Forecasting Realized Volatility: A Comparative Study of HAR-RV-CARMA and Neural Network Models

Chigozie Andy Ngwaba

Department of Economics, Bradley University, Peoria, IL, USA

E-mail address: cngwaba@bradley.edu

Abstract

This paper conducts a comprehensive out-of-sample comparison between a structurally grounded econometric hybrid model—HAR-RV-CARMA—and popular deep learning benchmarks, including RNN, LSTM, and GRU architectures. Using daily realized volatility forecasting for major U.S. equity ETFs (SPY, QQQ, and IWM), the models are estimated over a training period from 2015 to 2022 and a test period covering 2023 to 2024, a timeframe marked by increased macroeconomic and policy uncertainty. Forecast performance is evaluated using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Quasi-Likelihood (QLIKE), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). Additionally, the study assesses directional accuracy and conducts a Diebold-Mariano test to compare forecast performance. Empirical results suggest that the HAR-RV-CARMA hybrid model outperforms the neural network models in volatility forecasting across all evaluation criteria except directional accuracy. It achieves lower forecast errors and superior goodness-of-fit, with Diebold-Mariano test outcomes rejecting the null hypothesis of equal predictive ability at significant levels. These findings underscore the efficacy of the hybrid model in enhancing predictive accuracy and provide a robust framework for volatility modeling in financial markets.

Keywords: Volatility forecasting; Neural networks; Kalman filter; HAR-RV; CARMA

JEL classification: C22; C44; G17

1. Introduction

Forecasting financial market volatility remains a central problem in empirical finance, with direct implications for risk management, derivative pricing, portfolio allocation, and regulatory oversight. Accurate volatility prediction may assist fund managers in reducing their exposure to market risks, and the government can benefit from accurate volatility forecasts by assessing their impact on the economy, thereby formulating effective policy recommendations. Predicting volatility is challenging because it is a latent variable, meaning it cannot be observed directly and must be estimated. Researchers often estimate volatility with methods like realized volatility, which is the annualized standard deviation of log returns, and implied volatility, which is derived from options. Studies such as Andersen et al. (2003), Qiu et al. (2019), Korkusuz et al. (2023), Lin et al. (2022), and Dutta and Das (2022) have conducted comprehensive modeling of realized volatility. Similarly, Studies including Psaradellis and Sermpinis (2016), Vrontos et al. (2021), Aït-Sahalia et al. (2021), Michael et al. (2025), and Jiao et al. (2024) focused on predicting, modeling, or forecasting implied volatility.

Traditional econometric volatility models generally rely on frameworks such as the generalized autoregressive conditional heteroskedasticity (GARCH) introduced by Bollerslev (1986), the heterogeneous autoregressive (HAR-RV) introduced by Corsi (2009), and stochastic volatility models such as CARMA (Brockwell, 2001). Brockwell (2014) found that a Lévy-driven CARMA (2,1) process fits daily realized volatility series well. Ngwaba (2025) proposed a hybrid model, HAR-RV-CARMA, that combines the forecasting strengths of HAR-RV and CARMA to improve volatility prediction. The key innovation of the approach is the use of Kalman filter dynamic weighting to integrate the HAR-RV and CARMA components, rather than assigning fixed weights or combining forecasts arbitrarily. The proposed hybrid model was applied to the realized volatility of five covered call ETFs, and the results showed that it outperformed the standalone HAR-RV and CARMA models. This study modifies Ngwaba (2025); instead of using a standard CARMA model, we implement a CARMA (2,1) model within the HAR-RV-CARMA framework.

Simultaneously, deep learning techniques, notably recurrent neural networks, have gained considerable popularity in financial forecasting applications. Architectures such as Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), and Gated Recurrent Unit (GRU) are often regarded as suitable alternatives to conventional econometric models, owing to their ability to approximate nonlinear dynamics and handle sequential data. Although traditional econometric techniques have demonstrated effectiveness in short-term forecasting, they are associated with several limitations. Neural

network models were subsequently introduced to mitigate these issues (QIU et al., 2025). RNNs use activation functions to model nonlinear relationships; they retain information specific to each timestep and sequentially update it at each subsequent timestep. This makes them ideal for time series analysis. In practice, however, conventional RNNs suffer from the well-known vanishing (and exploding) gradient problem, whereby gradients propagated backward through long sequences decay rapidly, making it difficult for the network to learn long-horizon dependencies. LSTM and GRU were introduced to address this problem (Pascanu et al., 2013). LSTM networks were introduced to address this issue by incorporating gated memory cells with input, forget, and output gates that selectively control the flow of information and enforce near-constant error flow over long durations (Hochreiter and Schmidhuber, 1997). GRU networks provide a more parsimonious alternative by combining gating mechanisms into update and reset gates, reducing the number of parameters while retaining the ability to mitigate vanishing gradients and capture long−term structure (Das et al., 2023).

This paper provides a direct comparison between a structurally motivated econometric hybrid volatility model, HAR-RV-CARMA, and three widely used deep learning benchmarks: vanilla RNN, LSTM, and GRU. The analysis focuses on daily realized volatility forecasts for major U.S. equity ETFs (SPY, QQQ, and IWM). Models are estimated using a long training sample from 2015 through 2022 and evaluated out-of-sample strictly over the 2023–2024 period. The goal of the paper is to ask and answer a simple question: Does replacing a well-specified structural volatility model with a generic deep learning architecture improve forecast performance in practice? To answer this, we evaluate models along multiple dimensions: statistical fit (AIC and BIC), magnitude accuracy (MAE and RMSE), likelihood-based loss (QLIKE), Directional accuracy and Diebold-Mariano tests.

The HAR-RV-CARMA outperforms neural network models across both in-sample and out-of-sample forecast accuracy and is consistent across all assets. RNN shows modest improvements only in directional accuracy, but these gains are small and come at the cost of substantially worse magnitude and likelihood-based performance. Overall, the evidence favors structural modeling over black-box flexibility in realized volatility forecasting.

The paper is organized as follows: Section 2 reviews the literature; Section 3 presents the data and estimation methods; Section 4 discusses the estimation results; Section 5 offers the discussion; and Section 6 concludes the paper.

2. Literature Review

An increasing number of studies have investigated volatility forecasting through time-series, machine learning, deep learning, and hybrid models. The GARCH model remains essential in volatility prediction due to its strong, interpretable framework for capturing time-varying volatility and clustering (Engle 1982; Bollerslev 1986). Gunnarsson et al. (2024) provided a thorough review of the literature on the use of artificial intelligence and machine learning to predict realized and implied volatility indices. Ma et al. (2024) and Song et al. (2023) studied stock market volatility prediction using macroeconomic variables, employing hybrid models that combine deep learning techniques with mixed data sampling (MIDAS).

Ngwaba (2025) introduced the HAR-RV-CARMA hybrid model, combining the HAR-RV and the CARMA model. Its main innovation is a Kalman-filter-based dynamic weighting mechanism that integrates both discrete- and continuous-time models, harnessing their strengths. The hybrid model is applied to daily realized volatility data from five major covered call ETFs. The empirical results show that the proposed model consistently outperforms the standalone models across various statistical measures, including RMSE, MAE, QLIKE, AIC, and BIC. Additional tests, such as the Diebold–Mariano and directional accuracy tests, confirm the hybrid model's superior predictive performance.

Brockwell and Lindner (2013) improve the modeling of spot volatility in financial time series by integrating CARMA processes into the stochastic volatility framework, building on the foundational work of Barndorff-Nielsen and Shephard (2001). They suggest using CARMA processes driven by non-decreasing Lévy processes with non-negative kernels to model spot volatility, offering greater flexibility than the traditional Ornstein-Uhlenbeck (OU) process. CARMA processes offer a broader range of autocorrelation structures, enabling a more accurate depiction of the persistent and complex dependencies observed in realized volatility data. Empirically, their study applies the CARMA-based volatility model to high-frequency DEM/USD exchange rate data and finds a better fit than the OU-based model. The results emphasize the model's capacity to capture key features of financial volatility—such as volatility clustering and long-range dependence—while remaining mathematically manageable.

Corsi (2009) proposed a framework for capturing the persistent behavior observed in financial market volatility. Motivated by the heterogeneous market hypothesis, Corsi constructs an additive series of realized volatility components defined over multiple time horizons, leading to a Heterogeneous Autoregressive model of Realized Volatility (HAR-RV) that linearly aggregates daily, weekly, and monthly

realized volatilities. Although the HAR-RV lacks formal long memory in the strict statistical sense, simulation evidence indicates that its simple structure replicates key empirical features of financial returns. Corsi demonstrates that the model not only reproduces these stylized facts but also delivers strong out-of-sample forecasting performance relative to traditional volatility models, making it a widely adopted benchmark in realized volatility forecasting research.

Gunnarsson et al. (2024) reviewed studies on Artificial Intelligence (AI) and Machine Learning (ML) for forecasting volatility. They found memory-based neural networks like LSTM and GRU, as well as tree methods, often outperform traditional models, though HAR-RV and GARCH remain benchmarks. Hybrid models combining econometrics with ML perform well by capturing linear and nonlinear dynamics. Most applications focus on U.S. equity and oil markets, with forecast horizons from intraday to monthly. Neural networks excel at short-term forecasts, while longer-term forecasts benefit from external predictors. ML models outperform traditional ones during high volatility and with richer data. There is a gap in integrating explainable AI (XAI), despite concerns about the "black box" nature of ML. Future research should incorporate XAI and probabilistic ML to enhance transparency and account for forecast uncertainty.

Qiu et al. (2025) offer a comprehensive historical overview, benchmarking models from implied volatility and GARCH to LSTM and transformer architectures. Their findings substantiate the superior performance of LSTM but also highlight key limitations of deep models, including substantial data and computational demands, instability during periods of crisis, and limited interpretability. They advocate adopting hybrid and ensemble approaches; however, their analysis is confined to the S&P 500 index, raising concerns about the robustness of these models across various assets and market regimes.

3. Data and Estimation

3.1. Data

The study uses daily historical time series data from three major U.S. equity ETFs (SPY, QQQ, and IWM). Table 1 provides details about the assets included in the study. Broad market indices were chosen to reflect overall market behavior across different sectors.

SPY is the most well-known and oldest ETF listed in the U.S., often leading in assets under management (AUM) and trading volume. It tracks the widely followed S&P 500 index, which selects 500 securities to represent the U.S. large-cap market. The index committee chooses these securities, not necessarily the 500 with the highest market cap, which may result in some omissions. Nonetheless, the

index provides excellent exposure to the U.S. large-cap sector. SPY was launched on 01/22/1993 and has an expense ratio of 0.09% and $711.94b in AUM.

Table 1: Summary Information for the ETF

Ticker	Name	Issuer	Expense ratio	AUM ($ Billions)	Inception Date
SPY	SPDR S&P 500 ETF Trust	SPDR	0.09%	711.94	01/22/1993
QQQ	Invesco QQQ Trust	Invesco	0.20%	403.87	03/10/1999
IWM	iShares Russell 2000 ETF	iShares	0.19%	73.22	05/22/2000

Source: www.etf.com as of 6 January 2026

QQQ tracks a modified-market-cap-weighted index comprising 100 NASDAQ-listed stocks. It is among the most established and actively traded ETFs globally, although it remains somewhat unconventional. The fund is structured as a unit investment trust, a relatively rare structure among ETFs. According to its underlying index's criteria, the fund invests exclusively in nonfinancial stocks listed on NASDAQ and largely excludes other sectors, resulting in a significant deviation from a broad-based large-cap portfolio. Notably, QQQ possesses substantial exposure to the technology sector. QQQ was launched on 03/10/1999 and has an expense ratio of 0.20% and $403.87b in AUM.

IWM tracks a market-cap-weighted index of US small-cap stocks, primarily the Russell 2000. Its extensive holdings make it one of the most diversified funds in this category. The fund also invests in micro-cap stocks, which have often made it riskier than the neutral benchmark. Nevertheless, incorporating micro-caps is a legitimate and potentially advantageous strategy within small-caps, and IWM generally aligns with the benchmark in other aspects. IWM was launched on 05/22/2000 and has an expense ratio of 0.19% and $73.22b in AUM. (Source: etf.com)

Table 2 reports summary statistics for the daily volatility series of SPY, QQQ, and IWM over the sample period. The daily ETF prices were obtained from Yahoo Finance for the period January 2015 to December 2024. This time frame was chosen to capture a range of high-volatility market conditions, including the COVID-19 market crash, interest rate hikes, the Federal Reserve's quantitative tightening, and significant geopolitical events. Realized volatility is computed as the annualized standard deviation of log returns over a 21-day rolling window, corresponding to approximately one trading month. The training set spans from January 2015 to December 2022, while the test set covers January 2023 to December 2024.

Table 2: Summary Statistics for Daily Realized Volatility (January 2016 – December 2024)

Sample Statistics	SPY	QQQ	IWM
Mean	0.1514	0.1946	0.2010
Median	0.1242	0.1648	0.1854
Std Dev	0.1019	0.1098	0.1088
Skewness	3.7290	2.5452	3.5281
Kurtosis	25.0607	14.8259	23.0403
Min	0.0347	0.0499	0.0509
Max	0.9400	0.9255	1.0430
Jarque-Bera (p-value)	0.0000	0.0000	0.0000
ADF (p-value)	0.0100	0.0100	0.0100

The mean and median volatility are higher for QQQ and IWM than for SPY, reflecting their greater exposure to growth-oriented and smaller-capitalization firms. All three series exhibit substantial dispersion, with standard deviations large relative to their means, highlighting pronounced time variation in volatility. The distributions are strongly right-skewed and highly leptokurtic, with skewness exceeding 2.5 and kurtosis far above the Gaussian benchmark, indicating the presence of infrequent but extreme volatility realizations. Jarque–Bera test statistics strongly reject the null of normality for all assets, confirming the non-Gaussian nature of daily volatility. In contrast, augmented Dickey–Fuller test results reject the presence of a unit root at conventional significance levels, suggesting that the volatility series are stationary despite their high persistence. These properties are consistent with well-documented stylized facts of financial market volatility and motivate the use of models designed to accommodate persistence, asymmetry, and fat tails.

Figure 1 displays the daily volatility series for SPY, QQQ, and IWM from 2015 to 2024. It shows clear volatility clustering, with long periods of low volatility interrupted by sharp spikes during market stress.

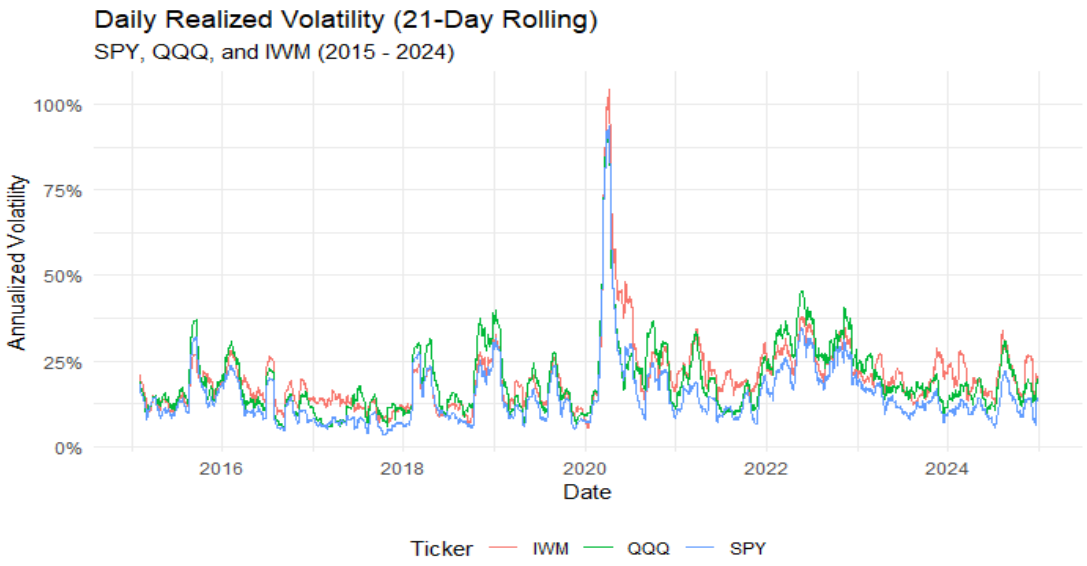


Figure 1: Daily Realized Volatility for the ETFs

The most noticeable spike occurs during the COVID-19 crisis, when volatility across all three assets rises quickly, then gradually returns to pre-crisis levels. Although these assets exhibit similar volatility patterns, QQQ and IWM tend to have larger, more frequent spikes than SPY, suggesting they are more sensitive to macroeconomic and financial shocks. Overall, the chart emphasizes the recurring yet episodic nature of volatility and highlights the need to model both short-term fluctuations and long-term dependence, providing visual context for the forecasting models discussed later.

3.2. Forecasting Model

3.2.1 The HAR-RV-CARMA Model

The HAR-RV model has emerged as the benchmark in the financial econometrics literature for realized volatility forecasts. The model can be expressed as follows:

$$RV_{HAR,t} = \beta_0 + \beta_1 RV_{t-1} + \beta_2 RV_{t-5} + \beta_3 RV_{t-22} + \epsilon_t \qquad (1)$$

RV_{t-1}, represents the lagged daily volatility. RV_{t-5}, represents the weekly volatility. RV_{t-22}, represents the monthly volatility. ϵ_t is the error term.

To model high-frequency noise and mean-reverting behavior that HAR-RV does not capture, we incorporate the CARMA model, known for its effectiveness in representing continuous-time volatility, particularly in financial data. The model can be expressed as follows:

$$dX_t = AX_t dt + BdW_t \quad (2)$$

$$Y_t = CX_t + v_t \quad (3)$$

Where X_t is the latent state process, W_t is the Brownian motion, A, B, and C are model parameters, Y_t is the spot volatility, and v_t is the white noise. Y_t is spot volatility; however, this study examines realized volatility. To obtain our realized volatility at time t, we take the integral of the spot volatility from time $t-1$ to t.

$$RV_{CARMA,t} = \int_{t-1}^{t} Y(s)ds \quad (4)$$

Equation (4) is the realized volatility for the CARMA model, which is essential for representing stochastic processes that develop continuously over time.

The Kalman filter for adaptive model combination can be expressed as follows:

$$w_t = w_{t-1} + \eta_t,\ Where\ \eta_t \sim N(0, Q) \quad (5)$$

Where $w_t\ and\ w_{t-1}$, are the weights at time t , and $t-1$ respectively, η_t captures measurement uncertainty, and Q is the state transition variance controlling the weight flexibility. Hence, the HAR-RV-CARMA model can be expressed as follows:

$$RV_t = w_t^{(1)} RV_{HAR,t} + w_t^{(2)} RV_{CARMA,t} + \epsilon_t \quad (6)$$

Where RV_t represents the hybrid realized volatility (HAR-RV-CARMA). $w_t^{(1)}$ and $w_t^{(2)}$ are the Kalman-updated weights at time t. ϵ_t represents the error term.

3.2.2. Recurrent Neural Networks (RNN)

The Recurrent Neural Network (RNN) is a specialized neural network designed for sequential data. Its architecture includes input and output layers that enable it to process and retain information over time, making it ideal for tasks such as language modeling and time-series prediction, where data order is crucial. The RNN maintains memory of previous inputs, enabling it to understand trends and patterns in data (Henrique et al., 2018). The model is expressed as follows.

$$h_t = f_h(u_t x_t + W_{t-1} h_{t-1}) \quad (7)$$

$$y_{t+T} = f_y(v_t h_t + b_y) \quad (8)$$

Where h_t represents the hidden layer vector, x_t is the input layer vector, y_{t+T} is the output layer, u_t is the input-to-hidden weight at time *t*, v_t is the hidden-to-output weight at time *t*, and W_{t-1} is the weight from the output state at time *t-1* to the hidden state at time *t*.

3.2.3. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are a class of recurrent neural networks designed to address the vanishing gradient problem that limits the ability of standard RNNs to capture long-range temporal dependencies. The LSTM architecture introduces a memory cell regulated by gating mechanisms that control the retention, update, and output of information over time. This gated structure allows the model to preserve relevant historical information while selectively forgetting irrelevant components, enabling effective modeling of persistent and nonlinear dynamics in time-series data. These properties make LSTMs particularly suitable for applications where long-term dependencies and temporal structure play a central role (Hochreiter & Schmidhuber, 1997).

3.2.4. Gated Recurrent Unit (GRU)

The GRU, introduced by Cho et al. (2014), addresses the vanishing gradient issue in standard RNNs. It is similar to LSTM but combines the forget and input gates into a single update gate. Additionally, the GRU merges the cell state and hidden state. A GRU unit comprises a cell that performs multiple operations, each of which may be a neural network. Both LSTM and GRU effectively combat the vanishing gradient problem in long sequence models. Compared to LSTMs, GRUs involve fewer tensor operations and train faster (Noh, 2021).

3.3. Assessment Indicators

This study evaluates the models' performance using a range of widely accepted statistical indicators: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Quasi-likelihood (QLIKE), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC).

MAE estimates discrepancies between predicted and actual values, offering a direct yet effective measure of prediction accuracy.

RMSE highlights larger errors by squaring the differences before averaging and taking the square root, making it particularly sensitive to extreme deviations.

QLIKE is a volatility-specific accuracy measure that penalizes both under- and over-prediction of conditional variance, making it particularly suitable for evaluating volatility forecasts.

AIC and BIC serve as model selection tools that balance goodness-of-fit with model complexity. Both penalize models with excessive parameters, but BIC applies a stricter penalty, favoring simpler models even more.

To ensure robustness, the study includes additional tests such as the directional accuracy test and the Diebold-Mariano test.

The directional accuracy test evaluates whether a model correctly predicts the sign of changes in volatility, thereby assessing its ability to capture market dynamics.

Finally, the study uses the Diebold and Mariano (1995) predictive accuracy test to assess if forecast improvements are statistically significant. This test compares two models' forecast errors to determine if differences are meaningful.

These metrics offer a comprehensive framework for assessing model accuracy and efficiency, enabling well-informed conclusions about their relative performance.

4. Results

TABLE 3: Out-of-Sample forecast results

Models	Metrics	**SPY**	**QQQ**	**IWM**
HAR-RV-CARMA	MAE	0.0201	0.0279	0.0316
	RMSE	0.0290	0.0397	0.0464
	QLIKE	-1.188	-0.8437	-0.7166
GRU	MAE	0.6116	0.9154	1.1617
	RMSE	0.6159	0.918	1.1654
	QLIKE	1.0756	1.1151	1.1925
LSTM	MAE	0.2018	0.2009	0.1917
	RMSE	0.2936	0.3028	0.2833
	QLIKE	0.0453	0.0441	0.0424
RNN	MAE	0.08054	0.0741	0.0674
	RMSE	0.1063	0.0968	0.0871
	QLIKE	0.02261	0.0201	0.0162

Table 3 reports out-of-sample forecast errors for the test period 2023–2024. HAR-RV-CARMA achieves lower MAE and RMSE values across all three ETFs. For SPY, HAR-RV-CARMA's MAE is 0.0201, compared to 0.0805 for RNN and 0.2018 for LSTM. The performance gap widens further for GRU, which exhibits larger forecast errors. This is consistent for QQQ and IWM.

The QLIKE results reinforce this conclusion. HAR-RV-CARMA is the only model producing strongly negative QLIKE values, indicating superior alignment with the conditional distribution of realized volatility. All neural network models produce positive QLIKE losses, reflecting poor density-level forecasts.

TABLE 4: Model Goodness of fit test

Models	Metrics	**SPY**	**QQQ**	**IWM**
HAR-RV-CARMA	AIC	-1935.07	-1670.59	-1563.76
	BIC	-1909.77	-1645.29	-1538.46
GRU	AIC	5069	9784.77	9994.02
	BIC	24096.53	28812.3	29021.55
LSTM	AIC	0.00896	0.00899	0.00892
	BIC	0.00027	0.000273	0.000272
RNN	AIC	-0.00357	-0.00341	-0.00324
	BIC	0.0056	0.00572	0.0059

Table 4 shows AIC and BIC values, with HAR-RV-CARMA outperforming all other models across different assets by having much lower information criteria. Notably, the GRU model displays extremely high BIC values, indicating serious overfitting due to excessive parameters. These findings highlight the importance of simplicity in volatility modeling, as more flexible models do not compensate for the lack of a clear structure.

TABLE 5: Directional Accuracy test (%)

Model	**SPY**	**QQQ**	**IWM**
HAR-RV-CARMA	56.4	53.2	53.8
GRU	52.2	51.4	50.6
LSTM	52.19	48.64	49.06
RNN	60.46	58.16	61.72

The results of directional accuracy exhibit a more nuanced perspective. RNN attains the highest directional accuracy, with 60.46% for SPY, 58.16% for QQQ, and 61.72% for IWM. The HAR-RV-CARMA model demonstrates moderate directional performance, achieving 56.4% for SPY, 53.2% for QQQ, and

53.8% for IWM. Nonetheless, these directional gains achieved by the RNN model are not accompanied by enhancements in magnitude or likelihood-based accuracy. In practical applications, the utility of accurately predicting the direction of volatility changes remains limited when the level of volatility itself is inadequately forecasted.

Table 6: Diebold Mariano test among forecast methods

Model	Benchmark		
	LSTM	RNN	HAR-RV-CARMA
GRU	-6.207***	-3.88***	-9.832***
LSTM		0.832	-10.719***
RNN			-16.4***

Note: *** denotes statistical significance at 1%.

To evaluate the predictive performance of different forecasting models, the study uses the Diebold-Mariano test, which determines whether there is a statistically significant difference in forecast accuracy between two competing models based on their errors. Table 6 displays the results of the Diebold-Mariano test. The Diebold-Mariano statistics show the size and direction of accuracy differences, while the p-value indicates statistical significance. The results reveal significant differences in forecast accuracy among the models. The HAR-RV model outperforms the LSTM, RNN and GRU models and is significantly different at a 1% significance level ($p < 0.01$). Similarly, the LSTM and RNN models outperform the GRU model and are significantly different at a 1% significance level ($p < 0.01$). Lastly, there is no significant difference between the LSTM and RNN models ($p > 0.1$). The Diebold-Mariano test results are consistent with the out-of-sample forecast results.

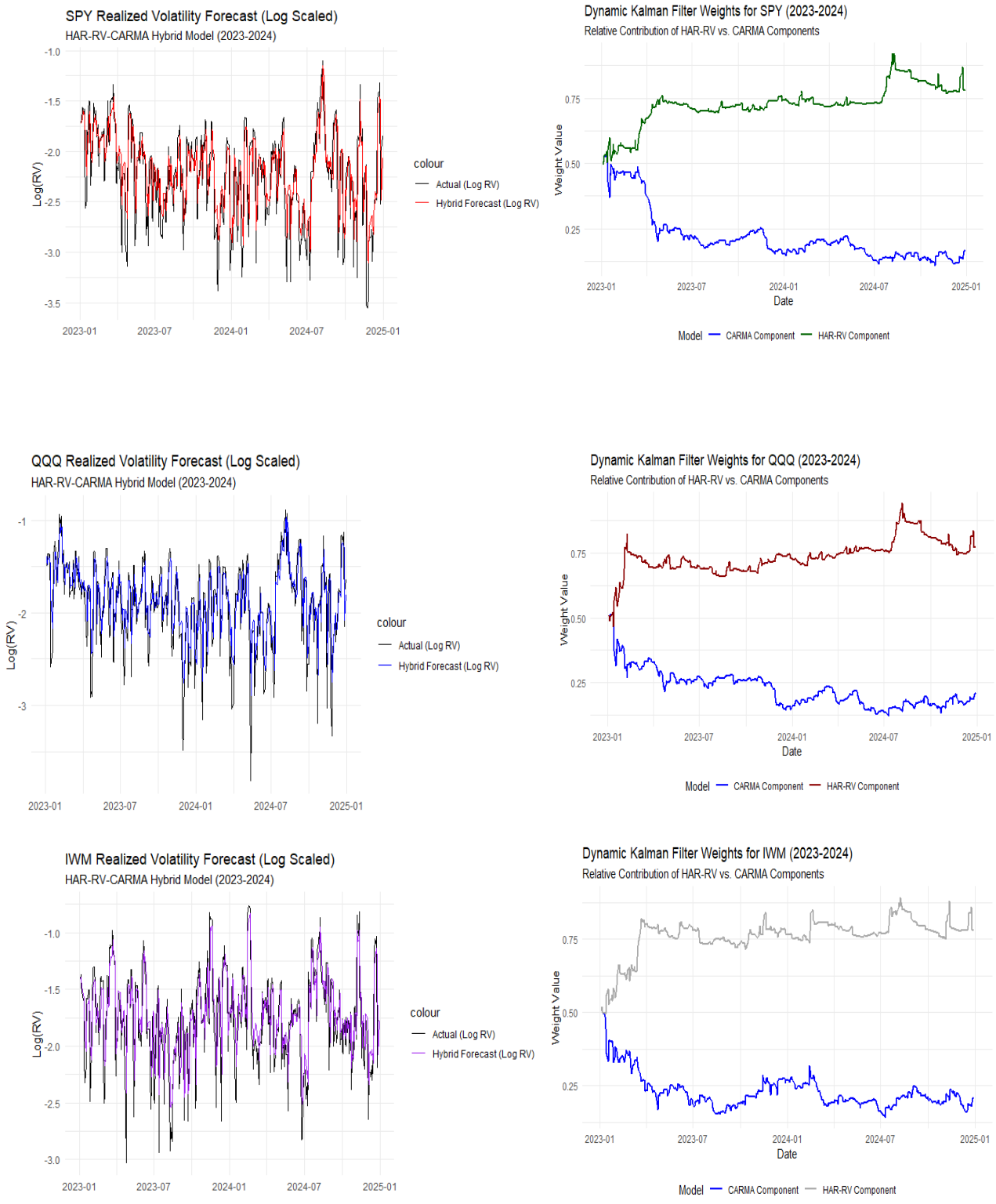


Figure 2: Forecast model comparison of the actual and predicted values (HAR-RV-CARMA)

Figure 2 shows the log-scaled realized volatility for the HAR-RV-CARMA model, the actual versus predicted volatility measured, and the dynamic Kalman filter weights assigned to each model for the test period 1 January 2023 to 31 December 2024. The actual-versus-predicted plot confirms the strong tracking ability of the HAR-RV-CARMA model, while the Kalman filter weights provide insight into how the model dynamically reallocates importance between discrete- and continuous-time volatility representations. This adaptive structure helps explain the improved forecasting performance observed in subsequent evaluation results and underscores the advantage of combining econometric models within a state-space framework. The weights vary smoothly rather than remaining fixed, indicating that the relative importance of the discrete-time HAR-RV structure and the continuous-time CARMA component changes across market conditions.

5. Discussion

Overall, this study accomplishes several goals. First, instead of using a standard CARMA model, we implement a CARMA (2,1) model within the HAR-RV-CARMA framework. Second, we apply the HAR-RV-CARMA model to broad market indices for the first time.[1] Third, we show that model complexity does not necessarily translate into model accuracy. Taken together, the results suggest that flexibility alone is not sufficient for effective volatility forecasting. Recurrent neural networks appear capable of capturing short-term directional movements but struggle to internalize the persistent, multi-horizon structure that characterizes realized volatility.

Similarly, while LSTM and GRU architectures are designed to mitigate the vanishing gradient problem, the Simple RNN demonstrated superior predictive accuracy on the 2023–2024 test set for SPY, QQQ, and IWM. This outperformance can be attributed to the short-memory nature of realized volatility and the regime-switching characteristics of financial markets. This is inconsistent with Gunnarson et al. (2024). Their study found that memory-based neural networks, such as LSTM and GRU, consistently rank among the top-performing models for volatility prediction

These findings do not imply that machine learning methods are unsuitable for volatility modeling. Rather, they suggest that generic deep learning architectures are unlikely to outperform well-specified econometric models unless they are augmented with structural constraints that reflect the persistent, multi-horizon nature of financial volatility.

[1] Ngwaba (2025) applied the HAR-RV-CARMA model to covered call ETFs, not individual stocks or broad market indices.

6. Conclusion

This paper compares a structural econometric volatility model, HAR-RV-CARMA, with recurrent neural network alternatives using realized volatility data for major U.S. equity ETFs. Using a long training sample (2015–2022) and a challenging post-pandemic test period (2023–2024), we show that HAR-RV-CARMA consistently outperforms deep learning models across information criteria, magnitude-based loss functions, and likelihood-based evaluation metrics. While neural networks offer modest gains in directional accuracy, these gains are economically limited and come at the cost of substantially worse forecast accuracy. The results reinforce the continued relevance of structure, parsimony, and interpretability in volatility forecasting.

The results of this study have important implications for practitioners, policymakers, and financial regulators. Enhanced predictive accuracy improves crucial areas of investment management, such as optimizing the timing of asset allocation, more effectively hedging downside risks, and accurately pricing derivatives embedded in option strategies. From a regulatory and policy standpoint, precise and prompt volatility forecasts are crucial for detecting emerging systemic risks and tracking liquidity stress in markets with structured equity products. For practitioners, the implication is simple: more complex models do not automatically deliver better volatility forecasts. For researchers, the findings point toward hybrid approaches that combine structural insight with machine-learning flexibility rather than relying solely on black-box architectures.

While the HAR-RV–CARMA hybrid model with Kalman-filter-based dynamic weighting demonstrates strong predictive performance, it also has limitations that offer opportunities for improvement. The current framework assumes linear state transitions and Gaussian noise, which may not fully capture the nonlinear behaviors often observed in financial markets. Future research could consider incorporating complexity measures to enhance predictive accuracy. Similarly, extending the model to include implied volatility could provide additional insights for practitioners, stakeholders, and policymakers.

References

Aït-Sahalia, Y., Li, C., & Li, C. X. (2021). Implied stochastic volatility models. *The Review of Financial Studies*, *34*(1), 394-450.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, *71*(2), 579-625.

Barndorff-Nielsen, O. E., & Shephard, N. (2001). Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 167-241.

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

Brockwell, P. J. (2001). Continuous-time ARMA processes. *Handbook of statistics*, *19*, 249-276.

Brockwell, P. J. (2014). Recent results in the theory and applications of CARMA processes. *Annals of the Institute of Statistical Mathematics*, *66*, 647-685.

Brockwell, P. J. (2014). Recent results in the theory and applications of CARMA processes. Annals of the Institute of Statistical Mathematics, 66, 647-685.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

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

Das, S., Tariq, A., Santos, T., Kantareddy, S. S., & Banerjee, I. (2023). Recurrent neural networks (RNNs): architectures, training tricks, and introduction to influential research. *Machine learning for Brain disorders*, 117-138.

Diebold, F., & Mariano, R. (1995). Predictive accuracy. *Journal of Business and Economic Statistics*, *13*(3), 253-263.

Dutta, A., & Das, D. (2022). Forecasting realized volatility: New evidence from time-varying jumps in VIX. *Journal of Futures Markets*, *42*(12), 2165-2189.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the econometric society*, 987-1007.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2018). Stock price prediction using support vector regression on daily and up to the minute prices. *The Journal of finance and data science*, *4*(3), 183-201.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Korkusuz, B., Kambouroudis, D., & McMillan, D. G. (2023). Do extreme range estimators improve realized volatility forecasts? Evidence from G7 stock markets. *Finance Research Letters*, 55, 103992.

Ma, W., Hong, Y., & Song, Y. (2024). On stock volatility forecasting under mixed-frequency data based on hybrid RR-MIDAS and CNN-LSTM models. *Mathematics*, *12*(10), 1538.

Michael, N., Cucuringu, M., & Howison, S. (2025). Options-driven volatility forecasting. *Quantitative Finance*, *25*(3), 443-470.

Ngwaba, C. A. (2025). HAR-RV-CARMA: A Kalman Filter-Weighted Hybrid Model for Enhanced Volatility Forecasting. *Risks*, *13*(11), 223.

Noh, S. H. (2021). Analysis of gradient vanishing of RNNs and performance comparison. *Information, 12*(11), 442.

Gunnarsson, E. S., Isern, H. R., Kaloudis, A., Risstad, M., Vigdel, B., & Westgaard, S. (2024). Prediction of realized volatility and implied volatility indices using AI and machine learning: A review. *International review of financial analysis*, *93*, 103221.

Jiao, Y., Guo, S., & Liu, Q. (2024). Implied volatility modeling and forecasting: evidence from China. *China Finance Review International*.

Lin, Y., Lin, Z., Liao, Y., Li, Y., Xu, J., & Yan, Y. (2022). Forecasting the realized volatility of stock price index: A hybrid model integrating CEEMDAN and LSTM. *Expert Systems with Applications*, *206*, 117736.

Psaradellis, I., & Sermpinis, G. (2016). Modelling and trading the US implied volatility indices. Evidence from the VIX, VXN and VXD indices. *International Journal of Forecasting*, *32*(4), 1268-1283.

Pascanu, R., Mikolov, T., & Bengio, Y. (2013, May). On the difficulty of training recurrent neural networks. In *International conference on machine learning* (pp. 1310-1318). Pmlr.

Song, Y., Tang, X., Wang, H., & Ma, Z. (2023). Volatility forecasting for stock market incorporating macroeconomic variables based on GARCH-MIDAS and deep learning models. *Journal of Forecasting*, *42*(1), 51-59.

Qiu, Z., Kownatzki, C., Scalzo, F., & Cha, E. S. (2025). Historical Perspectives in Volatility Forecasting Methods with Machine Learning. *Risks*, *13*(5), 98.

Qiu, Y., Zhang, X., Xie, T., & Zhao, S. (2019). Versatile HAR model for realized volatility: A least square model averaging perspective. *Journal of Management Science and Engineering*, *4*(1), 55-73.

Vrontos, S. D., Galakis, J., & Vrontos, I. D. (2021). Implied volatility directional forecasting: a machine learning approach. *Quantitative Finance*, *21*(10), 1687-1706.