Realized Volatility Forecasting with ML

Thesis Advisor: Prof. Dacheng Xiu

Yichen Ji

Department of Statistics
University of Chicago

May 1st, 2024

RV: a volatility measure using high-frequency data

$$RV_t = \sum_{i=1}^{M} r_{t,i}^2$$

where $r_{t,i}$ is the log return over the ith intraday period on day t.

- RV: Why interested?

RV: a volatility measure using high-frequency data

$$RV_t = \sum_{i=1}^{M} r_{t,i}^2$$

where $r_{t,i}$ is the log return over the ith intraday period on day t.

- RV: Why interested?
  - Non-parametric, model-free, more granular return variability

RV: a volatility measure using high-frequency data

$$RV_t = \sum_{i=1}^{M} r_{t,i}^2$$

where $r_{t,i}$ is the log return over the ith intraday period on day t.

- RV: Why interested?
  - Non-parametric, model-free, more granular return variability
  - Practically useful in options pricing, trading, and risk management e.g. Optiver RV Prediction Kaggle Competition

RV: a volatility measure using high-frequency data

$$RV_t = \sum_{i=1}^{M} r_{t,i}^2$$

where $r_{t,i}$ is the log return over the ith intraday period on day t.

- RV: Why interested?
  - Non-parametric, model-free, more granular return variability
  - Practically useful in options pricing, trading, and risk management e.g. Optiver RV Prediction Kaggle Competition
- Forecasting: How feasible?

RV: a volatility measure using high-frequency data

$$RV_t = \sum_{i=1}^{M} r_{t,i}^2$$

where $r_{t,i}$ is the log return over the ith intraday period on day t.

- RV: Why interested?
  - Non-parametric, model-free, more granular return variability
  - Practically useful in options pricing, trading, and risk management e.g. Optiver RV Prediction Kaggle Competition
- Forecasting: How feasible?
  - Availability of high-frequency intraday data

RV: a volatility measure using high-frequency data

$$RV_t = \sum_{i=1}^{M} r_{t,i}^2$$

where $r_{t,i}$ is the log return over the ith intraday period on day t.

- RV: Why interested?
  - Non-parametric, model-free, more granular return variability
  - Practically useful in options pricing, trading, and risk management e.g. Optiver RV Prediction Kaggle Competition
- Forecasting: How feasible?
  - Availability of high-frequency intraday data
  - Stylized facts → high signal-to-noise ratio compared to returns

RV: a volatility measure using high-frequency data

$$RV_t = \sum_{i=1}^{M} r_{t,i}^2$$

where $r_{t,i}$ is the log return over the ith intraday period on day t.

- RV: Why interested?
  - Non-parametric, model-free, more granular return variability
  - Practically useful in options pricing, trading, and risk management e.g. Optiver RV Prediction Kaggle Competition
- Forecasting: How feasible?
  - Availability of high-frequency intraday data
  - Stylized facts → high signal-to-noise ratio compared to returns
  - e.g. clustering, (local) mean-reverting, asymmetry, etc.

RV: a volatility measure using high-frequency data

$$RV_t = \sum_{i=1}^{M} r_{t,i}^2$$

where $r_{t,i}$ is the log return over the ith intraday period on day t.

- RV: Why interested?
  - Non-parametric, model-free, more granular return variability
  - Practically useful in options pricing, trading, and risk management e.g. Optiver RV Prediction Kaggle Competition
- Forecasting: How feasible?
  - Availability of high-frequency intraday data
  - Stylized facts $\rightarrow$ high signal-to-noise ratio compared to returns
  - e.g. clustering, (local) mean-reverting, asymmetry, etc.
- Machine learning: What potential? [KX+23]

RV: a volatility measure using high-frequency data

$$RV_t = \sum_{i=1}^{M} r_{t,i}^2$$

where $r_{t,i}$ is the log return over the ith intraday period on day t.

- RV: Why interested?
  - Non-parametric, model-free, more granular return variability
  - Practically useful in options pricing, trading, and risk management e.g. Optiver RV Prediction Kaggle Competition
- Forecasting: How feasible?
  - Availability of high-frequency intraday data
  - Stylized facts → high signal-to-noise ratio compared to returns
  - e.g. clustering, (local) mean-reverting, asymmetry, etc.
- Machine learning: What potential? [KX⁺23]
  - Presence of large conditioning panel information sets

RV: a volatility measure using high-frequency data

$$RV_t = \sum_{i=1}^{M} r_{t,i}^2$$

where $r_{t,i}$ is the log return over the ith intraday period on day t.

- RV: Why interested?
  - Non-parametric, model-free, more granular return variability
  - Practically useful in options pricing, trading, and risk management e.g. Optiver RV Prediction Kaggle Competition
- Forecasting: How feasible?
  - Availability of high-frequency intraday data
  - Stylized facts $\rightarrow$ high signal-to-noise ratio compared to returns
  - e.g. clustering, (local) mean-reverting, asymmetry, etc.
- Machine learning: What potential? [KX+23]
  - Presence of large conditioning panel information sets
  - Ambiguous functional forms

Two Questions

- **Q**: What information do market participants have and how do they use it?

Two Questions

- **Q**: What information do market participants have and how do they use it?
- **A**: We don't know, but machine learning models may uncover some complex patterns given adequate panel data and functional complexity. [KMZ24]

Two Questions

- **Q**: What information do market participants have and how do they use it?
- **A**: We don't know, but machine learning models may uncover some complex patterns given adequate panel data and functional complexity. [KMZ24]
- **Q**: Which of the many economic models available in the literature should we impose?

Two Questions

- **Q**: What information do market participants have and how do they use it?
- **A**: We don't know, but machine learning models may uncover some complex patterns given adequate panel data and functional complexity. [KMZ24]
- **Q**: Which of the many economic models available in the literature should we impose?
- **A**: Apply and compare the performance of each of its methods in familiar empirical problems. [GKX20]

Two Questions

- **Q**: What information do market participants have and how do they use it?
- **A**: We don't know, but machine learning models may uncover some complex patterns given adequate panel data and functional complexity. [KMZ24]
- **Q**: Which of the many economic models available in the literature should we impose?
- **A**: Apply and compare the performance of each of its methods in familiar empirical problems. [GKX20]
- **Objective**: Compare the out-of-sample predictive performance of machine learning models against structural time-series econometric models

"A good volatility model must be able to forecast volatility." [EP07]

- OLS-based models: HAR [Cor09], MIDAS [GSCV06], SHAR [PS15], HARQ [BPQ16], HEXP [BHHP18]
- Attempts using ML models: LASSO [AK16], random forest [LD18], feed-forward neural networks (FFNN) and recurrent neural networks (RNN) [Buc20], convolutional neural networks (CNN) [RBH22]
- Comparative analysis: [RP20], [LT22], [CSV23], [ZZCQ24]
- Robust realized measures: [BNS06, ZMAS05, Zha06, BNHLS08, PV09, ADS12, DX21],
- ML $\cap$ (Economics $\cup$ Finance): [A^{+}18, GKX20, GKX22, KMZ24, CPZ24]

Model Overview

Econometrics	Machine Learning
HAR	LASSO
MIDAS	Principal Component Regression (PCR)
SHAR	Random Forest (RF)
HARQ	Gradient Boosting Regression Tree (GBRT)
HEXP	Feed-forward Neural Networks (FFNN)

Quadratic Variation Theory

Assume the log price p_t within the active part of a trading day t follows a continuous semimartingale of the form:

$$p_t = \int_{t-1}^{t} \mu_s ds + \int_{t-1}^{t} \sigma_s dW_s$$

The quadratic variation (QV) of this log-price process, after some derivation, is:

$$QV_t = [p, p]_t = \int_{t-1}^{t} \sigma_s^2 ds$$

The true unobservable volatility construct that integrates the instantaneous volatility over time is called integrated volatility (IV):

$$IV_t = \int_{t-1}^{t} \sigma_s^2 ds$$

$QV_t = IV_t$ (without jumps)! Heads-up: Such nice coincidence doesn't happen in general e.g. jump-diffusion process.)

Consistency & Asymptotic Theory [BNS02]

Since we can only observe intraday price observations in discrete time...

$$RV_t = \sum_{i=1}^{M} r_{t,i}^2 \xrightarrow{p} IV_t$$

Moreover, the semimartingale theory provides CLT:

$$\sqrt{M}\left(\frac{RV_t - IV_t}{\sqrt{2IQ_t}}\right) \xrightarrow{d} N(0,1)$$

where $IQ_t = \int_{t-1}^{t} \sigma_s^4 ds$ denotes *integrated quarticity*, which is independent of the limiting Gaussian distribution and can be consistently estimated by the *realized quarticity* (RQ) statistic:

$$RQ_t = \frac{M}{3}\sum_{i=1}^{M} r_{t,i}^4 \xrightarrow{p} IQ_t$$

HAR [Cor09]

$$RV_t = \beta_0 + \beta_d RV^d_{t-1} + \beta_w RV^w_{t-1} + \beta_m RV^m_{t-1} + \beta_q RV^q_{t-1} + \epsilon_t$$

where $RV^l_{t-1} = \frac{1}{l}\sum_{i=1}^{l} RV_{t-i}, l = \{1, 5, 22, 63\}$ is the simple average of daily RVs over different lag horizons (daily, weekly, monthly, quarterly, respectively), and $\{\epsilon_t\}_t$ is a zero mean innovation process.

- Simple, parsimonious, easy to implement
- Serve as the benchmark model

MIDAS [GSCV06]

$$RV_t = \beta_0 + \beta_1 MIDAS_{t-1} + \epsilon_t,$$

$$MIDAS_t = \frac{1}{\sum_{i=1}^{L} a_i} \sum_{i=0}^{L} a_{i+1} RV_{t-i}$$

$$a_i = \left(\frac{i}{L}\right)^{\theta_1 - 1} \left(1 - \frac{i}{L}\right)^{\theta_2 - 1} \frac{\Gamma(\theta_1 + \theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)}, i = 1, ..., L$$

- Smoothly weighted moving average of lagged daily RVs
- Parametrize the coefficients/weights in a beta polynomial form

SHAR [PS15]

$$
\begin{aligned}
RV_t &= \beta_0 + \beta_d^+ RS_{t-1}^{d+} + \beta_d^- RS_{t-1}^{d-} \\
&\quad + \beta_w RV_{t-1}^w + \beta_m RV_{t-1}^m + \beta_q RV_{t-1}^q + \epsilon_t, \\
RS_t^+ &= \sum_{i=1}^{M} r_{t,i}^2 \mathbb{I}\left\{r_{t,i} > 0\right\}, RS_t^- = \sum_{i=1}^{M} r_{t,i}^2 \mathbb{I}\left\{r_{t,i} < 0\right\}.
\end{aligned}
$$

- Leverage realized semivariance (RS) estimator by [BNKS08]
- [PS15] found that the negative RS has more predictive power than its positive counterpart.

HARQ [BPQ16]

$$RV_t = IV_t + \eta_t, \eta_t \sim N(0, 2\Delta IQ_t)$$

$$RV_t = \beta_0 + (\beta_d + \phi_d \sqrt{RQ_{t-1}^d}) RV_{t-1}^d + (\beta_w + \phi_w \sqrt{RQ_{t-1}^w}) RV_{t-1}^w$$
$$+ (\beta_m + \phi_m \sqrt{RQ_{t-1}^m}) RV_{t-1}^m + (\beta_q + \phi_q \sqrt{RQ_{t-1}^q}) RV_{t-1}^q + \epsilon_t$$

- Exploit the heteroskedasticity in the measurement error η_t
- Compensate for uncertainty in RV measurements: low variance in measurement errors offers a stronger predictive signal

HEXP [BHHP18]

$$RV_t = \beta_0 + \beta_1 ExpRV_{t-1}^{1} + \beta_5 ExpRV_{t-1}^{5} + \beta_{25} ExpRV_{t-1}^{25} + \beta_{125} ExpRV_{t-1}^{125} + \epsilon_t,$$

$$ExpRV_t^{\mathrm{CoM}(\lambda)} = \sum_{i=1}^{500} \frac{e^{-i\lambda}}{e^{-\lambda} + e^{-2\lambda} + \cdots + e^{-500\lambda}} RV_{t-i+1}$$

$$\mathrm{CoM}(\lambda) = \frac{\sum_{t=0}^{\infty} e^{-\lambda t} t}{\sum_{t=0}^{\infty} e^{-\lambda t}} = \frac{e^{-\lambda}}{1 - e^{-\lambda}}$$

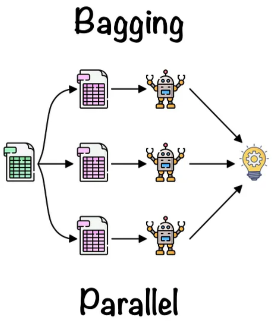
- *CoM* Center of Mass, defined as the weighted average period for the lags used; λ decay rate
- Use a mixture of exponentially weighted moving averages (EWMA) of lagged daily RVs as regressors

Linear Models: LASSO & PCR

- LASSO: sparsity, variable selection
- PCR: dimension reduction, but forms PCs before the forecasting step

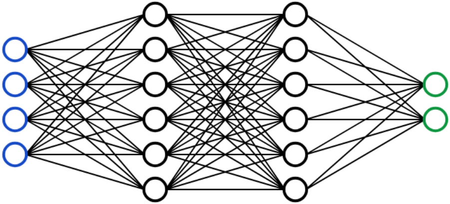
Tree-based Models: RF & GBRT

- RF (bagging): averaged over forecasts of separate trees trained on bootstrapped samples
- GBRT (boosting): each tree fitted on the residual errors of the preceding tree, correcting what earlier predictors don't capture



Feed-forward Neural Networks

- "Universal approximators" with layered structure
- Require large-scale training data and compute with engineering optimization and tuning tricks to success



Data and Variables

Data:

- 2 universes: 1000 S&P 500 stocks and 10014 U.S. stocks
- Sample period: January 1996 - December 2022 (27 years)
- Data source:
  - 1-min price observations from TAQ
  - options implied volatility data from OptionMetrics
  - overnight return and trading volume data from CRSP
- Collect call and put options with maturities ranging from $1, 2, 3$ months and absolute delta equal to $0.1, 0.15, ..., 0.9$

Features & Response Variable:

- 5-minute sampling frequency for intraday returns
- 122 features in total (15 realized + 102 implied + 4 price volume + 1 intercept)
- Response Variable: next-day RV (in logs)

Response Variable - S&P 500

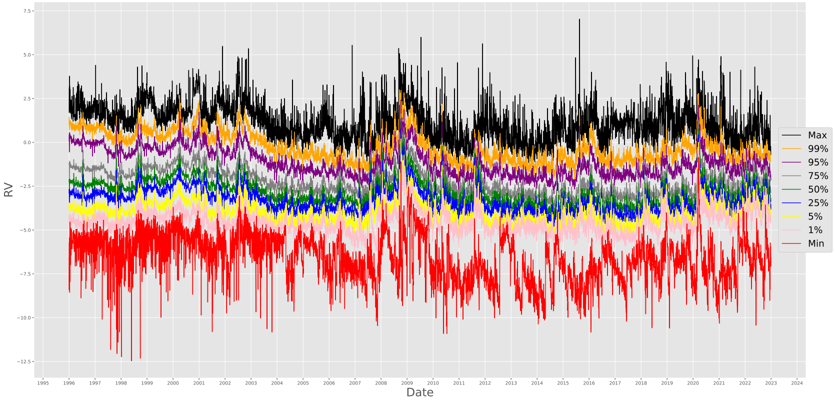


Figure 1: maximum, minimum, $99^{th}, 95^{th}, 75^{th}, 50^{th}, 25^{th}, 5^{th}, and 1^{st}$ percentiles of daily RV in log-scale for stocks in the S&P 500 universe from 1996 to 2022.

Response Variable - U.S. Stocks

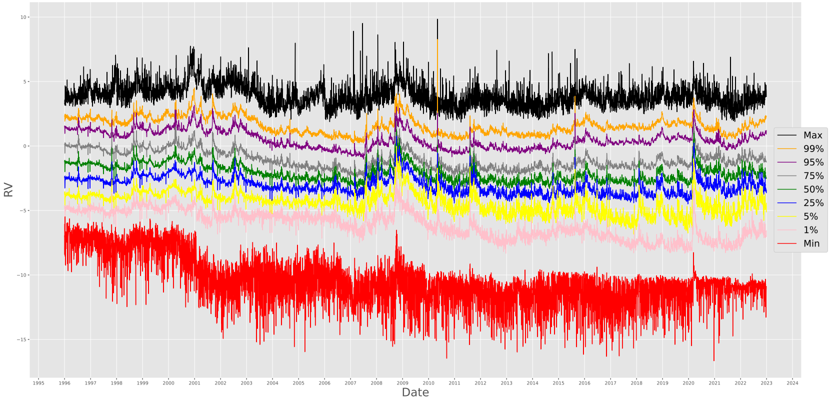
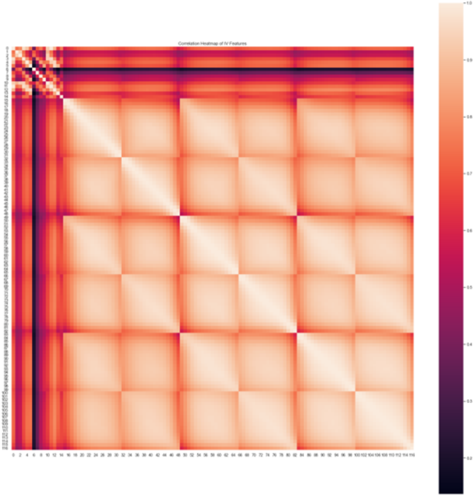


Figure 2: maximum, minimum, $99^{th}, 95^{th}, 75^{th}, 50^{th}, 25^{th}, 5^{th}, and 1^{st}$ percentiles of daily RV in log-scale for stocks in the U.S. stock universe from 1996 to 2022.

Feature Correlation Heatmap



Training Scheme & Evaluation Metrics

Training scheme:

- Rolling window: 5 training years + 1 validation year+ 1 test year
- Panel/pooled fitting

Evaluation metrics:

- R^2: $1 - \frac{\sum_{i,t}\left(RV_{i,t} - \widehat{RV_{i,t}}^{m}\right)^2}{\sum_{i,t}\left(RV_{i,t} - \widehat{RV_{i,t}}^{\text{benchmark}}\right)^2}$,
- Mean squared error (MSE):
  $\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\#\mathcal{T}_{\text{test}}}\sum_{t\in\mathcal{T}_{\text{test}}}\left(\mathrm{RV}_{i,t} - \widehat{\mathrm{RV}}_{i,t}\right)^2$
- Quasi-likelihood (QLIKE):
  $\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\#\mathcal{T}_{\text{test}}}\sum_{t\in\mathcal{T}_{\text{test}}}\left[\frac{\exp(\mathrm{RV}_{i,t})}{\exp(\widehat{\mathrm{RV}}_{i,t})} - \left(\mathrm{RV}_{i,t} - \widehat{\mathrm{RV}}_{i,t}\right) - 1\right]$

OOS Performance - S&P 500 ($* = 99.99^{th}$ percentile winsorized)

Model	R2	MSE	MSE*	QLike	QLike*
HAR	0.7052	0.3970	0.3962	0.4039	0.3737
MIDAS	0.6995	0.4047	0.4039	0.4018	0.3729
SHAR	0.7057	0.3963	0.3955	0.4029	0.3735
HARQ	0.7187	0.3787	0.3780	0.3912	0.3601
HEXP	0.7071	0.3944	0.3936	0.4015	0.3721
OLSRM	0.7201	0.3768	0.3761	0.3880	0.3583
OLSRM4	0.7202	0.3768	0.3761	0.3874	0.3578
OLSIV	0.6096	0.5257	0.5248	0.4471	0.4128
OLSALL	0.7276	0.3668	0.3660	0.3673	0.3366
LASSO	0.7276	0.3668	0.3661	0.3667	0.3368
PCR	0.7216	0.3748	0.3740	0.3734	0.3416
RF	0.7204	0.3765	0.3758	0.3681	0.3373
GBRT	0.7068	0.3948	0.3941	0.3854	0.3567
NN	**0.7321**	**0.3607**	**0.3599**	**0.3576**	**0.3245**

Table 1: **OOS Forecasting Performance, S&P 500 Stocks**

OOS Performance - U.S. stocks

Model	R2	MSE	MSE*	QLike	QLike*
HAR	0.7849	0.5708	0.5680	1.5628	0.4886
MIDAS	0.7815	0.5798	0.5771	1.4024	0.4914
SHAR	0.7850	0.5706	0.5678	1.6462	0.4883
HARQ	0.7884	0.5615	0.5587	1.6487	0.4864
HEXP	0.7863	0.5670	0.5643	1.4541	0.4827
OLSRM	0.7897	0.5580	0.5552	1.7167	0.4819
OLSRM4	0.7898	0.5578	0.5550	1.6645	0.4817
OLSIV	0.5109	1.2980	1.2951	1.4857	1.0282
OLSALL	0.7906	0.5557	0.5529	1.5204	0.4758
LASSO	0.7904	0.5563	0.5535	1.5025	0.4758
PCR	0.7861	0.5675	0.5647	1.3597	0.4781
RF	0.7905	0.5561	0.5533	1.2245	0.4594
GBRT	0.7756	0.5954	0.5926	**1.1476**	0.4855
NN	**0.7954**	**0.5428**	**0.5400**	1.4290	**0.4509**

Table 2: **OOS Forecasting Performance, US Stocks**

OLS Individual v.s. Pooled Fit - S&P 500

Model	R2		MSE*		QLike*	
	Individual	Pooled	Individual	Pooled	Individual	Pooled
HAR	0.6833	0.7052	0.4253	0.3962	0.4305	0.3737
MIDAS	**0.6907**	0.6995	**0.4158**	0.4039	**0.3798**	0.3729
SHAR	0.6834	0.7057	0.4252	0.3955	0.4335	0.3735
HARQ	0.6775	0.7187	0.4332	0.3780	0.5024	0.3601
HEXP	0.6693	0.7071	0.4442	0.3936	0.4701	0.3721
OLSRM	0.6734	0.7201	0.4383	0.3761	0.4894	0.3583
OLSRM4	0.6654	0.7202	0.4492	0.3761	0.5145	0.3578
OLSIV	0.4551	0.6096	0.7317	0.5248	0.8039	0.4128
OLSALL	0.5514	**0.7276**	0.6019	**0.3660**	0.7744	**0.3366**

Table 3: **Individual vs Pooled Fit, S&P 500 Stocks**

OLS Individual v.s. Pooled Fit - U.S. stocks

Model	R2		MSE*		QLike*	
	Individual	Pooled	Individual	Pooled	Individual	Pooled
HAR	0.6991	0.7849	0.7953	0.5680	6.1815	0.4886
MIDAS	**0.7434**	0.7815	**0.6777**	0.5771	**0.6883**	0.4914
SHAR	0.6992	0.7850	0.7947	0.5678	5.8682	0.4883
HARQ	0.6379	0.7884	0.9581	0.5587	26.9071	0.4864
HEXP	0.6427	0.7863	0.9452	0.5643	22.3139	0.4827
OLSRM	0.6032	0.7897	1.0501	0.5552	32.1466	0.4819
OLSRM4	0.5933	0.7898	1.0762	0.5550	36.0054	0.4817
OLSIV	0.3112	0.5109	1.8256	1.2951	45.2375	1.0282
OLSALL	0.4051	**0.7906**	1.5765	**0.5529**	64.0068	**0.4758**

Table 4: **Individual vs Pooled Fit, U.S. Stocks**

Diebold-Mariano (DM) Test - S&P 500

Model	HAR	MIDAS	SHAR	HARQ	HEXP	OLSRM	OLSRM4	OLSIV	OLSALL	LA
MIDAS	-50.2	-	-	-	-	-	-	-	-	
SHAR	28.0	55.6	-	-	-	-	-	-	-	
HARQ	135.5	153.6	131.0	-	-	-	-	-	-	
HEXP	43.5	72.5	28.5	-109.6	-	-	-	-	-	
OLSRM	141.6	169.0	142.6	34.5	126.7	-	-	-	-	
OLSRM4	139.9	167.4	140.9	33.2	125.1	1.3	-	-	-	
OLSIV	-27.4	-25.8	-27.6	-31.4	-28.0	-31.8	-31.8	-	-	
OLSALL	114.8	145.3	113.4	59.3	107.8	52.1	52.3	34.6	-	
LASSO	112.8	141.4	111.2	56.7	106.3	49.0	48.9	34.6	-1.0	
PCR	81.8	110.2	79.4	16.1	76.0	8.5	8.4	32.9	-82.8	-9
RF	83.0	112.1	80.0	10.2	72.4	1.4	1.2	32.0	-43.6	-4
GBRT	4.0	18.0	2.7	-30.5	-0.9	-34.3	-34.4	28.9	-60.3	-6
NN	131.2	159.7	129.2	84.1	120.7	74.7	73.7	35.4	28.6	2

Table 5: **Diebold-Mariano Test, S&P 500 Stocks**

Diebold-Mariano (DM) Test - U.S. Stocks

Model	HAR	MIDAS	SHAR	HARQ	HEXP	OLSRM	OLSRM4	OLSIV	OLSALL	LA
MIDAS	-96.8	-	-	-	-	-	-	-	-	
SHAR	19.2	99.6	-	-	-	-	-	-	-	
HARQ	138.8	197.1	137.7	-	-	-	-	-	-	
HEXP	109.5	145.9	96.4	-75.9	-	-	-	-	-	
OLSRM	167.5	242.2	170.8	100.2	125.8	-	-	-	-	
OLSRM4	143.9	218.4	145.1	61.7	106.8	4.4	-	-	-	
OLSIV	-103.9	-102.7	-104.0	-105.4	-104.4	-105.9	-105.9	-	-	
OLSALL	156.0	236.7	156.8	89.1	121.3	40.2	55.2	106.5	-	
LASSO	177.5	248.7	178.6	101.6	139.3	38.5	24.2	106.5	-11.7	
PCR	55.3	119.0	50.4	-66.3	-8.1	-105.6	-95.7	104.8	-124.9	-14
RF	80.6	135.8	79.2	30.0	61.4	10.9	9.5	106.6	-1.9	1
GBRT	-41.2	-25.9	-41.5	-56.7	-47.6	-62.6	-62.9	104.7	-67.5	-6
NN	197.4	278.7	194.9	134.4	176.5	110.1	108.7	108.1	101.5	10

Table 6: **Diebold-Mariano Test, U.S. Stocks**

1. Motivation
2. Literature Review
3. Methodology
4. Empirical Findings
5. **Conclusion**
6. References

Conclusion & Discussion

Empirical Conclusion: Shallow neural networks deliver superior out-of-sample predictive performance compared to existing OLS-based regression models.

Discussion:

- Inclusion of jumps and microstructure noise consideration
- How to impose economic structure based on domain knowledge of economic and finance theory
- Economic gain and implications from machine learning volatility forecast & real-world execution
- Engineering optimization tricks v.s. interpretability for more complex network architecture

Possible Future Directions

- Try jump-robust and microstructure noise-robust estimators as features
- Tweak nonlinear models to focus on stocks with lower arbitrage and transaction costs
- Tailed machine learning model and network architecture design

[A+18] Susan Athey et al.
The impact of machine learning on economics.
The economics of artificial intelligence: An agenda, pages 507–547, 2018.

[ADS12] Torben G Andersen, Dobrislav Dobrev, and Ernst Schaumburg.
Jump-robust volatility estimation using nearest neighbor truncation.
Journal of Econometrics, 169(1):75–93, 2012.

[AK16] Francesco Audrino and Simon D Knaus.
Lassoing the har model: A model selection perspective on realized volatility dynamics.
Econometric Reviews, 35(8-10):1485–1521, 2016.

[BHHP18] Tim Bollerslev, Benjamin Hood, John Huss, and Lasse Heje Pedersen.
Risk everywhere: Modeling and managing volatility.
The Review of Financial Studies, 31(7):2729–2773, 2018.

[BNHLS08] Ole E Barndorff-Nielsen, Peter Reinhard Hansen, Asger Lunde, and Neil Shephard.
Designing realized kernels to measure the ex post variation of equity prices in the presence of noise.
Econometrica, 76(6):1481–1536, 2008.

[BNKS08] Ole E Barndorff-Nielsen, Silja Kinnebrock, and Neil Shephard.
Measuring downside risk-realised semivariance.
CREATES Research Paper, (2008-42), 2008.

[BNS02] Ole E Barndorff-Nielsen and Neil Shephard.
Econometric analysis of realized volatility and its use in estimating stochastic volatility models.
Journal of the Royal Statistical Society Series B: Statistical Methodology, 64(2):253–280, 2002.

[BNS06] Ole E Barndorff-Nielsen and Neil Shephard.
Econometrics of testing for jumps in financial economics using bipower variation.
Journal of financial Econometrics, 4(1):1–30, 2006.

[BPQ16] Tim Bollerslev, Andrew J Patton, and Rogier Quaedvlieg.
Exploiting the errors: A simple approach for improved volatility forecasting.
Journal of Econometrics, 192(1):1–18, 2016.

[Buc20] Andrea Bucci.
Realized volatility forecasting with neural networks.
Journal of Financial Econometrics, 18(3):502–531, 2020.

[Cor09] Fulvio Corsi.
A simple approximate long-memory model of realized volatility.
Journal of Financial Econometrics, 7(2):174–196, 2009.

[CPZ24] Luyang Chen, Markus Pelger, and Jason Zhu.
Deep learning in asset pricing.
Management Science, 70(2):714–750, 2024.

[CSV23] Kim Christensen, Mathias Siggaard, and Bezirgen Veliyev.
A machine learning approach to volatility forecasting.
Journal of Financial Econometrics, 21(5):1680–1727, 2023.

[DX21] Rui Da and Dacheng Xiu.
When moving-average models meet high-frequency data: Uniform inference on volatility.
Econometrica, 89(6):2787–2825, 2021.

[EP07] Robert F Engle and Andrew J Patton.
What good is a volatility model?
In *Forecasting volatility in the financial markets*, pages 47–63. Elsevier, 2007.

[GKX20] Shihao Gu, Bryan Kelly, and Dacheng Xiu.
Empirical asset pricing via machine learning.
The Review of Financial Studies, 33(5):2223–2273, 2020.

[GKX22] Stefano Giglio, Bryan Kelly, and Dacheng Xiu.
Factor models, machine learning, and asset pricing.
Annual Review of Financial Economics, 14:337–368, 2022.

[GSCV06] Eric Ghysels, Pedro Santa-Clara, and Rossen Valkanov.
Predicting volatility: getting the most out of return data sampled at different frequencies.
Journal of Econometrics, 131(1-2):59–95, 2006.

[KMZ24] Bryan Kelly, Semyon Malamud, and Kangying Zhou.
The virtue of complexity in return prediction.
The Journal of Finance, 79(1):459–503, 2024.

[KX+23] Bryan Kelly, Dacheng Xiu, et al.
Financial machine learning.
Foundations and Trends® in Finance, 13(3-4):205–363, 2023.

[LD18] Chuong Luong and Nikolai Dokuchaev.
Forecasting of realised volatility with the random forests algorithm.
Journal of Risk and Financial Management, 11(4):61, 2018.

[LT22] Sophia Zhengzi Li and Yushan Tang.
Automated risk forecasting.
In *Automated Risk Forecasting: Li, Sophia Zhengzi| Tang, Yushan*. [Sl]: SSRN, 2022.

[PS15] Andrew J Patton and Kevin Sheppard.
Good volatility, bad volatility: Signed jumps and the persistence of volatility.
Review of Economics and Statistics, 97(3):683–697, 2015.

[PV09] Mark Podolskij and Mathias Vetter.
Bipower-type estimation in a noisy diffusion setting.
Stochastic processes and their applications, 119(9):2803–2831, 2009.

[RBH22] Rafael Reisenhofer, Xandro Bayer, and Nikolaus Hautsch.
Harnet: A convolutional neural network for realized volatility forecasting.
arXiv preprint arXiv:2205.07719, 2022.

[RP20] Eghbal Rahimikia and Ser-Huang Poon.
Machine learning for realised volatility forecasting.
Available at SSRN, 3707796, 2020.

[Zha06] Lan Zhang.
Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach.
Bernoulli, 12(6):1019–1043, 2006.

[ZMAS05] Lan Zhang, Per A Mykland, and Yacine Aït-Sahalia.
A tale of two time scales: Determining integrated volatility with noisy high-frequency data.
Journal of the American Statistical Association, 100(472):1394–1411, 2005.

[ZZCQ24] Chao Zhang, Yihuang Zhang, Mihai Cucuringu, and Zhongmin Qian.
Volatility forecasting with machine learning and intraday commonality.
Journal of Financial Econometrics, 22(2):492–530, 2024.

Thank You