Thesis or Antithesis? Synthesis! Using Ensembling Techniques To Combine Predictions of HAR and GARCH Models.

Andrey Sadchikov, Vladimir Yurchenko, and George Erykalkin

Moscow Institute of Physics and Technology, Moscow, Russia

Abstract. This paper investigates the use of ensembling techniques to enhance volatility forecasting in cryptocurrency markets by combining HAR and GARCH models. HAR-based models generally outperform GARCH(1,1) and naive models, though the choice of HAR specification is crucial. The main contribution is the novel use of stacking HAR, GARCH, and Naive models, significantly improving performance. In contrast, attempts to combine HAR and GARCH with ARIMA for residual prediction failed, as ARIMA couldn't capture the residuals' distribution. This study highlights the potential of stacking HAR and GARCH models, an underexplored area, and suggests further research into refining stacking methods and exploring advanced residual prediction models.

1 Introduction

Volatility forecasting is a cornerstone of financial econometrics, especially in dynamic and rapidly evolving markets like cryptocurrencies. The importance of volatility in determining option prices, Value at Risk (VaR), and the Sharpe ratio underscores its relevance to investors and risk managers. However, the extreme fluctuations and unpredictability of cryptocurrency prices present unique modeling challenges. Among the tools available, the Heterogeneous Autoregressive (HAR) model, first introduced by (6), offers a robust framework for capturing time-varying volatility across different time horizons.

This study evaluates the performance of HAR-type models in predicting the realized volatility of cryptocurrency returns, comparing their accuracy to GARCH-type models (2), naive benchmarks, and various combinations of these approaches.

Building upon our previous research, we employ a comprehensive methodology that integrates distributional analysis and rolling-window prediction techniques to assess model efficacy under diverse market conditions.

2 Historical Overview and Related Work

- **HAR Model (6), 2009**
  The Heterogeneous Autoregressive (HAR) model captures realized volatility

over daily, weekly, and monthly horizons. Its structure is:

$$\mathrm{RV}_t = \beta_0 + \beta_1 \mathrm{RV}_{t-1} + \beta_2 \mathrm{RV}^w_{t-1} + \beta_3 \mathrm{RV}^m_{t-1} + \varepsilon_t, \tag{1}$$

where RV_{t-1}, RV^w_{t-1}, and RV^m_{t-1} represent realized volatility over daily, weekly, and monthly timeframes, respectively.

- **Jump Components in Volatility Modeling**
  The jump component captures sudden, discrete, and extreme movements in asset prices, which are distinct from the smooth, continuous fluctuations typically modeled in financial time series. These components contribute to the fat tails of return distributions and significantly affect risk metrics like Value at Risk (VaR) and Expected Shortfall.
  The realized volatility (RV_t) can be decomposed into continuous (CV_t) and jump (J_t) components:

  $$\mathrm{RV}_t = \mathrm{CV}_t + \mathrm{J}_t,$$

  where:
  - CV_t: The continuous component, representing smooth price changes, often modeled using Brownian motion.
  - J_t: The jump component, capturing discrete price changes, modeled using a Poisson jump process.

  Estimation of Jump Components: Jumps are identified and estimated using high-frequency data:
  - **Bipower Variation (BPV):** A robust estimator of continuous volatility:

    $$\mathrm{BPV}_t = \sum_{i=2}^{n} |r_{t,i-1}| \cdot |r_{t,i}|,$$

    where $r_{t,i}$ are intraday returns.
  - **Threshold Method:** Returns exceeding a threshold (e.g., multiple of standard deviation) are classified as jumps.
  - **Integrated Volatility (IV):** Total variability is measured, and deviations from BPV are treated as jumps:

    $$\mathrm{J}_t = \mathrm{IV}_t - \mathrm{BPV}_t.$$

  Jumps are critical for capturing fat-tailed distributions, improving risk modeling, and enhancing the predictive performance of HAR variants.
- **HAR Variants and Related Models**
  - **HAR-CJ (1), 2007**
    Extends HAR by incorporating jump components:

    $$\mathrm{RV}_t = \beta_0 + \beta_1 \mathrm{RV}_{t-1} + \beta_2 \mathrm{RV}^w_{t-1} + \beta_3 \mathrm{RV}^m_{t-1} + \beta_4 \mathrm{J}_t + \varepsilon_t,$$

    where J_t represents the jump component.
  - **HAR-RV-C (5), 2012**
    Adds external covariates (C_t), such as macroeconomic indicators:

    $$\mathrm{RV}_t = \beta_0 + \beta_1 \mathrm{RV}_{t-1} + \beta_2 \mathrm{RV}^w_{t-1} + \beta_3 \mathrm{RV}^m_{t-1} + \beta_4 \mathrm{C}_t + \varepsilon_t.$$

- **HAR-Q Model (4), 2016**
The HAR-Q model extends the standard HAR model by incorporating the fourth moment (kurtosis) of the return distribution. This allows the model to capture the effects of heavy tails in volatility, improving its forecasting ability for extreme events and tail risk. The HARQ model can be specified as:

$$\begin{aligned}\mathrm{RV}_t &= \beta_0 + \beta_1 \mathrm{RV}_{t-1} + \beta_2 \mathrm{RV}^w_{t-1} + \beta_3 \mathrm{RV}^m_{t-1} \\ &\quad + \beta_4 \sqrt{\mathrm{RQ}_{t-1}} \cdot \mathrm{RV}_{t-1} + \beta_5 \sqrt{\mathrm{RQ}^w_{t-1}} \cdot \mathrm{RV}^w_{t-1} \\ &\quad + \beta_6 \sqrt{\mathrm{RQ}^m_{t-1}} \cdot \mathrm{RV}^m_{t-1} + \varepsilon_t\end{aligned}$$

- **HAR-Log (7), 2015**
Models the logarithm of realized volatility to handle heteroscedasticity:

$$\log(\mathrm{RV}_t) = \beta_0 + \beta_1 \log(\mathrm{RV}_{t-1}) + \beta_2 \log(\mathrm{RV}^w_{t-1}) + \beta_3 \log(\mathrm{RV}^m_{t-1}) + \varepsilon_t.$$

- **HAR-M (3), 2016**
Extends HAR to multivariate settings, capturing covariances between asset volatilities:

$$\mathrm{RV}_{t,i} = \beta_{0,i} + \beta_{1,i} \mathrm{RV}_{t-1,i} + \beta_{2,i} \mathrm{RV}^w_{t-1,i} + \beta_{3,i} \mathrm{RV}^m_{t-1,i} + \sum_{j \neq i} \gamma_{ij} \mathrm{RV}_{t,j} + \varepsilon_{t,i}.$$

3 Methodology

3.1 Data Collection

We obtained hourly closing price data for BTC-USD and ETH-USD from December 1, 2023, to December 1, 2024, using the yfinance API.

3.2 HAR Modeling

Building HAR and GARCH Predictions The HAR model uses daily, weekly, and monthly realized volatility values to predict the next value, as in Formula 1.
As GARCH modeling was the focus of our previous research, we will not delve into the details and only mention that the GARCH(p, q) model directly predicts the current volatility, represented as follows:

$$\widehat{\mathrm{RV}}_t^2 = \omega + \sum_{i=1}^{q} \alpha_i u^2_{t-1} + \sum_{j=1}^{p} \beta_j \tilde{\sigma}^2_{t-j} \tag{2}$$

Here, $\tilde{\sigma}^2_t$ denotes the conditional volatility at time t.
To prevent data leaks, the final prediction of the model was obtained using the expanding window method: the model was fitted on data from the moment $t = 0$ to $t - 1$ and then made an estimation $\widehat{\mathrm{RV}}_t$. The result of this process is shown in Figure 1.

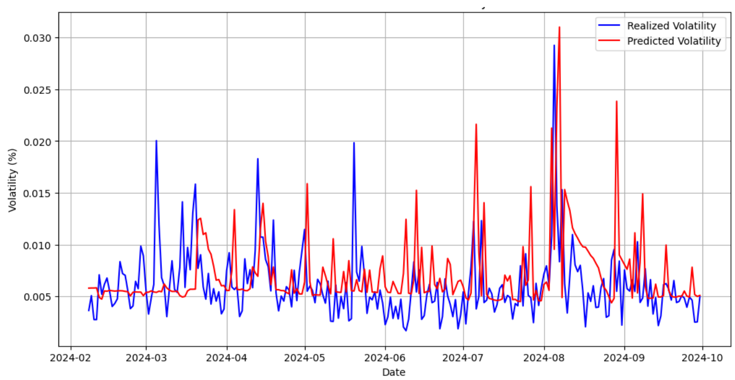

Fig. 1: Results of HAR Predictions

3.3 Evaluation

To estimate model performance, we computed the Mean Squared Error (MSE) and Mean Absolute Error (MAE) between the realized volatility values RV_t (i.e., ground truth) and GARCH predictions $\widehat{\mathrm{RV}}_t$.

$$\mathrm{MSE}(\{\mathrm{RV}_t\}_{t=1}^T, \{\widehat{\mathrm{RV}}_t\}_{t=1}^T) := \frac{1}{T}\sum_{t=1}^{T}(\mathrm{RV}_t - \widehat{\mathrm{RV}}_t)^2 \tag{3}$$

$$\mathrm{MAE}(\{\mathrm{RV}_t\}_{t=1}^T, \{\widehat{\mathrm{RV}}_t\}_{t=1}^T) := \frac{1}{T}\sum_{t=1}^{T}|\mathrm{RV}_t - \widehat{\mathrm{RV}}_t| \tag{4}$$

3.4 Realized Volatility Computation

To compute the realized volatility on day t, we calculated the daily standard deviation of hourly returns r_t^i, adjusting for the number of trading hours per day:

$$\text{Realized Volatility on day } t = \mathrm{RV}_t = \mathrm{Std}(r_t^1, \dots, r_t^{24}) \cdot \sqrt{24} \tag{5}$$

3.5 Model Comparison

We used the *volatility persistence model* as our benchmark naive model, which posits that future volatility can be predicted from its most recent value. Mathematically, this is expressed as:

$$\mathrm{RV}_{t+1} = \mathrm{RV}_t \tag{6}$$

In this equation, RV_{t+1} represents the expected future volatility, while RV_t is the current volatility. This model serves as a simple baseline against which we can measure the performance of more complex forecasting methods.

To effectively capture the difference between the performance of naive and advanced models, we introduce the Mean Scaled Squared Error (MSSE) and Mean Absolute Scaled Error (MASE) metrics:

$$\mathrm{MSSE}(\{\mathrm{RV}_t\}_{t=1}^T, \{\widehat{\mathrm{RV}}_t\}_{t=1}^T) := \frac{\sum_{t=2}^T (\mathrm{RV}_t - \widehat{\mathrm{RV}}_t)^2}{\sum_{t=2}^T (\mathrm{RV}_t - \mathrm{RV}_{t-1})^2} \tag{7}$$

$$\mathrm{MASE}(\{\mathrm{RV}_t\}_{t=1}^T, \{\widehat{\mathrm{RV}}_t\}_{t=1}^T) := \frac{\sum_{t=2}^T |\mathrm{RV}_t - \widehat{\mathrm{RV}}_t|}{\sum_{t=2}^T |\mathrm{RV}_t - \mathrm{RV}_{t-1}|} \tag{8}$$

Here, $\widehat{\mathrm{RV}}_t$ is the prediction from the advanced model.

4 Results

4.1 Autocorrelation Survey

We performed the Ljung-Box test with the purpose of determining the presence of autocorrelation for the error term $\varepsilon_t := \widehat{\mathrm{RV}}_t - \mathrm{RV}_t$. We present the results of the test in Table 1. We also present ACF and PACF plots in Figures 2 and 3.

Lag Number	**HAR**		**GARCH(1,1)**	
	LB-Statistic	**p-value**	**LB-Statistic**	**p-value**
1	14.03	0.0	0.01	0.904228
2	20.61	0.000033	1.43	0.489933
3	22.28	0.000057	16.09	0.001089
4	22.28	0.000176	16.32	0.002619
5	23.84	0.000236	16.46	0.005656

Table 1: Statistical Tests Results for HAR and GARCH models

As can be seen from these exhibits, strong autocorrelation exists in the residuals of the HAR model at all lags and in the residuals of the GARCH model at lags greater than 2. This potentially leaves room for improvement by creating a bagging-like model that utilizes an additional model (such as ARIMA or a more complex one) to predict the ε_t term.

4.2 Bagging with ARIMA

Both the HAR + ARIMA and GARCH + ARIMA combinations did not yield acceptable results because ARIMA could not successfully capture the underlying distribution of the ε_t term. This can be seen in Figure 4. Using a more sophisticated model instead of ARIMA seems like an interesting and promising experiment to the authors.

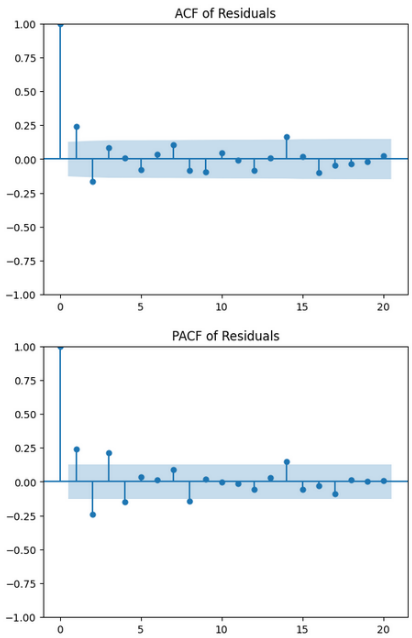

Fig. 2: Plots for HAR residuals

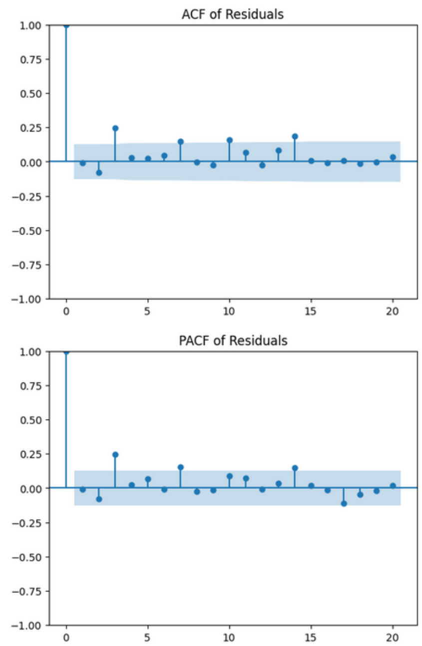

Fig. 3: Plots for GARCH residuals

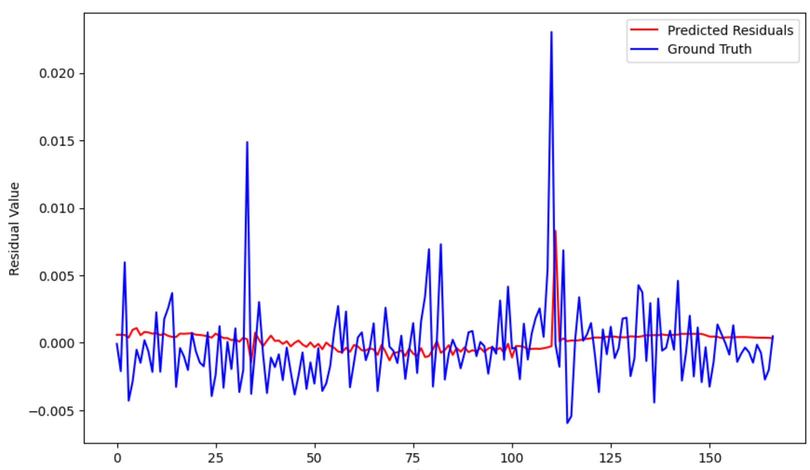

Fig. 4: ARIMA does not succeed in predicting the residual of RV

4.3 Ensembling HAR and GARCH

Due to the completely different nature of HAR and GARCH models, a natural idea emerges: combining their predictions into a new model. It was observed that the following estimates:

$$\widehat{\mathrm{RV}}_{\mathrm{Ensemble}_1} := \alpha_1 \cdot \widehat{\mathrm{RV}}_{\mathrm{HAR}} + \alpha_2 \cdot \widehat{\mathrm{RV}}_{\mathrm{GARCH}}$$

and

$$\widehat{\mathrm{RV}}_{\mathrm{Ensemble}_2} := \beta_1 \cdot \widehat{\mathrm{RV}}_{\mathrm{HAR}} + \beta_2 \cdot \widehat{\mathrm{RV}}_{\mathrm{GARCH}} + \beta_3 \cdot \widehat{\mathrm{RV}}_{\mathrm{Naive}}$$

sometimes outperform all single-model predictions. However, one should be careful with tuning the α's. In our experiments, we chose $\alpha_i := \frac{1}{2}$ and $\beta_i := \frac{1}{3}$.

4.4 Model Comparison

In order to maintain readability, we present only a few examples of model evaluation metrics in Tables 2 and 3. We also focused our attention on HAR-type models, since GARCH-based ones were already evaluated in previous work.

As can be seen from the Tables, both HAR and GARCH models typically outperform the naive volatility persistence model in our analysis. Among the various HAR specifications tested, the vanilla **HAR** and **HAR-J** models exhibited the best overall performance. However, it is important to note that the differences in performance among the various HAR and GARCH models were quite large, typically differing by up to 30%. Notably, the GARCH model failed when predicting the ETH-USD data, but ensembles with GARCH performed competitively. In the case of BTC-USD data, $\mathrm{Ensemble}_2$ significantly outperformed all other models.

Table 2: Performance Metrics for Model Evaluation on BTC-USD data

Model	$\mathbf{MSSE}^{-1}$	$\mathbf{MAE}^{-1}$
Naive	1.0000	1.0000
HAR	1.1499	1.1200
HAR-Q	1.0029	1.0382
HAR-J	1.1298	1.1124
GARCH	1.0517	1.0198
$\mathrm{Ensemble}_1$	1.1659	1.1034
$\mathrm{Ensemble}_2$	1.3380	1.1796

Table 3: Performance Metrics for Model Evaluation on ETH-USD data

Model	$\mathbf{MSE}^{-1}$	$\mathbf{MAE}^{-1}$
Naive	1.0000	1.0000
HAR	1.2038	1.1445
HAR-Q	1.1969	1.1080
HAR-J	1.2495	1.1709
GARCH	0.6564	0.8563
$\mathrm{Ensemble}_1$	1.0233	1.0248
$\mathrm{Ensemble}_2$	1.1947	1.1083

5 Conclusion

HAR-based models often outperform GARCH(1,1) model and always outperform Naive model, except for extreme cases. However, It is imporant to choose HAR model specification (HAR-J or HAR-Q) carefully for particular task.

However, the key result of the work is the survey of various combination of models, such as stacking and bagging.

- Stacking: Combining HAR, GARCH, and, surprisingly, Naive models into simple ensembles by averaging, such as Ensemble_1 and Ensemble_2, demonstrated significant improvements in volatility forecasting accuracy, with stacking ensemble outperforming all other models.
- Bagging: Attempts to combine HAR and GARCH models with ARIMA for residual prediction did not yield satisfactory results, as ARIMA failed to capture the underlying distribution of residuals (ε_t). These results suggest that while stacking offers promising improvements, bagging using ARIMA may not be the best choice for predicting residuals in this context.

The authors suggest that future research should focus on employing more advanced models for bagging and refining the weight optimization in stacking ensembles. Stacking, in particular, appears to be a promising approach, consistently demonstrating exceptional performance.

Bibliography

[1] Andersen, T.G., Bollerslev, T., Diebold, F.X.: Roughing it Up: Including Jump Components in the Measurement, Modeling and Forecasting of Return Volatility. NBER Working Papers 11775, National Bureau of Economic Research, Inc (Nov 2005), https://ideas.repec.org/p/nbr/nberwo/11775.html

[2] Bollerslev, T.: Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics **31**(3), 307–327 (1986). https://doi.org/https://doi.org/10.1016/0304-4076(86)90063-1, https://www.sciencedirect.com/science/article/pii/0304407686900631

[3] Bollerslev, T., Patton, A., Quaedvlieg, R.: Exploiting the errors: A simple approach for improved volatility forecasting. Journal of Econometrics **192** (11 2015). https://doi.org/10.1016/j.jeconom.2015.10.007

[4] Bollerslev, T., Patton, A., Quaedvlieg, R.: Exploiting the errors: A simple approach for improved volatility forecasting. Journal of Econometrics **192**(1), 1–18 (2016), https://EconPapers.repec.org/RePEc:eee:econom:v:192:y:2016:i:1:p:1-18

[5] Corsi, F., Audrino, F., Reno, R.: Har modeling for realized volatility forecasting. In: Handbook of Volatility Models and Their Applications, pp. 363 – 382. John Wiley & Sons, Inc, New Jersey, USA (2012). https://doi.org/10.1002/9781118272039.ch15, https://openaccess.city.ac.uk/id/eprint/4438/

[6] Corsi, F.: A simple approximate long-memory model of realized volatility. Journal of Financial Econometrics **7**(2), 174–196 (02 2009). https://doi.org/10.1093/jjfinec/nbp001, https://doi.org/10.1093/jjfinec/nbp001

[7] Sheppard, K., Patton, A.: Good volatility, bad volatility: Signed jumps and the persistence of volatility. Review of Economics and Statistics **97** (10 2011). https://doi.org/10.1162/REST_a_00503

6 Appendix

All the supporting code can be found at GitHub.