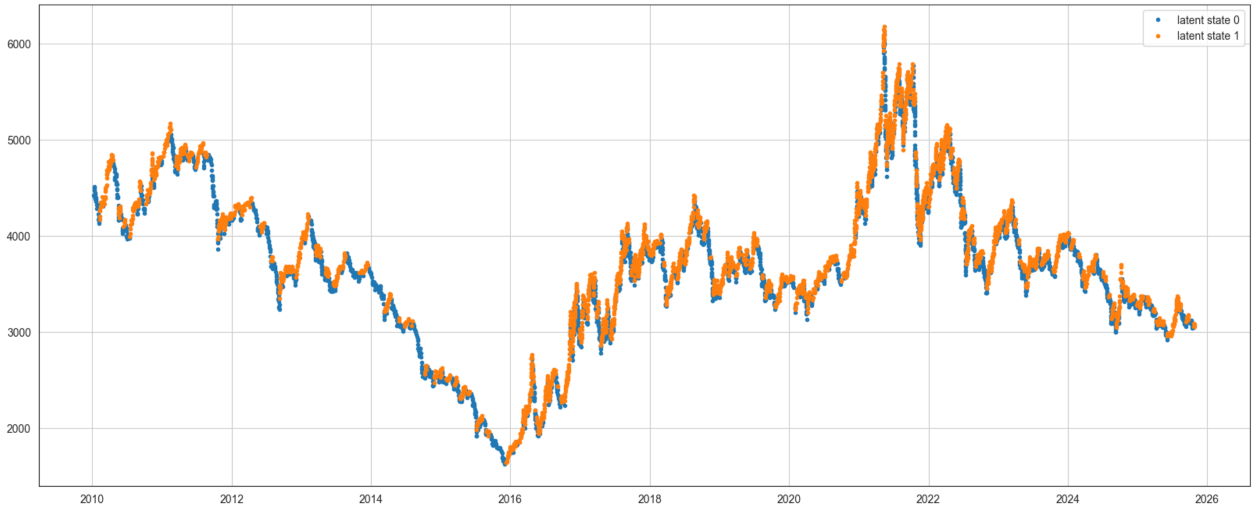

Figure 2. GMM-HMM market-state labels and the price trend of the underlying asset

2.3. Feature Engineering for Ensemble-Learning Models

Feature engineering uses the resampled OHLC data and GMM-HMM state labels. Open, Close, High, and Low are paired in all combinations, and first- through fifth-order logarithmic differences are calculated to form 80 base features.

The base features span different price attributes and differencing windows, jointly capturing short-term reversals and medium-term trends.

Figure 3 summarizes the construction of the 80-dimensional base-feature set.

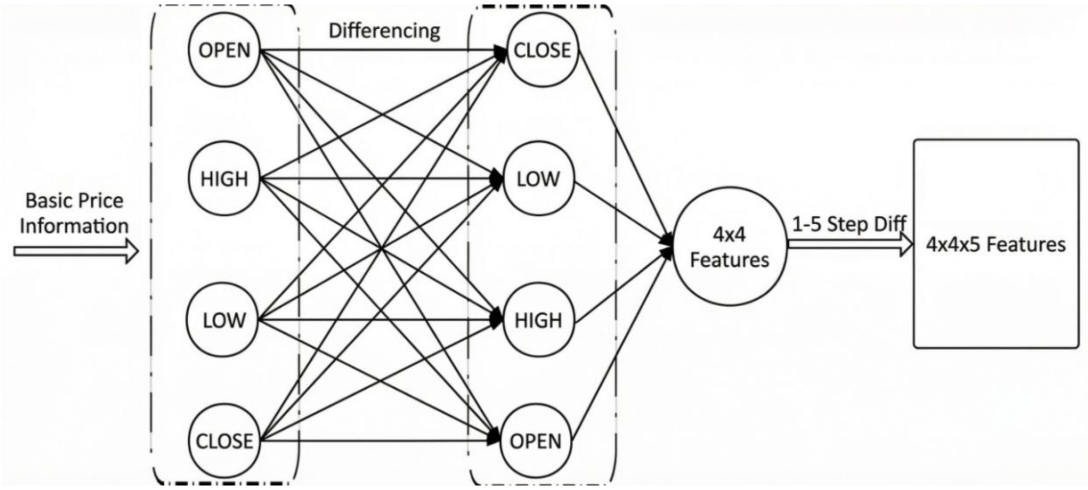

Figure 3. Schematic of the 80-dimensional base-feature construction

For each base feature, the current value and one- through four-period lags are retained, producing an N x 400 feature matrix. Principal component analysis (PCA) is then applied, and the retained components are passed to the supervised-learning models.

2.4. Multi-Model Training and Trading-Signal Generation

2.4.1. Model Training

Model training is implemented in PyCaret using stratified 10-fold cross-validation, with the number of PCA components varied from 1 to 400. Candidate models include XGBoost, CatBoost, and LightGBM. XGBoost records a mean cross-validation accuracy of 0.8117 and an AUC of 0.8968, and XGBoost and CatBoost are selected for the final consensus ensemble.

2.4.2. Trading-Signal Generation

The final signal follows a strict consensus rule: a long position is taken when both XGBoost and CatBoost predict an increase, a short position is taken when both predict a decrease, and no trade is executed when their predictions disagree.

The consensus rule is expressed as:

$$S_{final}(t) = \begin{cases} 1, & \text{if} \sum_{m=1}^{M} \mathbb{I}\,(S_m(t) = 1) = M \\ 0, & \text{if} \sum_{m=1}^{M} \mathbb{I}\,(S_m(t) = 0) = M \\ \text{null}, & \text{otherwise} \end{cases}$$

Where $S_m(t)$ is the prediction of model m at time t, M is the number of models in the ensemble, and Sfinal(t) is the final signal. The no-trade outcome serves as a risk-control state when the models disagree.

3. EMPIRICAL ANALYSIS

3.1. Experimental Setup and Comparison Design

The research subject is the main continuous rebar futures contract, covering 3,883 trading days. Data from January 4, 2010 to January 1, 2023 are used for regime detection, feature construction, and model training; January 2, 2023 to December 31, 2025 is strictly reserved for out-of-sample backtesting.

3.1.1. Backtesting Settings

Initial capital is RMB 1,000,000. Each trade uses a fixed position of 20 contracts, with a contract multiplier of 10 metric tons per contract. A commission of 0.05% of transaction value is charged on both entry and exit, and each trade includes one minimum price tick of slippage.

3.1.2. Benchmarks and Ablation Models

Performance measures include annualized return, maximum drawdown, annualized volatility, Sharpe ratio, Calmar ratio, Sortino ratio, win rate, and profit-loss ratio. The benchmarks are a buy-and-hold strategy and a 5-day/20-day dual-moving-average strategy.

The ablation study introduces four models sequentially. Model A uses conventional daily data and a single XGBoost model. Model B replaces the input with information-driven resampled bars. Model C further adds GMM-HMM regime filtering. Model D builds on Model C by requiring consensus between XGBoost and CatBoost.

3.2. Experimental Results and Overall Performance

Table 1 reports the core performance of each strategy over the 2023-2025 out-of-sample period.

Table 1. Comparison of core out-of-sample performance metrics across strategies

Strategy	Ann. Return	Max. Drawdown	Ann. Volatility	Sharpe Ratio	Calmar Ratio	Sortino Ratio	Win Rate	Profit-Loss Ratio	Trades
Buy and Hold	-6.02%	-24.58%	13.30%	-0.64	-0.24	-0.99	0.0%	0.00	0
Dual-MA Strategy	-1.70%	-15.64%	12.18%	-0.30	-0.11	-0.46	36.7%	1.59	49
Model A	2.91%	-9.43%	11.07%	0.07	0.31	0.11	51.3%	1.07	341
Model B	8.58%	-11.09%	17.13%	0.64	0.77	1.02	54.6%	1.01	130
Model C	8.32%	-11.87%	16.74%	0.58	0.70	0.91	51.5%	1.21	66
Model D	11.33%	-7.8%	16.69%	0.88	1.45	1.05	51.9%	1.27	54

3.2.1. Cumulative Net Value and Returns

Figure 4 shows that the rebar market generally moved sideways and downward during the test period. The buy-and-hold and dual-moving-average strategies recorded annualized returns of -6.02% and -1.70%, respectively. All machine-learning strategies produced positive returns, with Model D performing best at 11.33% per year.

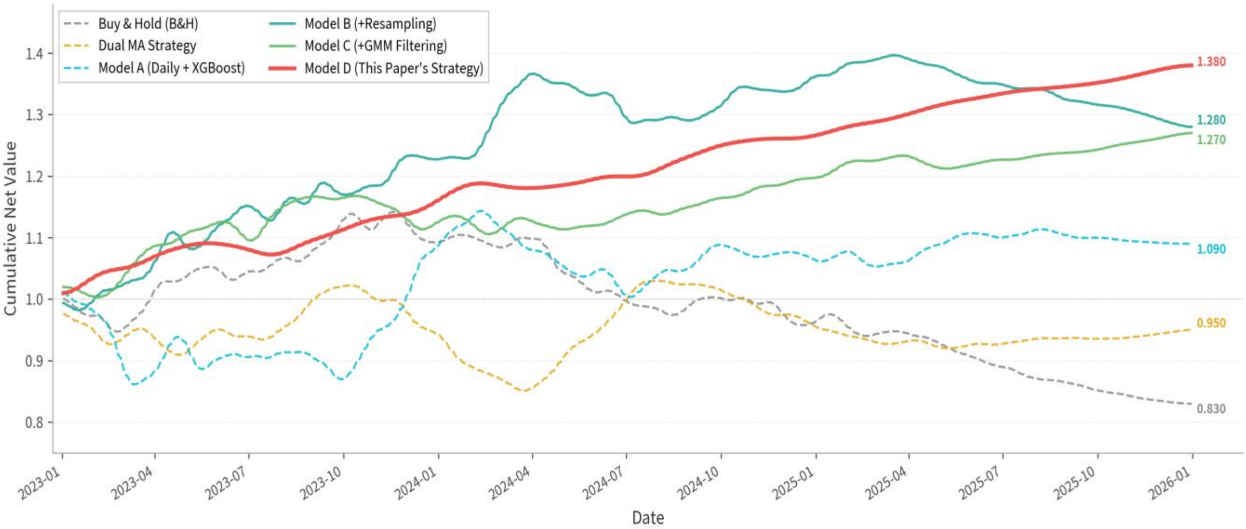

Figure 4. Out-of-sample cumulative net value comparison across strategies (2023.01-2025.12)

3.2.2. Risk Control and Drawdown

Figure 5 shows that the buy-and-hold strategy experienced a maximum drawdown of -24.58%. Through regime filtering and consensus decision-making, Model D limited maximum drawdown to -7.80% and achieved a Sharpe ratio of 0.88, outperforming all comparison strategies on a risk-adjusted basis.

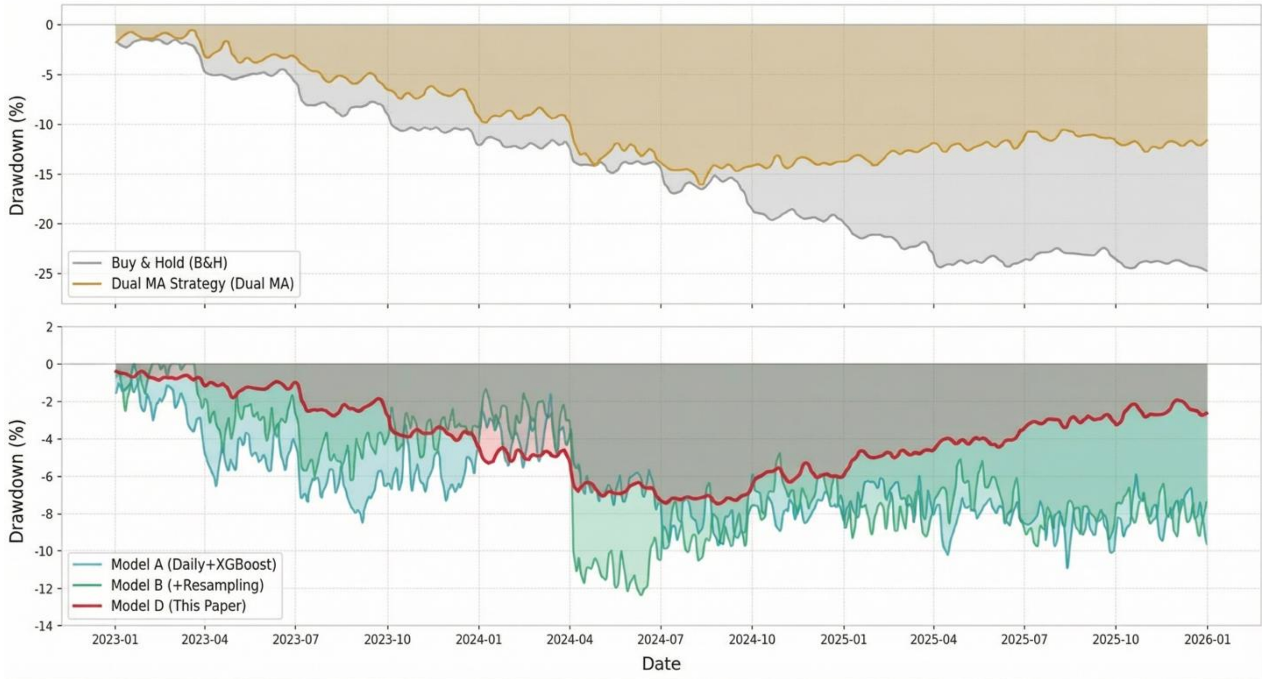

Figure 5. Dynamic drawdown comparison across strategies (2023.01-2025.12)

3.3. Ablation Analysis

Figure 6 shows the marginal contribution of each module:

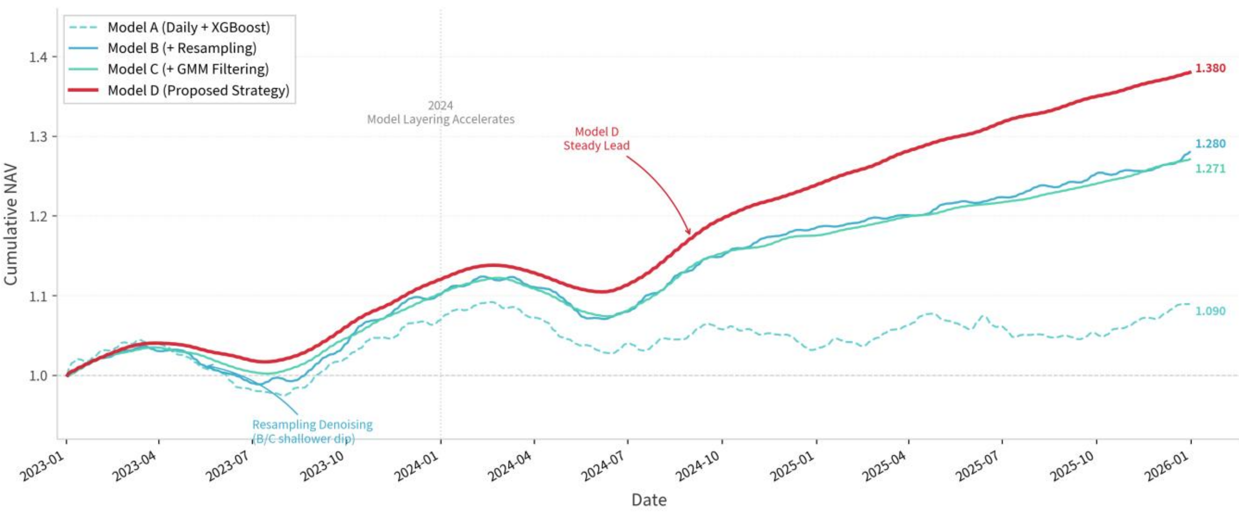

Figure 6. Net value comparison across ablation models (2023.01-2025.12)

Replacing conventional sampling in Model A with resampling in Model B increased annualized return from 2.91% to 8.58%, raised the win rate from 51.3% to 54.6%, and reduced the number of trades from 341 to 130. This indicates that the information axis filters part of the market noise.

After GMM-HMM filtering was added, Model C further reduced the number of trades to 66 and increased the profit-loss ratio from 1.01 to 1.21. Although annualized return declined slightly to 8.32%, trade quality improved.

The dual-model consensus rule increased Model D's annualized return to 11.33%, narrowed maximum drawdown to -7.80%, and raised the Sharpe ratio to 0.88. This indicates that remaining out of the market when the models disagree improves the balance between return and risk.

4. CONCLUSION

This study develops and tests a futures trading framework that combines information-driven resampling, GMM-HMM regime filtering, and multi-model consensus decision-making. The empirical results support three conclusions:

First, the return distribution after resampling is closer to the normal distribution, indicating that the hybrid information axis improves the information quality of the sample.

Second, regime filtering reduces the number of trades from 130 in Model B to 66 in Model C and increases the profit-loss ratio from 1.01 to 1.21.

Third, the complete strategy achieves an out-of-sample annualized return of 11.33%, a 7.80% maximum drawdown, and a Sharpe ratio of 0.88, confirming the complementary effects of regime awareness and multi-model consensus decision-making.

REFERENCES

[1] Mandelbrot, B., & Taylor, H. M. (1967). On the distribution of stock price differences. Operations Research, 15(6), 1057-1062.

[2] Easley, D., Kiefer, N. M., O'Hara, M., & Paperman, J. B. (1996). Liquidity, information, and infrequently traded stocks. Journal of Finance, 51(4), 1405-1436.

[3] De Prado, M. L. (2018). Advances in financial machine learning. John Wiley & Sons.

[4] Sokolovsky, A., Arnaboldi, L., Bacardit, J., et al. (2022). Volume-centred range bars: Novel interpretable representation of financial markets designed for machine learning applications. arXiv preprint arXiv:2103.12419.

[5] Zhang, R. (2021). Research on stock feature K-line chart prediction based on machine learning [Doctoral dissertation]. Dalian University of Technology.

[6] Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica, 57(2), 357-384.

[7] Liu, Z. Q. (2021). Stock price volatility prediction based on Markov chain. Co-Operative Economy & Science, (24), 64-67.

[8] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

[9] Gupta, R., Kapoor, S., Gupta, H., et al. (2025). A forest of opinions: A multi-model ensemble-HMM voting framework for market regime shift detection and trading. Data Science in Finance and Economics, 5(4), 466.