MULTI-MODEL PREDICTIVE ANALYSIS OF REALIZED VOLATILITY USING THE GARMAN–KLASS ESTIMATOR

Xinting Li, Hongtao Ma & Junbo Qian
Peking University, National School of Development
lixt@stu.pku.edu.cn, mahongtao@stu.pku.edu.cn, AlanQ@stu.pku.edu.cn

Introduction

This study leverages multi-source data including the SPY ETF, VIX volatility index, and the U.S. 10-year Treasury yield, constructing daily volatility proxies using the Garman–Klass estimator, alongside log returns and lagged features, to build three distinct models—Heterogeneous Autoregressive (HAR), Random Forest, and Long Short-Term Memory (LSTM)—for daily volatility forecasting.

- **Bridging Classic and Modern Volatility Forecasting:** This study combines classical econometric (HAR), machine learning (Random Forest), and deep learning (LSTM) approaches, creating a unified framework to compare linear and nonlinear models in volatility forecasting.
- **Multi-Source Feature Construction for Robust Prediction:** By incorporating the Garman-Klass estimator, high-frequency data, multi-scale lagged features, VIX, and macro indicators like Treasury yields, the study enhances model inputs and improves volatility prediction's market adaptability.
- **Evidence of Machine Learning Superiority:** Data-driven models like Random Forest and LSTM outperform traditional HAR models in accuracy and trend detection, especially during market turbulence, highlighting modern algorithms' value in financial analysis.

Research Process

Data and Volatility Measurement

- **Data Sources**
  - SPY ETF OHLC data (daily Open, High, Low, Close)
  - VIX index (volatility index)
  - U.S. 10-year Treasury yield (DGS10)
- **Volatility Label Construction**

1. **Log Returns:**

$$\text{Returns}_t = \ln\left(\frac{C_t}{C_{t-1}}\right).$$

2. **Garman–Klass Volatility Proxy** (primary):

$$\sigma^2_{\text{GK},t} = \ln\left(\frac{H_t}{C_t}\right)\ln\left(\frac{H_t}{O_t}\right) + \ln\left(\frac{L_t}{C_t}\right)\ln\left(\frac{L_t}{O_t}\right), \quad \text{GK\_vol}_t = \sqrt{\sigma^2_{\text{GK},t} \times 252}.$$

If any step fails, jump to the fallback method.

3. **Alternative Daily Volatility Proxy:**

$$\widetilde{\text{RV}}_t^{\text{alt}} = \left|\text{Returns}_t\right| \sqrt{252}.$$

Selected Models

HAR Model (Heterogeneous Autoregressive)

- **Modeling Idea & Theory:** The HAR model uses averaged daily, weekly, and monthly realized volatilities as regressors to capture multi-scale volatility persistence. Grounded in the heterogeneous market hypothesis, it reflects how high-, medium-, and low-frequency investors collectively drive volatility dynamics. The simple HAR(3) structure approximates long-memory behavior and achieves robust out-of-sample forecasting[1].

$$\text{RV}_t = \beta_0 + \beta_1 \text{RV}_{t-1} + \beta_2 \frac{1}{5}\sum_{i=1}^{5} \text{RV}_{t-i} + \beta_3 \frac{1}{22}\sum_{i=1}^{22} \text{RV}_{t-i} + \varepsilon_t.$$

One can also optionally add extra lag terms such as $\text{RV}_{t-2}, \text{RV}_{t-3}$, etc.

- **Advantages:**
  - Few parameters and fast fitting.
  - Provides clear economic interpretation for realized volatility.

Random Forest Regressor

- **Modeling Idea & Theory:** Random Forest combines multiple decision trees via Bootstrap sampling and random feature selection, capturing complex financial market patterns without distributional assumptions. Its "diversity + aggregation" approach mitigates overfitting, ensuring robust predictions in high-dimensional, noisy financial data[2].
- **Advantages / Uses:**
  - Insensitive to high-dimensional or noisy features.
  - Provides feature importance scores to analyze which lagged information is most critical for volatility forecasting.

LSTM Network (Long Short-Term Memory)

- **Modeling Idea:** Feed the sequential features (returns, VIX, Treasury yield, daily volatility proxy) directly into two stacked LSTM layers along the time dimension, learning both short- and long-term dependencies to output the next-period volatility[3].
- **Advantages / Uses:**
  - Excels at capturing the dynamic evolution of time-series data, even in the presence of noise.

Feature Engineering

Feature engineering is the process of creating, transforming, selecting, and combining raw data into meaningful features that improve model performance. In this study, the feature engineering involves the following categories:

Table 1: Feature Engineering

Category	Feature Name(s)
Lag Features	Returns_Lag1, Returns_Lag2, …, Returns_Lag5
	VIX_Lag1, VIX_Lag2, …, VIX_Lag5
	DGS10_Lag1, DGS10_Lag2, …, DGS10_Lag5
	StdVol_Lag1, StdVol_Lag2, …, StdVol_Lag5
Statistical Features	Standard Deviation Volatility
	Garman-Klass Volatility Estimator
Delta Features	VIX_Change
HAR Features	Daily_RV (lag-1)
	Weekly_RV (5-day avg)
	Monthly_RV (22-day avg)

Result

According to **Figure 1**, overall, both deep learning and machine learning models demonstrate significantly better numerical accuracy and trend-tracking ability compared to the traditional linear HAR model. This advantage is especially evident during extreme market conditions when volatility experiences abrupt changes.

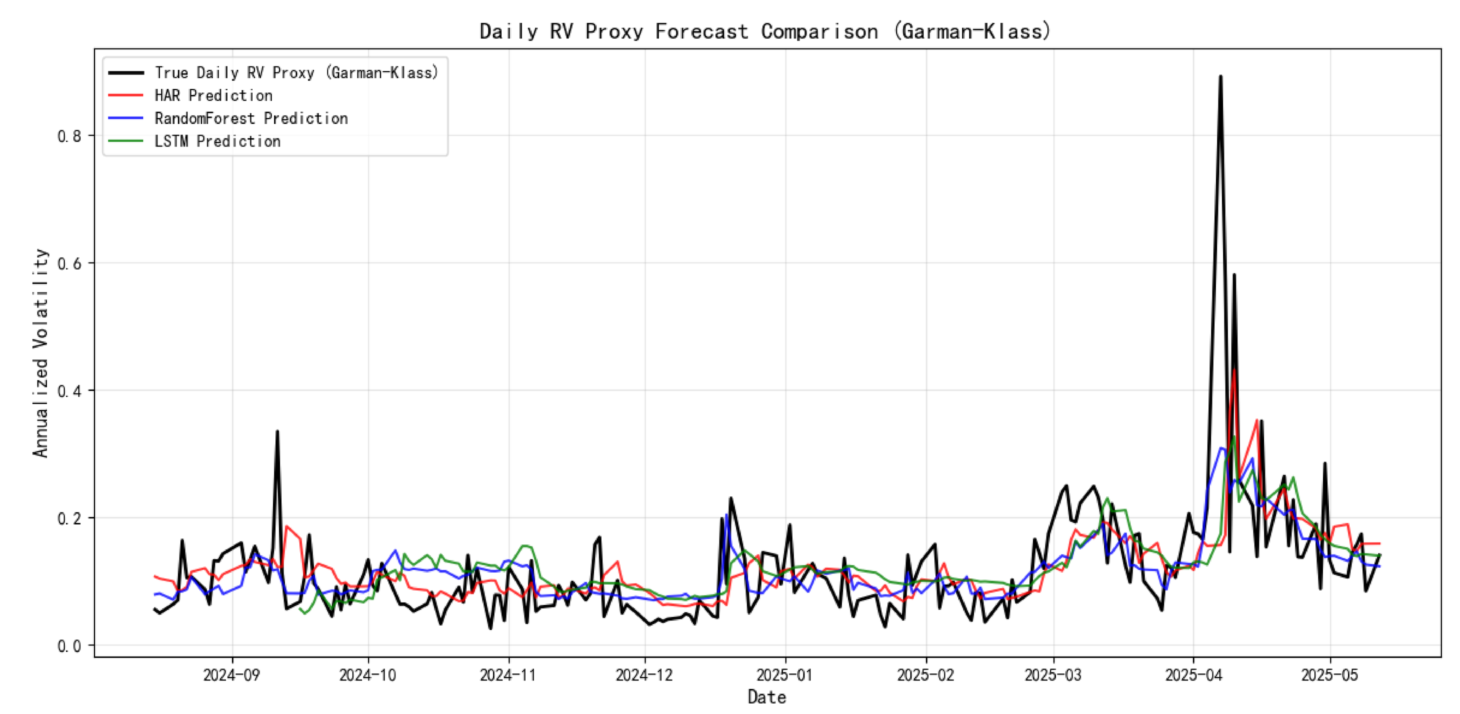

Figure 1: Volatility Forecast Comparison

According to the results in **Table 2**

- LSTM achieves the lowest MSE (0.003636) and MAE (0.038757), indicating the highest prediction accuracy.
- Random Forest ranks second in error metrics and attains the highest direction accuracy (0.4891).
- Both Random Forest and LSTM outperform the traditional HAR model in volatility forecasting, especially in predictive accuracy.

Table 2: Model Performance Metrics

Model	MSE	MAE	Direction Accuracy
HAR	0.006817	0.046630	0.4674
Random Forest	0.005103	0.043793	0.4891
LSTM	0.003636	0.038757	0.4479

Conclusion

- Deep learning and machine learning models (e.g., Random Forest and LSTM) clearly outperform the traditional HAR in numerical accuracy and trend tracking, especially in extreme volatility scenarios, although they still tend to underestimate and lag during sudden spikes.
- In stable markets, prioritize HAR or Random Forest for interpretability; when capturing abrupt risks, rely more on LSTM, and consider combining multiple models to enhance early-warning and adaptability to extreme events.

References

[1]Corsi, F. (2009). *A simple approximate long-memory model of realized volatility. Journal of Financial Econometrics*, 7(2), 174–196.

[2]Luong, C., & Dokuchaev, N. (2018). *Forecasting of realised volatility with the random forests algorithm. Journal of Risk and Financial Management*, 11(4), 61.

[3]Kim, H. Y., & Won, C. H. (2018). *Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. Expert Systems with Applications*, 103, 25–37.