*Journal of Financial Econometrics*, 2024, Vol. 22, No. 2, 492–530
https://doi.org/10.1093/jjfinec/nbad005
Advance Access Publication Date: 20 March 2023
Article

OXFORD

# Volatility Forecasting with Machine Learning and Intraday Commonality*

**Chao Zhang [1,2,3], Yihuang Zhang[2,4], Mihai Cucuringu[1,2,3,5], and Zhongmin Qian[2,4]**

[1]Department of Statistics, University of Oxford, Oxford, UK, [2]Mathematical Institute, University of Oxford, Oxford, UK, [3]Oxford-Man Institute of Quantitative Finance, University of Oxford, UK, [4]Oxford Suzhou Centre for Advanced Research, Suzhou, China, and [5]The Alan Turing Institute, London, UK

Address correspondence to Chao Zhang, Department of Statistics, University of Oxford, Oxford, UK, or e-mail: chao.zhang@stats.ox.ac.uk.

The first two authors contributed equally to this work.

Received February 7, 2022; revised February 15, 2023; editorial decision February 18, 2023; accepted February 24, 2023

## Abstract

We apply machine learning models to forecast intraday realized volatility (RV), by exploiting commonality in intraday volatility via pooling stock data together, and by incorporating a proxy for the market volatility. Neural networks dominate linear regressions and tree-based models in terms of performance, due to their ability to uncover and model complex latent interactions among variables. Our findings remain robust when we apply trained models to new stocks that have not been included in the training set, thus providing new empirical evidence for a universal volatility mechanism among stocks. Finally, we propose a new approach to forecasting 1-day-ahead RVs using past intraday RVs as predictors, and highlight interesting time-of-day effects that aid the forecasting mechanism. The results demonstrate that the proposed methodology yields superior out-of-sample forecasts over a strong set of traditional baselines that only rely on past daily RVs.

**Key words:** commonality, intraday volatility forecasting, neural networks, realized volatility

**JEL classification:** C45, C53, G17

*We would like to thank two anonymous referees, the associate editor and the editor, Dacheng Xiu, for their valuable comments. We are grateful to Rama Cont, Alvaro Cartea, Blanka Horvath, and participants at the 11th Bachelier World Congress 2022 and the 2022 Asian Finance Association Annual Conference for helpful comments. We also thank the Oxford Suzhou Centre for Advanced Research for providing the computational facility. The first author acknowledges the support from Clarendon Fund. The second author acknowledges the support from EPSRC Centre for Doctoral Training in Mathematics of Random Systems: Analysis, Modelling and Simulation (EP/S023925/1).

Forecasting and modeling stock return volatility has been of interest to both academics and practitioners. Recent advances in high-frequency trading (HFT) highlight the need for robust and accurate intraday volatility forecasts. For example, Deutsche Börse, one of the world's leading data and technology service providers, launched the "Intraday Volatility Forecast" project in 2015 to provide intraday volatility forecasts up to 30-min for DAX, EURO STOXX 50, and Euro-Bund.

Engle and Sokalska (2012) pointed out that intraday volatility forecasts are important for managing risk, pricing derivatives, and devising quantitative strategies, especially in HFT. Stroud and Johannes (2014) also demonstrated that intraday measures are useful for market makers, HFT, and option traders. Specifically, intraday volatility forecasts may support traders in assessing the likelihood of price changes and therefore better understanding the risk involved in certain automated trading strategies (see Bates 2019). Screen traders could leverage volatility forecasts to support live trading and enhance the pre-trade transaction cost analysis from the risk assessment of price slippage. Intraday volatility forecasts are also helpful for practitioners screening for high-volatility opportunities and trading corresponding option strategies (see Ni, Pan, and Poteshman 2008).

To the best of our knowledge, unlike daily volatility forecasting, intraday volatility has not yet received much attention in the research literature. It is pointed out by Andersen and Bollerslev (1997) that conventional parametric models, such as Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) and stochastic volatility (SV) models, may fail to reveal certain features of intraday returns. In Andersen et al. (2003) and Corsi (2009), high-frequency data are used to estimate daily realized volatility (RV) by summing squared intraday returns. Some related literature on this aspect will be reviewed briefly in the next section. These methods may reveal important information about characteristics of daily returns and volatilities, but do not easily lend themselves applicable to the task of forecasting intraday volatility.

In this article, we study several non-parametric machine learning (ML) models for forecasting *multi-asset intraday volatility* by leveraging high-frequency data from the U.S. equity market. We first propose a measure for evaluating the commonality in intraday volatility. The results demonstrate that, by taking advantage of commonality in intraday volatility, the forecasting performance of these ML models improves significantly. Neural networks (NNs) yield both statistically and economically significant improvements in out-of-sample performance over linear regressions and tree-based models, due to their ability to uncover the non-linearity and model complex latent interactions among variables. The improvements remain robust when we apply trained models to new stocks that have not been included in the training set, thus alleviating the overfitting concerns of NNs and providing new evidence toward a certain universality phenomenon in modeling volatility. In the end, our findings reveal that past intraday volatilities provide additional useful information for forecasting daily volatility, and reveal subtle time-of-day effects that aid the forecasting mechanism.

We augment our proposed methodology with a very thorough set of numerical experiments. The data covered in this work span the period from July 2011 to June 2021, and include the top 100 most liquid components of the Standard & Poor's (S&P) 500 index, and the 10-min, 30-min, 65-min, and daily (without overnight information) forecasting horizons are analyzed.

More specifically, a measure for evaluating the commonality in intraday volatility is proposed, which is the adjusted *R*-squared value from linear regressions of the RVs of a given stock against the market RVs. It is demonstrated that commonality over the daily horizon is turbulent over time, although the commonality in intraday RVs is strong and stable. The analysis of the high-frequency data from the real market reveals the following interesting phenomena. During a trading session, commonality achieves a peak near closing sessions, in contrast to the diurnal volatility pattern.

Second, in order to assess the benefits of incorporating commonality into models used to predict intraday volatility, multiple ML algorithms (including autoregressive integrated moving averages [ARIMA], heterogeneous autoregressive [HAR], ordinary least squares [OLS], least absolute shrinkage and selection operator [LASSO], XGBoost, multilayer perceptron [MLP], and long short-term memory [LSTM]) are implemented under three different schemes: (i) *Single*: training specific models for each asset; (ii) *Universal*: training one model with pooled data for all assets; (iii) *Augmented*: training one model using pooled data with an additional predictor which takes into account the impact of market RV. It is revealed that for most models, the incorporation of intraday commonality likely leads to better out-of-sample performance, based on the pooled data together with additional information of the market volatility.

The empirical results we present in the article demonstrate that NNs can be superior to other techniques. Empirical evidence is provided to demonstrate the capability of NNs for capturing complex interactions among predictors. Furthermore, to alleviate the concerns of over-fitting, a stringent out-of-sample test is conducted, where the trained models are evaluated on completely new stocks which have not been included in the training sample. Our results reveal that NNs can outperform other approaches. By comparing the result with the performance obtained by OLS models trained for each new stock, we show the validity of a universal volatility mechanism among stocks. Similar findings are reported in Sirignano and Cont (2019) concerning universal features of price formation in equity markets.

We conclude the article by proposing a new approach for predicting daily volatility, in which the past intraday volatilities rather than the past daily volatilities are used as predictors. This approach fully utilizes the available high-frequency data, and therefore contributes to the improvement over traditional methods of modeling daily volatilities. The results presented in this article demonstrate that ML models, where past intraday volatilities are used as predictors, tend to outperform the traditional models with past daily volatilities (e.g., HAR of Corsi [2009], SHAR of Patton and Sheppard [2015], and HARQ of Bollerslev, Patton, and Quaedvlieg [2016]). We believe that our approach brings a novel perspective on research that studies the effectiveness of past intraday volatilities in forecasting future daily volatility, providing new insights into the understanding of volatility dynamics.

The remainder of this article is structured as follows. We begin with Section 1 by reviewing some closely related literature, which however should not be considered as a comprehensive survey of the subject. In Section 2, the data and the definition of RV are described. In Section 3, we discuss the commonality in intraday volatility. Various ML models and three training schemes for predicting future intraday volatility are introduced in Section 4. Section 5 provides the forecasting results and discusses the empirical findings. In Section 6, a new approach to forecasting daily volatility using past intraday volatility as predictors is proposed. Finally, we summarize our study and discuss further avenues of investigation in Section 7.

## 1 Related Literature

Our study is built upon several research streams proposed by various authors over the recent years. The first stream is related to the research on the commonality in financial markets. Chordia, Roll, and Subrahmanyam (2000) have recognized the existence of commonality in liquidity and Karolyi, Lee, and Van Dijk (2012) have suggested that commonality in liquidity is related to market volatility, in particular, the presence of international investors and trading activity. Dang, Moshirian, and Zhang (2015) have made an observation that the news commonality is associated with stock return co-movement and liquidity commonality.

The co-movement in daily volatility is well known from the previous literature. Traditional GARCH and SV models (e.g., Andersen et al. 2006; Calvet, Fisher, and Thompson 2006) all make use of the volatility spillover effects. Herskovic et al. (2016) have provided empirical evidence of the co-movement in volatility across the equity market. Bollerslev et al. (2018) have observed strong similarities in daily RV and have utilized them to forecast the daily RV. Engle and Sokalska (2012) have emphasized that pooled data are useful for intraday volatility forecasting and Herskovic et al. (2020) have reported that volatilities co-move strongly over time. However, there is still a void of research related to commonality in intraday volatility and its implications for managing intraday risks, especially for forecasting purposes.

Second, there are numerous contributions in the existing literature on the topic of forecasting daily volatility. However, most methods proposed by various researchers for modeling and forecasting return volatility largely rely on the parametric GARCH or SV models, which provide forecasts of daily volatility from daily return. As pointed out by Andersen et al. (2003, 2006) and Engle and Patton (2007), these models employed to predict daily volatility cannot take advantage of high-frequency data, and suffer from the curse of high-dimensionality when dealing with multiple assets simultaneously. Due to the availability of high-frequency data, RV, computed from summing squared intraday returns, has gained popularity in recent years. Andersen et al. (2003) have proposed an AutoRegressive Fractionally Integrated Moving Average (ARFIMA) model for forecasting daily RVs, which outperforms conventional GARCH and related approaches. Corsi (2009) has put forward a parsimonious AR-type model, termed HAR, for predicting daily RVs using various RV components over different time horizons. Recently, Izzeldin et al. (2019) have made a comparison investigation for the forecasting performance of ARFIMA and HAR, and have concluded that their performance is essentially indistinguishable. See Section 6 for more models to predict daily volatility.

Nonetheless, little attention has been paid to the role of forecasting intraday volatility. Taylor and Xu (1997) have proposed an hourly volatility model based on an ARCH specification and Engle and Sokalska (2012) have constructed a GARCH model for intraday financial returns, by specifying the variance as a product of daily, diurnal, and stochastic intraday components. These models, such as traditional GARCH and SV, are potentially restrictive due to their parametric nature, and are not able to effectively take into account the non-linear and highly complex relationships among different financial variables.

Third, ML models have demonstrated great potential in finance, such as their applications in asset pricing. The high-dimensional nature of ML methods allows for better approximations to unknown and potentially complex data-generating processes, in contrast

with traditional economic models. Gu, Kelly, and Xiu (2020) have pointed out the superior performance of ML models for empirical asset pricing. Recently, Xiong, Nichols, and Shen (2015) have applied LSTMs to forecast S&P 500 volatility, with Google domestic trends as predictors, and Bucci (2020) has demonstrated that recurrent NNs (RNNs) are able to outperform all the traditional econometric methods in forecasting monthly volatility of the S&P index. Rahimikia and Poon (2020) have compared ML models with HAR models for forecasting daily RV by using variables extracted from limit order books and news. Li and Tang (2020) have proposed a simple average ensemble model combining multiple ML algorithms for forecasting daily (and monthly) RV and Christensen, Siggaard, and Veliyev (2021) have examined the performance of ML models in forecasting 1-day-ahead RV with firm-specific characteristics and macroeconomic indicators.

## 2 Data and RV

### 2.1. Data

We use the Nasdaq ITCH data from LOBSTER[1] to compute intraday returns via mid-prices. We select the top 100 components of S&P 500 index, for the period between July 1, 2011 and June 30, 2021. After filtering out the stocks for which the dataset does not span the entire sample period, we are left with 93 stocks. Table 1 presents the number of stocks in each sector, according to the Global Industry Classification Standard (GICS) sector division.[2]

### 2.2. Realized Volatility

In a general form, $P_{i,t}$ denotes the price process of a financial asset $i$ and it follows:

$$\mathrm{d}\log P_{i,t} = \mu_i \mathrm{d}t + \sigma_{i,t}\mathrm{d}W_t, \tag{1}$$

where $\mu_i$ is the drift, $\sigma_{i,t}$ is the instantaneous volatility, and $W_t$ is the standard Brownian motion. The theoretical integrated variance (IV) of stock $i$ during $(t-h, t]$ is estimated as

$$\mathrm{IV}_{i,t}(h) = \int_{t-h}^{t} \sigma_{i,s}^2 \mathrm{d}s, \tag{2}$$

where $h$ is the look-back horizon, such as 10 min, 30 min, 1 day, etc.

In this article, we consider the minutely logarithmic mid-price return for asset $i$ during $(t-1, t]$ as

$$r_{i,t} := \log\left(\frac{P_{i,t}}{P_{i,t-1}}\right). \tag{3}$$

Here, $P_{i,t}$ is the mid-price at time $t$, that is, $P_{i,t} = \frac{P^b_{i,t}+P^s_{i,t}}{2}$ and $P^b_{i,t}$ (respectively, $P^a_{i,t}$) represents the best bid (respectively, ask) price.

Andersen et al. (2001) and Barndorff-Nielsen and Shephard (2002) showed that the sum of squared intraday returns is a consistent estimator of the unobservable IV. Because of the availability of high-frequency intraday data, we choose to compute RV as a proxy for the unobserved IV (see Andersen et al. 2001; Bollen and Inder 2002; Hansen and Lunde 2006).

1 https://lobsterdata.com/ (accessed on February 28, 2022).

2 The GICS is an industry taxonomy developed in 1999 by MSCI and S&P.

**Table 1.** Components in each sector

| Sector | Number | Tickers |
|---|---|---|
| Information Technology | 20 | AAPL ACN ADBE ADP AVGO CRM CSCO FIS FISV IBM INTC INTU MA MSFT MU NVDA ORCL QCOM TXN V |
| Health Care | 19 | ABT AMGN BDX BMY BSX CI CVS DHR GILD ISRG JNJ LLY MDT MRK PFE SYK TMO UNH VRTX |
| Financials | 15 | AXP BAC BLK BRK.B C CB CME GS JPM MMC MS PNC SCHW USB WFC |
| Industrials | 9 | BA CAT CSX GE HON LMT MMM UNP UPS |
| Consumer Discretionary | 8 | AMZN HD LOW MCD NKE SBUX TGT TJX |
| Consumer Staples | 8 | CL COST KO MO PEP PG PM WMT |
| Communication Services | 6 | CMCSA DIS GOOG NFLX T VZ |
| Others | 8 | AMT CCI COP CVX D DUK SO XOM |

To reduce the impact of extreme values, we consider the logarithm, in line with Andersen et al. (2003), Bucci (2020) and Herskovic et al. (2016). Specifically, during a period $(t-h,t]$, the RV is defined as follows[3]:

$$\mathrm{RV}_{i,t}^{(h)} := \log\left[\sum_{s=t-h+1}^{t} r_{i,s}^2\right]. \tag{4}$$

As pointed out by Pascalau and Poirier (2021), there are no conclusive methods to incorporate the overnight session's information content into the daily volatility. In line with Engle and Sokalska (2012), overnight information is excluded from our empirical analysis of daily volatility. For simplicity, we refer to this daily scenario (excluding the overnight) as the "1-day" scenario, throughout the rest of this article.

## 2.3. Summary Statistics

To mitigate the effect of possibly spurious data errors, for each stock, we set the values of return/volatility below the 0.5 percentile equal to the respective 0.5 percentile, and the values above the 99.5 percentile is set equal to the 99.5 percentile, a process commonly referred to as *winsorization*. Figure 1 illustrates the pairwise Pearson and Spearman correlations of returns and realized volatilities. This figure depicts the empirical distribution of pairwise correlation coefficients over the entire sample period. We generally observe higher correlations in RV than the counterparts in return. Figure 1 also reveals that, on average, as the horizon gets longer, RV's correlations increase from 0.598 (10-min) to 0.731 (30-min) to 0.766 (65-min). However, when turning to daily RV, correlations in RVs become weaker, with an average of 0.514. This indicates that the connections between stocks in terms of intraday volatility may be more stable and tight than the ones in daily volatility.

Figure 2 plots the daily RV over time. Stocks demonstrate similar time-series patterns, consistent with Herskovic et al. (2016) and Bollerslev et al. (2018). Additionally, the width shrinks

3 Liu, Patton, and Sheppard (2015) demonstrate that no sub-sampling frequency significantly outperforms a 5-min interval in terms of forecasting daily RVs, making it a widely accepted time interval in the literature. In this article, we use 1-min returns since our main focus is intraday RVs, such as 10-min RVs.

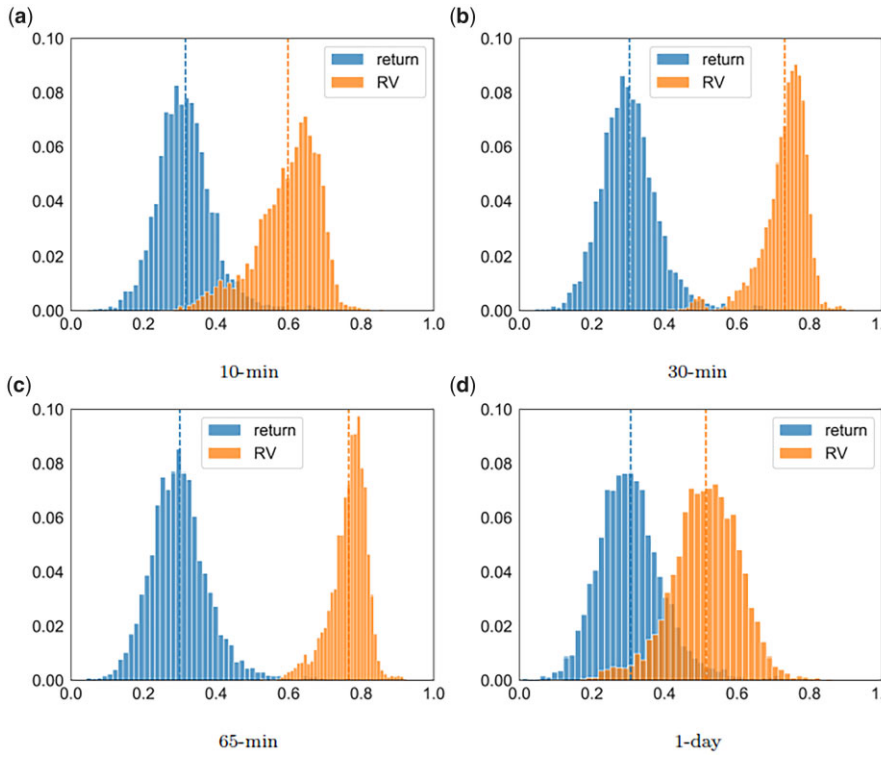


**Figure 1.** Histograms of pairwise correlations of realized volatilities and returns.

*Notes:* The panels (**a**)–(**d**) are based on observations in the frequency of 10-min, 30-min, 65-min, and 1-day, respectively. The dashed vertical lines represent the average correlation values of RVs and returns.

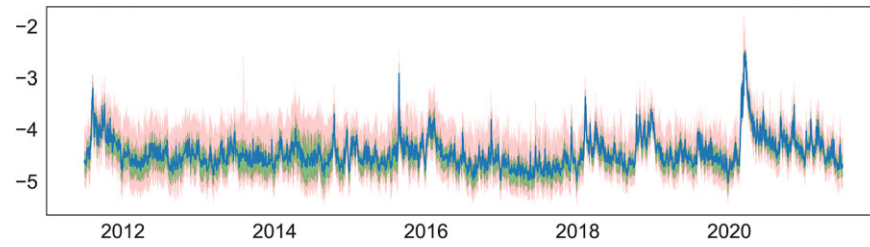


**Figure 2.** Daily RV (in logs).

*Notes:* The blue curve represents cross-sectional average of daily RV across stocks, with the inner area covering the 25th percentile to the 75th percentile, and the outer area covering the 5th percentile to the 95th percentile.

during the periods of higher volatility, such as, stock market crashes in August 2011 (European sovereign debt crisis), between June 2015 and June 2016 (Chinese stock market turbulence and Brexit), in March 2018 (China–U.S. trade war), in March 2020 (COVID-19). Figure 3 shows

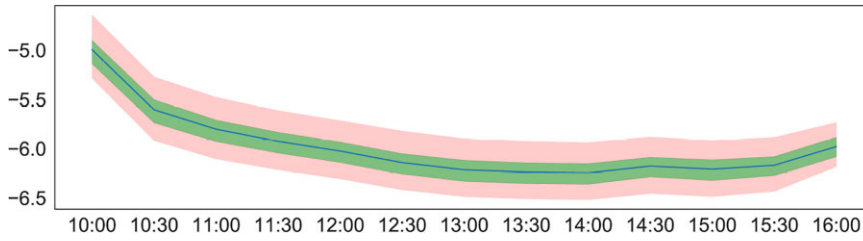


**Figure 3.** Diurnal RV (in logs).

*Notes:* The blue curve represents cross-sectional average of 30-min RV across stocks and days, with the inner area covering the 25th percentile to the 75th percentile and the outer area covering the 5th percentile to the 95th percentile.

that the diurnal volatility forms a so-called reverse-J-shape, namely larger fluctuations near the open and close (see Harris 1986; Engle and Sokalska 2012).

## 3 Commonality Estimation

Inspired by prior studies (e.g., Chordia, Roll, and Subrahmanyam 2000; Morck, Yeung, and Yu 2000; Karolyi, Lee, and Van Dijk 2012; Dang, Moshirian, and Zhang 2015), we follow an analogous procedure to estimate the commonality in volatility. Specifically, we use the average adjusted $R^2$ value from the following regressions across stocks, as a measure of commonality in volatility (denoted as $R^2_{(h)}$)[4]

$$\mathrm{RV}_{i,t}^{(h)} = \alpha_i + \beta_i \mathrm{RV}_{M,t}^{(h)} + \epsilon_{i,t}, \tag{5}$$

where $\mathrm{RV}_{M,t}^{(h)}$ (see Bollerslev et al. 2018) is the contemporaneous market volatility during $(t-h,t]$ for stock $i$, which is calculated as the equally weighted average[5] of all individual stock volatilities during $(t-h,t]$, that is,

$$\mathrm{RV}_{M,t}^{(h)} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{RV}_{i,t}^{(h)}. \tag{6}$$

Figure 4 presents the commonality in RV, averaged across stocks for each month. To create this figure, we use the observations in each month to obtain the $R^2$ value from Equation (5). We notice that commonality effects in intraday scenarios (especially 30-min and 65-min) are substantially larger than the daily ones. For example, as reported in Table 2, the average commonality in 65-min data is around 74.3%, while only 35.5% in daily data. Moreover, $R^2_{(h)}$ is much more turbulent at the daily frequency. The last column

4 We also perform another regression, where except for contemporaneous market volatility, the lag one (thus $t-1$ in Equation (5)) and lead one (thus $t+1$ in Equation (5), hence not computable in real time due to the forward looking bias) in market volatility are also included, in order to explain non-contemporaneous trading, in line with Chordia, Roll, and Subrahmanyam et al. (2000); Karolyi, Lee, and Van Dijk (2012); and Dang, Moshirian, and Zhang (2015). The $R^2$ values are similar to the ones of Equation (5).

5 We also implemented the value-weighted market volatility and the results are similar to the equally weighted market volatility.

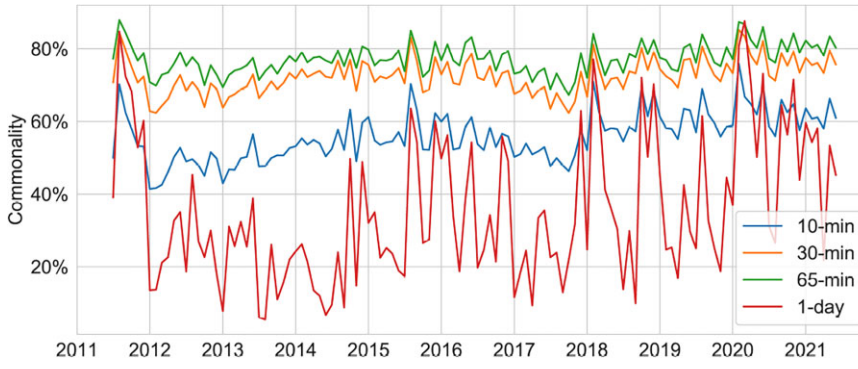


**Figure 4.** Commonality in RV.

*Notes:* The commonality is averaged across stocks for each month during the sample period of July 2011–June 2021.

**Table 2.** Statistics of the monthly average commonality in volatility

| | Mean | Std | Corr. with VIX |
|---|---|---|---|
| 10-min | 0.560 | 0.068 | 0.536 |
| 30-min | 0.725 | 0.048 | 0.574 |
| 65-min | 0.743 | 0.041 | 0.609 |
| 1-day | 0.355 | 0.198 | 0.690 |

*Note:* VIX represents the market volatility from the Chicago Board Options Exchange.

in Table 2 also reports the results of the relation between the average commonality and the market volatility. As the horizon extends, the average commonality has a higher correlation with the market volatility.[6]

Figure 5 reports the averaged values and standard deviations (black vertical lines) of commonality for each half-hour in the trading session. To create this figure, we use the observations in a given interval, such as [09:30, 10:00], to fit Equation (5). We observe a gradual increase in commonality throughout the trading session as we get closer to market close, in sharp contrast to the diurnal volatility pattern in Figure 3.

## 4 Methodology

In this section, we leverage the commonality for the task of predicting cross-asset volatility. We construct the prediction model as follows:

$$\begin{aligned} \mathrm{RV}_{i,t+h}^{(h)} &= F_i(\mathbf{u};\theta) + \epsilon_{i,t+h} \\ &= F_i\left(\mathrm{RV}_{i,t}^{(h)},\ldots,\mathrm{RV}_{i,t-(p-1)h}^{(h)},\mathrm{RV}_{M,t}^{(h)},\ldots,\mathrm{RV}_{M,t-(p-1)h}^{(h)};\theta\right) + \epsilon_{i,t+h}, \end{aligned} \tag{7}$$

where $\mathrm{RV}_{i,t+h}^{(h)}$ is the volatility of asset $i$ during $(t, t+h]$. $\mathbf{u}$ represents the input features, which can be further separated into two categories: (i) a multi-dimensional vector of

6 We refer the reader to additional analysis on commonality in Appendix A.

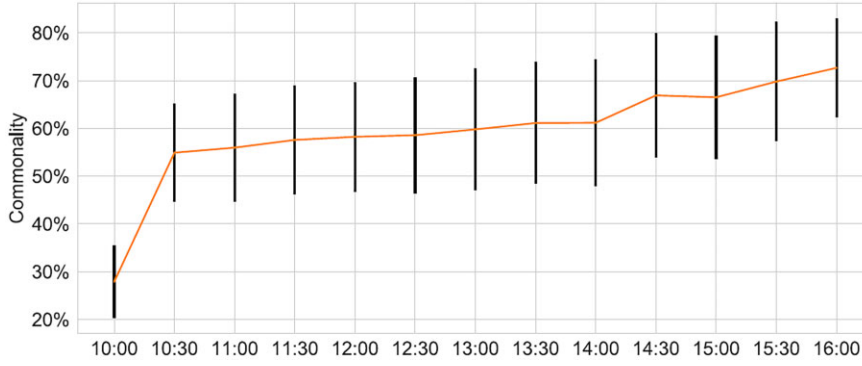


**Figure 5.** Commonality in RV.

*Notes:* The commonality is averaged across stocks for each half-hour during the sample period of July 2011–June 2021.

predictor variables for a specific stock $i$ available up to time $t$, denoted as *individual features*, such as $\left(\mathrm{RV}_{i,t}^{(h)}, \ldots, \mathrm{RV}_{i,t-(p-1)h}^{(h)}\right)'$ and (ii) a vector of features for all stocks in the studied universe up to $t$, denoted as *market features*, such as $\left(\mathrm{RV}_{M,t}^{(h)}, \ldots, \mathrm{RV}_{M,t-(p-1)h}^{(h)}\right)'$. $\theta$ refers to the parameters that need to be estimated. Whenever is clear from the context and no ambiguity arises, we use also use $\theta$ to denote the forecasting model. We are aiming to find a function of variables that minimizes the out-of-sample errors for future RV.

## 4.1. Models

This section summarizes the collection of ML models employed in our numerical experiments.

### 4.1.1 Seasonal ARIMA

The ARIMA model is a popular forecasting method for univariate time-series data, where an initial differencing step can be applied one or more times to eliminate the non-stationarity of the trend. An ARIMA$(p, d, q)$ is given by

$$\varphi(L)(1-L)^d \mathrm{RV}_{i,t}^{(h)} = \rho(L)\varepsilon_{i,t}, \tag{8}$$

where $\varphi(L) = 1 - \sum_{k=1}^{p} \varphi_i L^i$ and $\rho(L) = 1 - \sum_{j=1}^{q} \rho_j L^j$ are the AR and MA lag polynomials, and $\epsilon_{i,t}$ is the error which is distributed as $\mathcal{N}(0, \sigma_i^2)$. Following Christensen and Prabhala (1998) and Ribeiro et al. (2021), we adopt ARIMA$(1, 1, 1)$ to model the daily RV.

When the time series exhibits seasonality, the seasonal differencing could be applied to eliminate the seasonal component, which is denoted as seasonal ARIMA (SARIMA). As revealed in Figure 3, intraday volatility time series possesses a seasonal component. For modeling intraday RV, we choose the SARIMA, where the seasonal period is the corresponding number of intraday time buckets in a day and other parameters related to the seasonal pattern are set as zero (for more details about SARIMA, see Sheppard 2010).

### 4.1.2 HAR with diurnal effects

Corsi (2009) proposed a volatility model, named as HAR, which considers realized volatilities over different interval sizes. HAR has shown remarkably good forecasting

performance on daily data (Patton and Sheppard 2015; Izzeldin et al. 2019). For day $t$, the forecast of HAR is based on

$$\mathrm{RV}_{i,t+1}^{(d)} = \alpha_i + \beta_i^{(d)}\mathrm{RV}_{i,t}^{(d)} + \beta_i^{(w)}\mathrm{RV}_{i,t}^{(w)} + \beta_i^{(m)}\mathrm{RV}_{i,t}^{(m)} + \epsilon_{i,t+1}, \tag{9}$$

where $\mathrm{RV}_{i,t}^{(d)}$ denotes the daily RV in the past day, and $\mathrm{RV}_{i,t}^{(w)} = \frac{1}{5}\sum_{l=1}^{5}\mathrm{RV}_{i,t-l}^{(d)}$ and $\mathrm{RV}_{i,t}^{(m)} = \frac{1}{21}\sum_{l=1}^{21}\mathrm{RV}_{i,t-l}^{(d)}$ denote the weekly and monthly lagged RV, respectively. The choice of a daily, weekly, and monthly lag is aiming to capture the long-memory dynamic dependencies observed in most RV series.

However, very little attention has been paid to forecasting intraday volatility with HAR. One closely connected model is that of Engle and Sokalska (2012), who proposed an intraday volatility forecasting model, where they interpret that conditional volatility of high-frequency returns is a product of daily, diurnal, and stochastic intraday components. After the decomposition of raw returns, the authors apply a GARCH model (Engle 1982) to learn the stochastic intraday volatility components.

Following the spirit of Engle and Sokalska (2012), we extend the daily HAR model to intraday scenarios by adding diurnal effect and previous intraday component, as follows[7]:

$$\mathrm{RV}_{i,t+h}^{(h)} = \alpha_i + \beta_i^{(\tau)}D_{i,\tau_{t+h}} + \beta_i^{(s)}\mathrm{RV}_{i,t}^{(h)} + \beta_i^{(d)}\mathrm{RV}_{i,t}^{(d)} + \beta_i^{(w)}\mathrm{RV}_{i,t}^{(w)} + \beta_i^{(m)}\mathrm{RV}_{i,t}^{(m)} + \epsilon_{i,t+h}, \tag{10}$$

where $D_{i,\tau_{t+h}}$ denotes the average diurnal RV in the bucket-of-the-day $\tau_{t+h}$ computed from the last 21 days. For example, when $t = 10{:}30$ and $h = 30$ min, then $\tau_{t+h}$ corresponds to the bucket 10:30–11:00. $\mathrm{RV}_{i,t}^{(h)}$ represents the lag = 1 intraday RV. $\mathrm{RV}_{i,t}^{(d)}$ ($\mathrm{RV}_{i,t}^{(w)}$, $\mathrm{RV}_{i,t}^{(m)}$) denotes the aggregated daily (weekly, monthly) RV. When we consider the daily scenarios, Equation (10) becomes the standard HAR model (Equation 9), by removing the diurnal term and the intraday component.

#### 4.1.3 Ordinary least squares

Instead of using aggregated RV, we apply OLS to original features, as follows, with its loss function being the sum of squared errors. Recall $\mathbf{u} = (u_1, \ldots, u_p)'$ represent the vector of input features, such as past intraday RVs, and (perhaps) market RVs. Notice that the model only incorporating the past intraday RVs as features is actually an autoregressive (AR) model:

$$\mathrm{RV}_{i,t+h}^{(h)} = \alpha_i + \sum_{l=1}^{p}\beta_l u_l + \epsilon_{i,t+h}. \tag{11}$$

#### 4.1.4 Least absolute shrinkage and selection operator

When the number of predictors approaches the number of observations, or there are high correlations among predictor variables, the OLS model tends to overfit noise rather than signals. This is particularly burdensome for the volatility forecasting problem, where the features could be highly correlated.

LASSO is a linear regression method that can avoid overfitting via adding a penalty of parameters to the objective function. As pointed out by Hastie, Tibshirani, and Friedman

7 Since we use the log-version realized volatility, the multiplication of daily, diurnal, and stochastic intraday components in Engle and Sokalska (2012) translates to the addition in our model (10).

(2009), LASSO performs both variable selection and regularization, therefore enhances the prediction accuracy and interpretability of regression models. The objective function of LASSO is the sum of squared residuals and an additional $l_1$ constraint on the regression coefficients, as shown in Equation (12). Here, the hyperparameter $\lambda$ controls the penalty weight. In our experiments, we provide a set of hyperparameter values and then choose the one with the best performance on the validation data, as our forecasting model:

$$L_i = \sum_t \left[ \mathrm{RV}_{i,t+h}^{(h)} - \alpha_i - \sum_{l=1}^{p} \beta_l u_l \right]^2 + \lambda \sum_{l=1}^{p} \left\| \beta_l \right\|_1 . \tag{12}$$

#### 4.1.5 XGBoost

Linear models are unable to capture the possible non-linear relations between the dependent variable and the predictors, and the interactions among predictors. As pointed by Bucci (2020), RVs are subject to structural breaks and regime-switching, hence the need to consider non-linear models. One way to add non-linearity and interactions is the decision tree, see more in Hastie, Tibshirani, and Friedman (2009).

XGBoost is a decision-tree-based ensemble algorithm, implemented under a distributed gradient boosting framework by Chen and Guestrin (2016). There is abundant empirical evidence showing the success of XGBoost, such as in a large number of Kaggle competitions. In this work, we only review the essential idea behind XGBoost—tree boosting model. For more details about other important features of XGBoost, such as the scalability in various scenarios, parallelization, distributed computing, feature importance to enhance interpretability, etc., the reader may refer to Chen and Guestrin (2016). Let $\mathbf{u}$ represent the vector of input features,

$$F_i(\mathbf{u}) = \sum_{l=1}^{B} f_l(\mathbf{u}), \quad f_l \in \mathcal{F}, \tag{13}$$

where $\mathcal{F}$ is the space of regression trees. An example of the tree ensemble model is depicted in Figure 6. The tree ensemble model in Equation (13) is trained sequentially. Boosting (see Friedman 2001) means that new models are added to minimize the errors made by existing models, until no further improvements are achieved.

#### 4.1.6 Multilayer perceptron

Another non-linear method is the NN, which has become increasingly popular for ML problems, for example, in computer vision and natural language processing, due to its flexibility to learn complex interactions. Hill, O'Connor, and Remus (1996) and Zhang (2003) suggest that NN is a promising non-linear alternative to the traditional linear methods in time-series forecasting. We introduce two commonly implemented NNs in the following sections.

MLP is a class of feedforward NNs and a "universal approximator" that can learn any smooth functions (see Hornik, Stinchcombe, and White 1989). MLP has been applied to many fields, for example, computer vision and natural language processing. MLPs are composed of an input layer to receive the raw features, an output layer that makes forecasts about the input, and in-between those two, an arbitrary number of hidden layers that are non-linear transformations. MLPs perform a static mapping between an input space and an

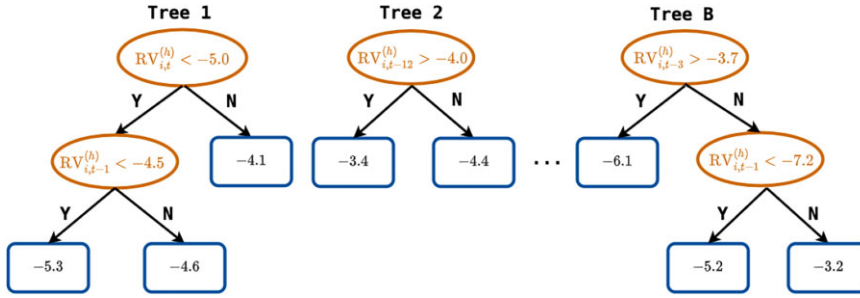


**Figure 6.** Illustration of a tree ensemble model.

*Notes:* B represents the number of trees. The final prediction of a tree ensemble model is the sum of predictions from each tree, as shown in Equation (13).

output space (Bucci 2020). The parameters in MLPs can be updated via stochastic gradient descent. In this work, we use Adam (see Kingma and Ba 2014), which is based on adaptive estimates of lower-order moments. Let $\mathbf{u} \in \mathbb{R}^p$ represent the input variables

$$F_i(\mathbf{u};\theta) = \mathbf{W}_L \cdot \sigma\Big(\mathbf{W}_{L-1} \ldots \sigma(\mathbf{W}_1\mathbf{u} + \mathbf{b}_1) \ldots + \mathbf{b}_{L-1}\Big) + \mathbf{b}_L, \tag{14}$$

where $\theta := (\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_L, \mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_L)$ represents the parameters in the NN. $\mathbf{W}_l \in \mathbb{R}^{n_l \times n_{l-1}}$, $\mathbf{b}_l \in \mathbb{R}^{n_l \times 1}$ for $l = 1, 2, \ldots, L$, and $n_0 = p$. For the activation function $\sigma(\cdot)$, we choose the rectified linear unit, that is, $\sigma(x) = \max(x, 0)$.

### 4.1.7 Long short-term memory

A RNN requires that historic information is retained to forecast future values. Therefore, RNN is well-suited for processing, classifying, and making predictions based on time-series data (Bucci 2020). LSTM, proposed by Hochreiter and Schmidhuber (1997), is an extension of the RNN architecture by replacing each hidden unit in RNNs with a memory block to capture the long-term effect. LSTMs have received considerable success in natural language processing, time series, generative models, etc.

For simplicity, we consider the time series for a given stock and remove the subscript for stock identity. The standard transformation in each unit of LSTM is defined as follows. For a more detailed discussion, we refer the reader to Hochreiter and Schmidhuber (1997):

$$\begin{aligned}
\mathbf{f}_t &= \sigma_g(\mathbf{W}_f\mathbf{u}_t + \mathbf{U}_f\mathbf{h}_{t-1} + \mathbf{b}_f) \\
\mathbf{i}_t &= \sigma_g(\mathbf{W}_i\mathbf{u}_t + \mathbf{U}_i\mathbf{h}_{t-1} + \mathbf{b}_i) \\
\mathbf{o}_t &= \sigma_g(\mathbf{W}_o\mathbf{u}_t + \mathbf{U}_o\mathbf{h}_{t-1} + \mathbf{b}_o) \\
\tilde{\mathbf{c}}_t &= \sigma_c(\mathbf{W}_c\mathbf{u}_t + U_c\mathbf{h}_{t-1} + \mathbf{b}_c) \\
\mathbf{c}_t &= \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \tilde{\mathbf{c}}_t \\
\mathbf{h}_t &= \mathbf{o}_t \circ \sigma_h(\mathbf{c}_t),
\end{aligned} \tag{15}$$

where $\mathbf{u}_t$ is the input vector, $\mathbf{f}_t$ is the forget gate's activation vector, $\mathbf{i}_t$ is the update gate's activation vector, $\mathbf{o}_t$ is the output gate's activation vector, $\tilde{\mathbf{c}}_t$ is the cell input activation vector, $\mathbf{c}_t$ is the cell state vector, and $\mathbf{h}_t$ is the hidden state vector, that is, output vector of the LSTM unit. $\circ$ is the Hadamard product function. $\sigma_g$ is the sigmoid function, and $\sigma_c, \sigma_h$ are

hyperbolic tangent function. $\mathbf{W}_{f(i,o,c)}, \mathbf{b}_{f(i,o,c)}$ refer to weight matrices and bias vectors that need to be estimated.

To summarize, we first consider a traditional time-series model ARIMA, then include three linear regression models, that is, HAR(-D), OLS, and LASSO. To account for the nonlinear impact of individual predictors on future volatilities and the interactions among predictors, we choose an ensemble tree model XGBoost and two NNs, that is, MLP and LSTM. The primary difference between MLP and LSTM is that LSTM has feedback connections, which allow learning the dependencies in the input sequences.

### 4.2. Training Scheme

Motivated by the strong commonality in volatility across stocks, we consider the following three different schemes for model training.

- *Single* denotes that we train customized models $F_i$ for each stock $i$, as in Bucci (2020) and Hansen and Lunde (2005). We use a stock's own past RVs only as predictor features, namely

$$\mathbf{u} = \left(\mathrm{RV}_{i,t}^{(h)}, \ldots, \mathrm{RV}_{i,t-(p-1)h}^{(h)}\right)'$$

  and no market features, where $p$ represents the number of lags.
- *Universal* denotes that we train models with the pooled data of all stocks in our universe. That is, $F_i$ is same for all stocks in Equation (7). As in the *Single* scheme, we use a stock's own past RVs only as predictor features and no market features. Sirignano and Cont (2019) showed that the model trained on the pooled data outperforms asset-specific models trained on time series of any given stock, in the sense of forecasting the direction of price moves. Bollerslev et al. (2018) and Engle and Sokalska (2012) suggested that models estimated under the *Universal* setting yield superior out-of-sample risk forecasts, compared with models under the Single setting, when forecasting daily RV.
- *Augmented* denotes that we train models with the pooled data of all stocks in our universe, but in addition, we also incorporate a predictor that takes into account the impact of the market RV (e.g., Bollerslev et al. 2018) in order to leverage the commonality in volatility shown in Section 3. Namely, $F_i$ is same for all stocks in Equation (7). We use both individual features and market features as predictors, that is, $\mathbf{u} = \left(\mathrm{RV}_{i,t}^{(h)}, \ldots, \mathrm{RV}_{i,t-(p-1)h}^{(h)}, \mathrm{RV}_{M,t}^{(h)}, \ldots, \mathrm{RV}_{M,t-(p-1)h}^{(h)}\right)'$. Note that for HAR with diurnal (HAR-D) models under the *Augmented* setting, we include aggregated market features as additional features and use the OLS to estimate the parameters.

In summary, compared with the benchmark Single setting, we gradually incorporate cross-asset and market information into the training of models. The hyperparameters for each model are summarized in Appendix B.

### 4.3. Performance Evaluation

To assess the predictive performance for RV forecasts, we compute the following metrics on the rolling out-of-sample tests (see Patton and Sheppard 2009; Patton 2011; Engle and

Sokalska 2012; Bollerslev et al. 2018; Bucci 2020; Rahimikia and Poon 2020; Pascalau and Poirier 2021). Both functions measure losses, so lower values are preferred. Patton and Sheppard (2009) demonstrate that QLIKE has the highest power in the Diebold–Mariano (DM) test. Consequently, we focus more on the QLIKE rather than the MSE:

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

where $\widehat{\text{RV}}_{i,t}^{(h)}$ represents the predicted value of $\text{RV}_{i,t}^{(h)}$, the RV for stock $i$ during $(t-h,t]$. $N$ is the number of stocks in our universe, $\mathcal{T}_{\text{test}}$ is the testing period, and $\#\mathcal{T}_{\text{test}}$ is the length of the testing period.

#### 4.3.1 Model confidence set

MCS was proposed by Hansen, Lunde, and Nason (2011) to identify a subset of models $\mathcal{M}^*$ with significantly superior performance from model candidates $\mathcal{M}_0$, at a given level of confidence. The iterative elimination is based on sequentially testing the following hypothesis:

$$H_0 : \mathbb{E}(\Delta L_{ij,t}) = 0, \text{ forall } i,j \in \mathcal{M}^*, \tag{16}$$

where $L_{ij,t}$ is the loss difference between models $i$ and $j$ at day $t$ in terms of a specific loss function $L$, such as MSE and QLIKE. The model confidence set (MCS) procedure renders it possible to make statements about the statistical significance from multiple pairwise comparisons. For additional details, we refer to the studies of Hansen, Lunde, and Nason (2011).

### 4.4. Utility Benefits

We have demonstrated that one can assess the out-of-sample statistical performance for each model via the above metrics and tests. However, in such an approach, the economic magnitude of the gain from complex risk models is ignored. Bollerslev et al. (2018) have proposed a utility-based framework, which gauges the utility benefits of an investor with mean–variance preferences investing in an asset with time-varying volatility and a constant Sharpe ratio. We implement this framework to measure the volatility forecasts. For a more detailed description, we refer the reader to Bollerslev et al. (2018).

The expected utility of a mean–variance investor at $t$ can be approximated as

$$\mathbb{E}_t\Big(u(W_{t+1})\Big) = \mathbb{E}_t(W_{t+1}) - \frac{1}{2}\gamma^A \text{Var}_t(W_{t+1}), \tag{17}$$

where $W_t$ denotes the wealth and $\gamma^A$ denotes the absolute risk aversion of the investor.

Assume that the investor allocates a fraction $x_t$ of the current wealth to a risky asset with return $r_{t+1}$ and the rest to a risk-free money market account earning $r_t^f$. Then, the wealth at $t+1$ becomes $W_{t+1} = W_t(1 + r_t^f + x_t r_{t+1}^e)$, where $r_{t+1}^e \equiv r_{t+1} - r_t^f$. After dropping constant terms, the expected utility in Equation (17) amounts to

$$\mathbb{E}_t\Big(u(W_{t+1})\Big) := U(x_t) = W_t\left(x_t\mathbb{E}_t(r^e_{t+1}) - \frac{\gamma}{2}x_t^2\mathbb{E}_t\Big(\exp(\mathrm{RV}_{t+1})\Big)\right), \tag{18}$$

where $\gamma \equiv \gamma^A W_t$ denotes the investor's relative risk aversion.

Suppose the conditional Sharpe ratio $\mathrm{SR} \equiv \mathbb{E}_t(r^e_{t+1})/\sqrt{\mathbb{E}_t\Big(\exp(\mathrm{RV}_{t+1})\Big)}$ is constant, so that the expected utility is

$$U(x_t) = W_t\left(x_t\mathrm{SR}\sqrt{\mathbb{E}_t\Big(\exp(\mathrm{RV}_{t+1})\Big)} - \frac{\gamma}{2}x_t^2\mathbb{E}_t\Big(\exp(\mathrm{RV}_{t+1})\Big)\right). \tag{19}$$

The optimal portfolio that maximizes this utility is obtained by investing the following fraction of wealth to the risky asset:

$$x_t^* = \frac{\mathrm{SR}/\gamma}{\sqrt{\mathbb{E}_t\Big(\exp(\mathrm{RV}_{t+1})\Big)}}. \tag{20}$$

To determine the utility gains based on different risk models, the expectation based on model $\theta$ is denoted by $\mathbb{E}_t^\theta(\cdot)$. Assuming that the investor uses model $\theta$, then the position $x_t^\theta = \frac{\mathrm{SR}/\gamma}{\sqrt{\mathbb{E}_t^\theta\Big(\exp(\mathrm{RV}_{t+1})\Big)}}$ is chosen. By plugging $x_t^\theta$ into Equation (19) and replacing $\mathbb{E}_t(\exp(\mathrm{RV}_{t+1}))$ with the RV $\exp(\mathrm{RV}_{t+1})$, the expected utility per unit of the wealth (called realized utility, or in short RU) is given by

$$\mathrm{RU}_t = \frac{\mathrm{SR}^2}{\gamma} \times \frac{\sqrt{\exp(\mathrm{RV}_{t+1})}}{\sqrt{\mathbb{E}_t^\theta(\exp(\mathrm{RV}_{t+1}))}} - \frac{\mathrm{SR}^2}{2\gamma} \times \frac{\exp(\mathrm{RV}_{t+1})}{\mathbb{E}_t^\theta(\exp(\mathrm{RV}_{t+1}))}. \tag{21}$$

If a risk model is ideal, that is, it predicts perfectly the realized volatilities $\mathbb{E}_t^\theta(\exp(\mathrm{RV}_{t+1})) = \exp(\mathrm{RV}_{t+1})$, then its realized utility is $\frac{\mathrm{SR}^2}{2\gamma}$. Alternatively, the investor is willing to give up $\frac{\mathrm{SR}^2}{2\gamma}$ of the wealth in order to utilize the perfect risk model instead of investing only in the risk-free asset. In this article, the same Sharpe ratio ($\mathrm{SR} = 0.4$) and the same coefficient of risk aversion ($\gamma = 2$) are applied as in Bollerslev et al. (2018) and Li and Tang (2020).

The previous comparisons are based on a frictionless setting, ignoring the trading cost. The case of incorporating the effect of transaction costs is also considered. Following Bollerslev et al. (2018) and Li and Tang (2020), we assume that transaction costs are linear in the absolute magnitude of the change in the positions, and use the full median bid–ask spread for each of the assets over the last 90 trading days. The realized utility with trading costs deducted, denoted as RU-TC, is simply the realized utility after subtracting the simulated costs. We evaluate this realized utility (with and without trading cost) empirically by averaging the corresponding realized expressions over stocks and the same rolling out-of-sample forecasts.

## 5 Experiments

### 5.1. Implementation

For each data set, we divide the observations into three non-overlapping periods and maintain their chronological order: training, validation, and testing. For a given trading day $t$,

the training data, including the samples in the first period [July 1, 2011, $t-251$], are used to estimate models subject to a given architecture. Validation data, including the recent 1-year samples $[t-250, t]$, are deployed to tune the hyperparameters of the models. Finally, testing data are samples in the next year $[t+1, t+251]$; they are out-of-sample in order to provide objective assessments of the models' performance. Due to limited computational resources, models are updated annually. In other words, when we retrain the models in the next calendar year, the training data expand by one year, whereas the validation samples are rolled forward to include the samples in the most recent 1-year period, following Gu, Kelly, and Xiu (2020). To examine the effect of model update frequency, we perform a robustness check for HAR-D models in Appendix D and we conclude that the update frequency has insignificant effect on the model's performance. Our testing period starts from July 1, 2015 until June 30, 2016, and the corresponding training and validation samples are [July 1, 2011, June 30, 2014] and [July 1, 2014, June 30, 2015], respectively. When we predict the RV in [July 1, 2016, June 30, 2017], the training and validation samples are [July 1, 2011, June 30, 2015] and [July 1, 2015, June 30, 2016], respectively. Therefore, our testing sample includes 6 years, from July 2015 to June 2021.

For HAR-D and OLS, we use both the training and validation data for training, due to no requirement of hyperparameter tuning. Given the stochastic nature of NNs, we apply an ensemble approach to MLPs and LSTMs for improving their robustness (see Hansen and Salamon 1990; Gu, Kelly, and Xiu 2020). Specifically, we train multiple NNs with different random seeds for initialization, and construct final predictions by averaging the outputs of all networks. For more information on the model settings, see Appendix B.

In all of the models, the features are based on the observations in the last 21 days. Prior to inputting variables in the models, at each rolling window estimation, we normalize them by removing the mean and scaling to unit variance.

### 5.2. Main Results

Table 3 presents the results of each model under three training schemes.[8] Due to limited computation power, MLPs and LSTMs are only performed under the *Universal* and *Augmented* settings. We draw the following conclusions from Table 3.

We begin by comparing the SARIMA with HAR-D under the Single setting.[9] The results show that the ARIMA model achieves similar performance as HAR over the 1-day horizon, consistent with Izzeldin et al. (2019). However, HAR-D yields more accurate out-of-sample forecasts than SARIMA across intraday horizons, that is, 10-min, 30-min, and 65-min.

Regarding linear models, we observe that for HAR-D, Universal shows no improvement in forecasting, compared with Single. HAR-D models trained under Augmented significantly outperform the ones trained under the other two schemes, across all horizons in our study. The average reduction in QLIKE of Augmented compared with Single is 0.031, −0.005, 0.004, and 0.010 over 10-min, 30-min, 65-min, and 1-day, respectively.

Generally speaking, there are significant improvements when moving from HAR-D models to OLS models, over 10-min, 30-min, and 65-min horizons. For example, QLIKEs

8 To more formally assess the statistical significance of the differences in out-of-sample volatility forecasts, Table C.1 in Appendix C also reports the results of all DM tests in terms of QLIKE.

9 Note that (S)ARIMA is for Single time series.

**Table 3.** Out-of-sample performance

| Panel A | | 10-min | | 30-min | | 65-min | | 1-day | |
|---|---|---|---|---|---|---|---|---|---|
| Statistical performance | | MSE | QLIKE | MSE | QLIKE | MSE | QLIKE | MSE | QLIKE |
| SARIMA | Single | 1.319 | 0.617 | 0.524 | 0.338 | 0.362 | 0.231 | 0.268 | 0.190 |
| HAR-D | Single | 1.013 | 0.484 | 0.332 | 0.222 | 0.270 | 0.190 | 0.269 | 0.190 |
| | Universal | 1.021 | 0.518 | 0.333 | 0.230 | 0.270 | 0.191 | 0.269 | 0.190 |
| | Augmented | 0.995 | 0.453 | 0.323 | 0.227 | 0.262 | 0.186 | 0.257 | 0.180 |
| OLS | Single | 1.009 | 0.490 | 0.307 | 0.222 | 0.251 | 0.176 | 0.263 | 0.192 |
| | Universal | 1.008 | 0.507 | 0.307 | 0.221 | 0.250 | 0.175 | 0.260 | 0.191 |
| | Augmented | 0.962 | 0.430 | 0.293 | 0.204 | 0.241 | 0.171 | 0.254* | 0.178* |
| LASSO | Single | 1.053 | 0.492 | 0.325 | 0.224 | 0.252 | 0.180 | 0.263 | 0.195 |
| | Universal | 1.012 | 0.511 | 0.309 | 0.222 | 0.251 | 0.176 | 0.261 | 0.192 |
| | Augmented | 0.961 | 0.428 | 0.293 | 0.204 | 0.242 | 0.172 | 0.255* | 0.179* |
| XGBoost | Single | 1.047 | 0.539 | 0.345 | 0.236 | 0.297 | 0.201 | 0.358 | 0.217 |
| | Universal | 0.968 | 0.417 | 0.290 | 0.191 | 0.242 | 0.170 | 0.268 | 0.192 |
| | Augmented | 0.968 | 0.422 | 0.297 | 0.190 | 0.249 | 0.174 | 0.285 | 0.187 |
| MLP | Single | – | – | – | – | – | – | – | – |
| | Universal | 0.947 | 0.397 | 0.284 | 0.181 | 0.232 | 0.163 | 0.260 | 0.191 |
| | Augmented | 0.945 | 0.386 | 0.280* | 0.179 | 0.229* | 0.162 | 0.257 | 0.180 |
| LSTM | Single | – | – | – | – | – | – | – | – |
| | Universal | 0.950 | 0.393 | 0.287 | 0.179 | 0.232 | 0.162 | 0.261 | 0.188 |
| | Augmented | 0.934* | 0.376* | 0.279* | 0.171* | 0.229* | 0.160* | 0.258 | 0.182 |
| Panel B | | 10-min | | 30-min | | 65-min | | 1-day | |
| Realized utility | | RU | RU-TC | RU | RU-TC | RU | RU-TC | RU | RU-TC |
| SARIMA | Single | 2.095 | 1.473 | 3.041 | 2.624 | 3.314 | 2.861 | 3.550 | 3.515 |
| HAR-D | Single | 2.690 | 2.065 | 3.457 | 3.040 | 3.542 | 3.095 | 3.548 | 3.515 |
| | Universal | 2.574 | 1.975 | 3.429 | 3.016 | 3.541 | 3.095 | 3.547 | 3.514 |
| | Augmented | 2.790 | 2.280 | 3.428 | 3.022 | 3.552 | 3.107 | 3.571 | 3.536 |
| OLS | Single | 2.660 | 1.901 | 3.506 | 3.027 | 3.579 | 3.118 | 3.543 | 3.504 |
| | Universal | 2.601 | 2.039 | 3.432 | 2.984 | 3.580 | 3.127 | 3.546 | 3.513 |
| | Augmented | 2.845 | 2.271 | 3.485 | 3.036 | 3.587 | 3.130 | 3.576 | 3.536 |
| LASSO | Single | 2.631 | 1.893 | 3.526 | 3.061 | 3.567 | 3.108 | 3.545 | 3.501 |
| | Universal | 2.593 | 2.044 | 3.432 | 2.989 | 3.578 | 3.126 | 3.543 | 3.512 |
| | Augmented | 2.852 | 2.292 | 3.487 | 3.046 | 3.586 | 3.132 | 3.575 | 3.537 |
| XGBoost | Single | 2.492 | 1.552 | 3.408 | 2.888 | 3.520 | 3.039 | 3.508 | 3.449 |
| | Universal | 2.890 | 2.200 | 3.532 | 3.067 | 3.592 | 3.116 | 3.546 | 3.505 |
| | Augmented | 2.864 | 2.212 | 3.545 | 3.083 | 3.581 | 3.109 | 3.571 | 3.524 |
| MLP | Single | – | – | – | – | – | – | – | – |
| | Universal | 2.952 | 2.380 | 3.564 | 3.119 | 3.607 | 3.139 | 3.543 | 3.506 |
| | Augmented | 2.993 | 2.442 | 3.569 | 3.126 | 3.609 | 3.145 | 3.571 | 3.534 |
| LSTM | Single | – | – | – | – | – | – | – | – |
| | Universal | 2.975 | 2.455 | 3.575 | 3.144 | 3.610 | 3.149 | 3.552 | 3.514 |
| | Augmented | 3.028 | 2.532 | 3.595 | 3.170 | 3.614 | 3.166 | 3.567 | 3.533 |

*Notes:* The table reports the out-of-sample results for predicting future RV over multiple horizons using different models under three training schemes. For each horizon, the model with the best (second best) out-of-sample performance in QLIKE (in Panel A)/RU (in Panel B) is highlighted in red (blue), respectively. An asterisk (*) indicates models that are included in the MCS at the 5% significance level.

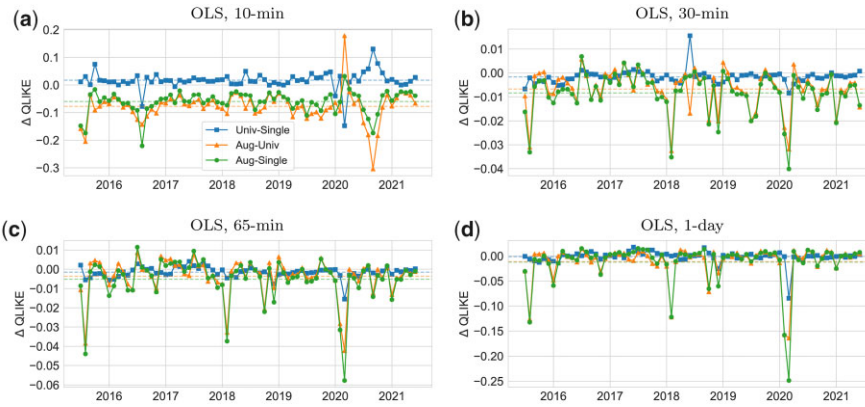


**Figure 7.** Pairwise ΔQLIKE of the OLS model across three training schemes.

*Notes:* ΔQLIKE is averaged across stocks in each month during the testing period July 2015–June 2021. The dashed horizontal lines represent the average reductions in QLIKE.

are reduced from 0.453 (respectively, 0.227, 0.186) with the best HAR-D model (i.e., under Augmented) to 0.430 (respectively, 0.204, 0.171) with the best OLS model (i.e., under Augmented), across the three horizons (i.e., 10-, 30-, and 65-min), respectively. Within the OLS models, conclusions are similar with HAR-D models, that is, no benefits from Universal while significant benefits from Augmented. We also observe similar findings in LASSO as in OLS, suggesting that regularization does not further aid performance. On the other hand, MLPs and LSTMs achieve state-of-the-art accuracy across all measures and intraday horizons (i.e., 10-, 30-, and 65-min), implying the complex interactions between predictors. Further analysis is provided in Section 5.3.

Interestingly, linear models slightly outperform MLPs and LSTMs at the 1-day horizon. This is perhaps expected, and might be due to the availability of only a small amount of data at the 1-day horizon, rendering the NNs to underperform due to lack of training data.

Echoing the findings from Panel A, OLS based on the 21-day rolling daily RVs deliver the higher utility than the HAR-type models, consistent with Bollerslev et al. (2018). NNs still perform the best, with the highest realized utility achieved by LSTMs.

Let us now consider the OLS model as an illustrative example for understanding the relative reduction in error. We compare its QLIKEs under these three schemes, at a monthly level, as shown in Figure 7. For better readability, we report the reduction in error of Universal relative to Single (denoted as Univ–Single), the reduction of Augmented relative to Universal (denoted as Aug–Univ), and the reduction of Augmented relative to Single (denoted as Aug–Single). Note that Aug–Single = (Aug–Univ) + (Univ–Single). Negative values of ΔQLIKE indicate an improvement on out-of-sample data and positive values indicate degradation. To arrive at this figure, we average the ΔQLIKE values in each month, across stocks. Figure 7 reveals that the improvement of Universal compared with Single is relatively small but consistent. In terms of the benefits of Augmented, it is typically the case that incorporating the market volatility as an additional feature helps improve the forecasting performance, especially for turmoil periods.

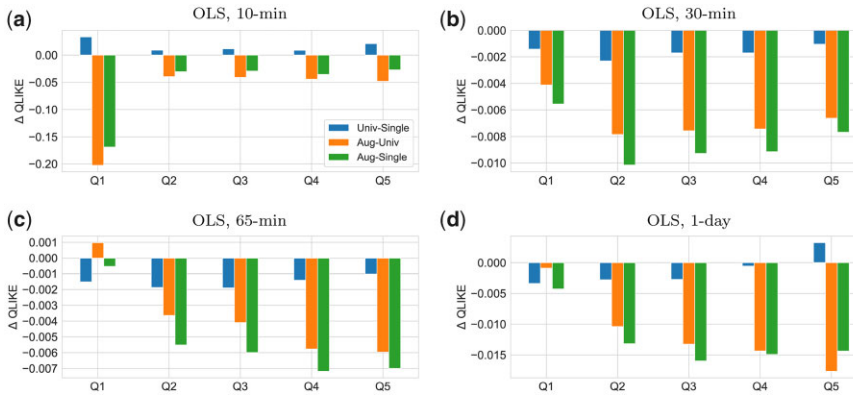


**Figure 8.** Pairwise ΔQLIKE of the OLS model sorted by commonality.

*Notes:* Q1, respectively Q5, denotes the subset of stocks with the lowest, respectively, highest, 20% values for the commonality.

An interesting question to investigate is whether the improvements of Universal or Augmented for individual stocks are associated with their commonality with the market volatility. To this end, we present the results in Figure 8 for each quintile bucket, sorted by stock commonality (computed from Equation 5). From this figure, we observe that the reduction of Augmented in out-of-sample QLIKE relative to Universal is explained by commonality to a large extent. Generally, the out-of-sample QLIKE is expected to decline steadily for stocks with higher commonality. Another interesting result arises from Figure 8(a), where we observe that Universal and Augmented actually reduce the out-of-sample QLIKE more for stocks (in the Q1 bucket) that are most loosely connected to the market in terms of 10-min volatility. Further research should be undertaken to investigate this finding.

## 5.3. Variable Importance and Interaction Effects

This section provides intuition for why NNs perform as strongly as they do, with an eye toward explainability. Due to the use of non-linear activation functions and multiple hidden layers, NNs enjoy the benefit of allowing for potentially complex interactions among predictors, albeit at the cost of considerably reducing model interpretability. To better understand such a "black-box" technique, we provide the following analysis to help illustrate the inner workings of NNs and explain their competitive performance.

### 5.3.1 Relative importance of predictors

In order to identify which variables are the most important for the prediction task at hand, we construct a metric (see Sadhwani, Giesecke, and Sirignano 2021) based on the sum of absolute partial derivatives (Sensitivity) of the predicted volatility. In particular, to quantify the importance of the $k$-th predictor, we compute

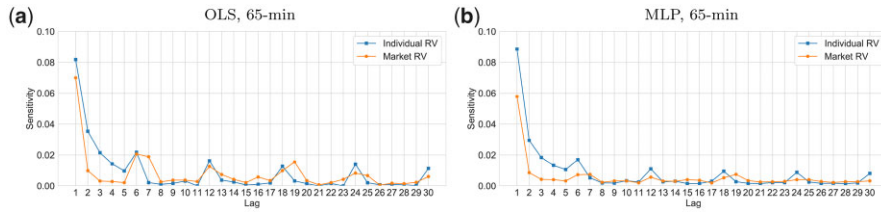


**Figure 9.** Relative importance of lagged individual and market RVs.

*Notes:* For ease of readability, we only report the sensitivity values for the most recent 30 lagged RVs (i.e., in the last five days for 65-min horizon).

$$\text{Sensitivity}_k = \sum_{i=1}^{N} \sum_{t \in \mathcal{T}_{\text{train}}} |\frac{\partial F}{\partial u_k}|_{\mathbf{u}=\mathbf{u}_{i,t}} \tag{22}$$

Here, $F$ is the fitted model under the Augmented scheme, $\mathbf{u}$ represents the vector of predictors, and $u_k$ is the $k$-th element in $\mathbf{u}$. $\mathbf{u}_{i,t}$ represents the input features of stock $i$ at time $t$. We normalize the sensitivity of all variables such that they sum up to one. In a special case of linear regression, the sensitivity measure is the normalized absolute slope coefficient.

Considering the 65-min scenario as an example, Figure 9 reveals that for both OLS and MLP, there has been a tendency of the lagged features to decline in terms of sensitivity, as the lag increases. Additionally, we observe that the sensitivity values rise to a high point at every six lags, corresponding to one day. A distinct difference between the sensitivity values implied by OLS and the ones implied by MLP is that the latter places more weight on the lag = 1 individual RV (Sensitivity = 0.90) and less on the lag = 1 market RV (Sensitivity = 0.059). On the other hand, for OLS, the sensitivities of lag = 1 individual (respectively, market) RV are 0.081 (respectively, 0.069).

### 5.3.2 Interaction effects

To analyze the interactions between the two most significant features implied by NNs, we adopt an approach (e.g., Gu, Kelly, and Xiu 2020; Choi, Jiang, and Zhang 2021) that focuses on the partial relations between a pair of input variables and responses, while fixing other variables at their mean values

$$F_{j|i}(u_j|u_i = q) = F(u_j, u_i = q, u_k = \overline{u}_k, k \neq i, j), \tag{23}$$

where $q$ represent the quantile values for the $i$-th predictor $u_i$.

Figure 10 illustrates how predicted volatility (i.e., the fitted values) varies as a function of the pair of predictor variables $\text{RV}_{i,t}^{(h)}$ and $\text{RV}_{M,t}^{(h)}$, over their support.[10] In particular, we analyze the interaction of the lag = 1 individual RV ($\text{RV}_{i,t}^{(h)}$), with the lag = 1 market RV ($\text{RV}_{M,t}^{(h)}$). As shown in Figure 10(a), a parallel shift between the different curves occurs if there are no interaction effects between $\text{RV}_{i,t}^{(h)}$ and $\text{RV}_{M,t}^{(h)}$. Figure 10(b) first reveals that the predicted volatility is non-linear in $\text{RV}_{i,t}^{(h)}$ and the slope of that relationship becomes higher after $\text{RV}_{i,t}^{(h)}$ exceeds a certain threshold (around 0.5). Furthermore, it demonstrates clear

10 Recall that the variables are normalized by removing the mean and scaling to unit variance.

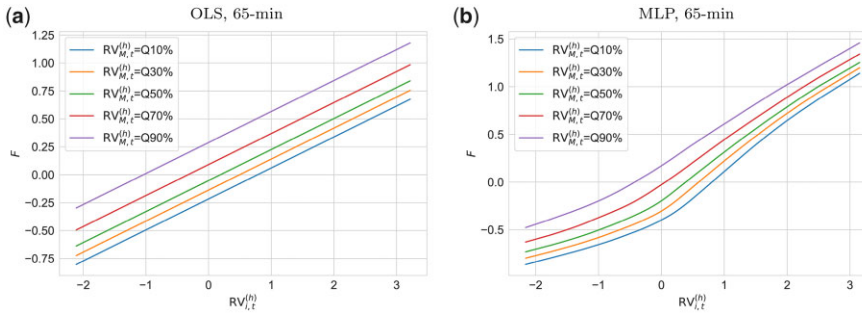


**Figure 10.** Interactions between the lagged individual and market RV.

*Notes:* The figure plots the pattern of predicted RV (*y*-axis) as a function of the lag = 1 individual RV (*x*-axis) conditioned on various lag = 1 market RV quantile values (keeping all other variables at their mean values).

interaction effects between $\mathrm{RV}^{(b)}_{i,t}$ and $\mathrm{RV}^{(b)}_{M,t}$. As it can be observed from the rightmost region of Figure 10(b), the distances between the curves become relatively smaller, conveying the message that, when an individual stock is very volatile, the market effect on it weakens.

## 5.4. Forecasting RVs of Unseen Stocks

To examine the ability to generalize and address concerns regarding overfitting, we perform a stringent out-of-sample test, that is, using the existent trained models to forecast the volatility of new stocks that have not been included in the training sample, in the spirit of Sirignano and Cont (2019) and Choi, Jiang, and Zhang (2021). For better distinction, we denote the stocks used for estimating ML models as *raw stocks* and those new stocks not in the training sample as *unseen stocks*.[11] We follow the procedure of training, validation, and testing periods described in Section 5.1. Specifically, to predict the RVs of unseen stocks in a particular year, we train and validate the models using the past data of raw stocks exclusively.

In this experiment, we choose OLS models trained for each unseen stock as the baseline. The results are shown in Table 4. Note that models trained under Single cannot be applied to forecast unseen stocks, since they are trained for each specific raw stock individually. From Table 4, we conclude that NNs trained on the pooled data of raw stocks have better forecasting performance compared with baselines, across all horizons. This presents new empirical evidence for a *universal volatility mechanism* among stocks. Furthermore, NNs significantly outperform other methods across three metrics, over 10-min, 30-min, and 65-min forecasting horizons, thus validating their robustness. Concerning the 1-day scenario, NNs obtain comparable results (QLIKE = 0.252) to the best non-NN model (MSE = 0.249, attained by LASSO). The realized utility of the different risk models echoes that of the out-of-sample QLIKE.

11 The set of unseen stocks includes the following 16 tickers: AMAT, APD, BIIB, COF, DE, EQIX, EW, GPN, HUM, ICE, ILMN, ITW, NOC, NSC, PLD, and SLB.

**Table 4.** Performance on unseen stocks

| Panel A | | 10-min | | 30-min | | 65-min | | 1-day | |
|---|---|---|---|---|---|---|---|---|---|
| Statistical performance | | MSE | QLIKE | MSE | QLIKE | MSE | QLIKE | MSE | QLIKE |
| OLS | Unseen | 0.664 | 0.372 | 0.329 | 0.219 | 0.287 | 0.205 | 0.348 | 0.254 |
| OLS | Universal | 0.678 | 0.410 | 0.328 | 0.223 | 0.286 | 0.206 | 0.343 | 0.260 |
| | Augmented | 0.639 | 0.359 | 0.317 | 0.222 | 0.278 | 0.208 | 0.327* | 0.249* |
| LASSO | Universal | 0.683 | 0.419 | 0.330 | 0.225 | 0.286 | 0.207 | 0.344 | 0.261 |
| | Augmented | 0.639 | 0.359 | 0.317 | 0.222 | 0.278 | 0.208 | 0.327* | 0.249* |
| XGBoost | Universal | 0.655 | 0.476 | 0.314 | 0.206 | 0.278 | 0.201 | 0.353 | 0.266 |
| | Augmented | 0.654 | 0.509 | 0.320 | 0.221 | 0.282 | 0.206 | 0.364 | 0.255 |
| MLP | Universal | 0.623 | 0.328 | 0.306 | 0.203* | 0.266 | 0.193* | 0.342 | 0.266 |
| | Augmented | 0.623 | 0.332 | 0.301* | 0.203* | 0.263* | 0.194* | 0.329 | 0.252 |
| LSTM | Universal | 0.637 | 0.348 | 0.311 | 0.211 | 0.267 | 0.195 | 0.339 | 0.265 |
| | Augmented | 0.622* | 0.326* | 0.303 | 0.205 | 0.263* | 0.194 | 0.332 | 0.255 |
| Panel B: Realized utility | | 10-min | | 30-min | | 65-min | | 1-day | |
| | | RU | RU-TC | RU | RU-TC | RU | RU-TC | RU | RU-TC |
| OLS | Unseen | 3.107 | 1.996 | 3.475 | 2.672 | 3.503 | 2.715 | 3.385 | 3.320 |
| OLS | Universal | 2.988 | 2.280 | 3.461 | 2.700 | 3.498 | 2.712 | 3.363 | 3.298 |
| | Augmented | 3.138 | 2.355 | 3.459 | 2.710 | 3.487 | 2.712 | 3.389 | 3.311 |
| LASSO | Universal | 2.959 | 2.270 | 3.457 | 2.704 | 3.496 | 2.712 | 3.359 | 3.296 |
| | Augmented | 3.137 | 2.376 | 3.458 | 2.720 | 3.485 | 2.716 | 3.389 | 3.315 |
| XGBoost | Universal | 2.688 | 1.640 | 3.510 | 2.711 | 3.511 | 2.701 | 3.349 | 3.269 |
| | Augmented | 2.563 | 1.578 | 3.464 | 2.688 | 3.495 | 2.680 | 3.388 | 3.302 |
| MLP | Universal | 3.233 | 2.396 | 3.515 | 2.736 | 3.529 | 2.730 | 3.340 | 3.266 |
| | Augmented | 3.221 | 2.444 | 3.514 | 2.749 | 3.522 | 2.735 | 3.378 | 3.302 |
| LSTM | Universal | 3.167 | 2.415 | 3.493 | 2.769 | 3.523 | 2.737 | 3.345 | 3.271 |
| | Augmented | 3.238 | 2.533 | 3.507 | 2.787 | 3.524 | 2.762 | 3.371 | 3.302 |

*Notes:* The table reports the out-of-sample results for predicting future RV of unseen stocks over multiple horizons using different models under three training schemes. The row OLS Unseen represents the baseline results based on OLS models estimated for each unseen stock. Other rows represent the results of models estimated on raw stocks under the Universal and Augmented settings. For each horizon, the model with the best (second best) out-of-sample performance in terms of QLIKE (in Panel A)/RU (in Panel B) is highlighted in red (blue), respectively. An asterisk (*) indicates models that are included in the MCS at the 5% significance level.

## 6 Forecasting Daily RVs with Intraday RVs

Given the fact that intraday volatility exhibits a high and stable commonality (see Sections 2 and 3), we are interested in the potential benefits of using past intraday RVs to forecast daily RVs.

### 6.1. Closely Related Literature

Generally speaking, there are two broad families of models used to forecast daily volatility: (i) GARCH and SV models that employ daily returns and (ii) models that use daily RVs. Previous well-established studies have shown that due to the utilization of available

intraday information, daily RV is a superior proxy for the unobserved daily volatility, when compared with the parametric volatility measures generated from the GARCH and SV models of daily returns (see Barndorff-Nielsen and Shephard 2002; Andersen et al. 2003; Izzeldin et al. 2019). It is worth noting that in these traditional forecasting daily RV models (e.g., ARFIMA of Andersen et al. 2003; HAR of Corsi 2009; SHAR of Patton and Sheppard 2015; and HARQ of Bollerslev, Patton, and Quaedvlieg 2016), only past daily RVs (or their alternatives) are included as predictors. Even though this is a mainstream approach in the literature, it does not benefit to the full extent from the availability of intraday data. In the presidential address of SoFiE 2021, Bollerslev (2022) also pointed out that "semivariation measured over shorter interday time intervals may afford additional useful information."

Intraday RV information is also studied for forecasting the 1-day-ahead volatility in several previous works, for example, the mixed data sampling (MIDAS) approach of Ghysels, Santa-Clara, and Valkanov (2005, 2006) and the "Rolling" approach of Pascalau and Poirier (2021). In particular, the classic MIDAS approach uses smooth-distributed lag polynomials of high-frequency predictors to forecast the low-frequency target variables, in the form $\mathrm{RV}_{i,t+1}^{(d)} = \beta_{i,0} + \beta_{i,1}[a(1)^{-1}a(L)]\mathrm{RV}_{i,t}^{(d)} + \epsilon_{i,t+1}$, where the $a(L)$ lag polynomial is defined by scaled beta functions. Ghysels, Santa-Clara, and Valkanov (2006) find that the direct use of high-frequency data does not improve volatility predictions compared with the forecasts from a model based on daily RVs only. We reckon it is due to the restricted flexibility of MIDAS models, usually with one or two parameters determining the pattern of the weights, therefore missing the time-of-day effect of intraday RVs. Pascalau and Poirier (2021) increase the training samples by rolling a fixed window of intraday returns over consecutive trading days by adding and dropping one intraday return at each end. They claim that their proposed "Rolling" approach could potentially capture the changing dynamics of serial correlation throughout the trading day; thus, leading to improved volatility forecasts.

### 6.2. Proposed Approach

In Section 4.1, we introduced a set of commonly used models, where the daily variables (lagged daily RVs) are employed as predictors when forecasting 1-day RVs. For simplicity, we refer to these models as *traditional* approaches in this section. Previous sections, such as Figure 9, concluded that the most recent RV plays a more important role in forecasting future volatility. Motivated by the fact that intraday volatility has a high and stable commonality, we propose a new prediction approach for forecasting daily volatility using past intraday RVs as predictors, denoted by *Intraday2Daily* approach.

In contrast to Ghysels, Santa-Clara, and Valkanov (2006) and Pascalau and Poirier (2021), our *Intraday2Daily* approach takes the time-of-day effect into account in an explicit way and posits the model

$$\mathrm{RV}_{i,t+1}^{(d)} = F_i\left(\mathrm{RV}_{i,t}^{(h)}, \ldots, \mathrm{RV}_{i,t-(p-1)h}^{(h)}, \mathrm{RV}_{i,t-1}^{(d)}, \ldots, \mathrm{RV}_{i,t-(p-1)}^{(d)}; \theta\right) + \epsilon_{i,t+1}. \tag{24}$$

Here, $\left(\mathrm{RV}_{i,t}^{(h)}, \ldots, \mathrm{RV}_{i,t-(k-1)h}^{(h)}\right)$ represent the past RVs for stock $i$ computed over shorter intraday time horizons $h$ at day $t$ and $\left(\mathrm{RV}_{i,t-1}^{(d)}, \ldots, \mathrm{RV}_{i,t-(p-1)}^{(d)}\right)$ are past daily RVs of stock $i$ up to day $t-1$. Departing from traditional models where all the variables are computed in the daily frequency, we decompose the lag-one daily $\mathrm{RV}_{i,t}^{(d)}$ to sub-sampled RVs, that is, $\left(\mathrm{RV}_{i,t}^{(h)}, \ldots, \mathrm{RV}_{i,t-(k-1)h}^{(h)}\right)$. Under the Augmented training scheme, we also incorporate the

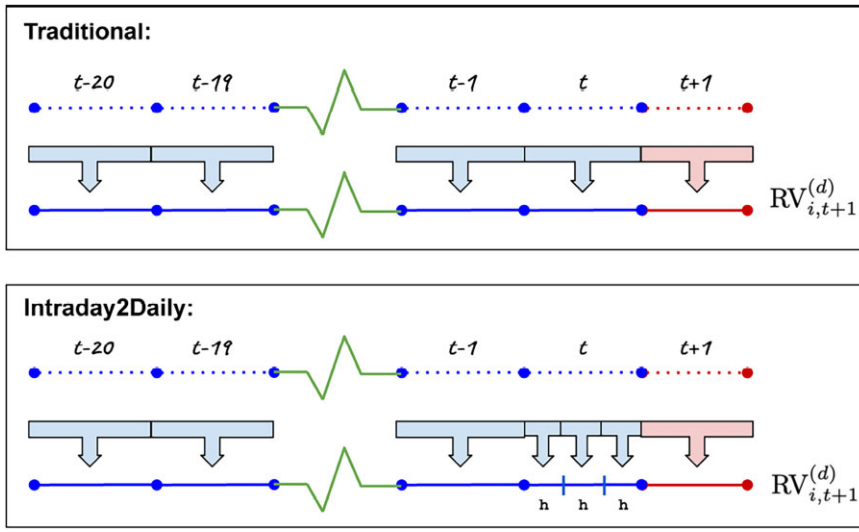


**Figure 11.** Illustration of two prediction approaches for future daily volatility (rightmost segment at day $t$+1).

*Notes:* In each box, dots in the top line represent the intraday returns. The traditional approaches employ the aggregated daily (or weekly, or monthly) RVs (the remaining left segments) as predictors, while the Intraday2Daily approach employs intraday RVs (the short segments marked with $h$ at day $t$). $h$ represents the horizon of intraday RVs. In this example, $h = 130$ min.

market volatilities into models. Figure 11 illustrates the comparison between the traditional approach and our Intraday2Daily approach.

The advantages of the Intraday2Daily approach over traditional approaches can be summarized as follows. First, the Intraday2Daily approach significantly enriches the information content of daily volatility. Second, it contributes to the literature in the modeling of daily volatility by examining the coefficients of intraday RVs. Third, the essential idea underlying the Intraday2Daily approach can be possibly applied to estimate other daily risk measures, such as value-at-risk (VaR), etc. For example, one may use half-hour VaRs to forecast the 1-day-ahead VaR. Finally, practitioners can better adjust their portfolios with more accurate forecasts from the Intraday2Daily approach rather than traditional approaches. To the best of our knowledge, this is the first study to explicitly investigate the predictive power of intraday RVs on daily volatility and to demonstrate the additional accuracy improvements it brings to the forecasting task.

### 6.3. Experiments

The forecasting performance of traditional approaches with daily variables is already summarized in the column "1-day" of Table 3. Table 5 reports the results of models combined with the Intraday2Daily approach.[12] In other words, models in Table 5 use sub-sampled intraday

12 We observe similar findings when applying the Intraday2Daily approach to forecast the raw volatilities (not in logs).

**Table 5.** Out-of-sample performance of the Intraday2Daily approach

| Panel A: | | 10-min | | 30-min | | 65-min | |
|---|---|---|---|---|---|---|---|
| Statistical performance | | MSE | QLIKE | MSE | QLIKE | MSE | QLIKE |
| HAR | Single | 0.259 | 0.189 | 0.252 | 0.185 | 0.252 | 0.185 |
| | Universal | 0.270 | 0.197 | 0.255 | 0.187 | 0.253 | 0.186 |
| | Augmented | 0.256 | 0.179 | 0.249 | 0.174 | 0.249 | 0.175 |
| OLS | Single | 0.255 | 0.186 | 0.252 | 0.186 | 0.253 | 0.186 |
| | Universal | 0.253 | 0.186 | 0.252 | 0.187 | 0.252 | 0.186 |
| | Augmented | 0.249 | 0.173 | 0.248 | 0.173 | 0.249 | 0.173 |
| LASSO | Single | 0.262 | 0.194 | 0.251 | 0.185 | 0.253 | 0.186 |
| | Universal | 0.261 | 0.191 | 0.248 | 0.187 | 0.248 | 0.186 |
| | Augmented | 0.273 | 0.203 | 0.247 | 0.173 | 0.247 | 0.173 |
| XGBoost | Single | 0.323 | 0.204 | 0.330 | 0.201 | 0.332 | 0.200 |
| | Universal | 0.261 | 0.177 | 0.257 | 0.173 | 0.257 | 0.173 |
| | Augmented | 0.264 | 0.179 | 0.261 | 0.176 | 0.266 | 0.176 |
| MLP | Single | – | – | – | – | – | – |
| | Universal | 0.243* | 0.171* | 0.242* | 0.171* | 0.246 | 0.172 |
| | Augmented | 0.247 | 0.174 | 0.246 | 0.175 | 0.247 | 0.176 |
| LSTM | Single | – | – | – | – | – | – |
| | Universal | 0.247 | 0.174 | 0.244 | 0.171* | 0.244 | 0.171 |
| | Augmented | 0.258 | 0.184 | 0.249 | 0.175 | 0.250 | 0.176 |
| Panel B: | | 10-min | | 30-min | | 65-min | |
| Realized utility | | RU | RU-TC | RU | RU-TC | RU | RU-TC |
| HAR | Single | 3.558 | 3.524 | 3.567 | 3.533 | 3.566 | 3.531 |
| | Universal | 3.539 | 3.521 | 3.563 | 3.534 | 3.565 | 3.532 |
| | Augmented | 3.562 | 3.532 | 3.573 | 3.541 | 3.570 | 3.536 |
| OLS | Single | 3.565 | 3.526 | 3.565 | 3.530 | 3.564 | 3.529 |
| | Universal | 3.565 | 3.534 | 3.564 | 3.535 | 3.565 | 3.533 |
| | Augmented | 3.581 | 3.548 | 3.583 | 3.551 | 3.583 | 3.547 |
| LASSO | Single | 3.561 | 3.526 | 3.565 | 3.531 | 3.564 | 3.529 |
| | Universal | 3.561 | 3.524 | 3.564 | 3.535 | 3.573 | 3.536 |
| | Augmented | 3.551 | 3.515 | 3.565 | 3.531 | 3.585 | 3.548 |
| XGBoost | Single | 3.532 | 3.474 | 3.545 | 3.486 | 3.550 | 3.489 |
| | Universal | 3.584 | 3.532 | 3.593 | 3.543 | 3.590 | 3.547 |
| | Augmented | 3.579 | 3.527 | 3.586 | 3.535 | 3.589 | 3.538 |
| MLP | Single | – | – | – | – | – | – |
| | Universal | 3.586 | 3.543 | 3.592 | 3.549 | 3.593 | 3.550 |
| | Augmented | 3.562 | 3.521 | 3.583 | 3.541 | 3.581 | 3.540 |
| LSTM | Single | – | – | – | – | – | – |
| | Universal | 3.586 | 3.543 | 3.592 | 3.549 | 3.593 | 3.550 |
| | Augmented | 3.562 | 3.521 | 3.583 | 3.541 | 3.581 | 3.540 |

*Notes:* The table reports the out-of-sample results for predicting future daily RV using different models under three training schemes when combined with the Intraday2Daily approach. The columns ("10-min," "30-min," and "65-min") represent the frequency of predictor features and the dependent variable in this table always corresponds to future daily volatility. The model with the best (second best) out-of-sample performance in QLIKE (in Panel A)/RU (in Panel B) is highlighted in red (blue), respectively. An asterisk (*) indicates models that are included in the MCS at the 5% significance level.

RVs rather than the lag-one total RV in the column "1-day" of Table 3. For example, the lag-one total RV in HAR (Equation 9) is replaced by non-overlapped intraday RVs.

By comparing the column "1-day" of Table 3 with Table 5, we establish that the Intraday2Daily approach generally helps improve the out-of-sample performance of benchmark models. For example, under the Single setting, compared with OLS using daily RVs (QLIKE = 0.192), 65-min RVs improve the out-of-sample performance (QLIKE = 0.186).

MLPs with intraday RVs again achieve the best out-of-sample performance. For example, the QLIKEs of MLPs under Universal are {0.171, 0.171, 0.172} using {10-min, 30-min, 65-min} RVs as predictors, respectively. The superior performance of MLPs over linear regressions when using intraday RVs further demonstrates the advantages of NNs to learn unknown dynamics in financial markets.

In general, the improvements of the Intraday2Daily approach lead to higher utilities. For instance, when considering the column "1-day" of Panel B in Table 3, we observe that the RU (respectively, RU-TC) values of OLS under Augmented are 3.576% (respectively, 3.536%). OLS with 65-min RVs as predictors obtain higher RU (respectively, RU-TC) values of 3.583% (respectively, 3.547%). Overall, MLPs deliver the highest utility (RU = 3.593%, RU-TC = 3.550%) based on the 65-min intraday RVs, followed by LSTMs, thus hinting at potential non-linearity and complex interactions inherent in the data.

## 6.4. Robustness Check

In this section, we present the empirical analysis of examining the robustness of the Intraday2Daily approach when incorporating new types of predictors (including semi-RV of Patton and Sheppard [2015] and realized quarticity [RQ] of Bollerslev, Patton, and Quaedvlieg [2016]).

### 6.4.1 Semi-variance-HAR

Patton and Sheppard (2015) proposed the semi-variance-HAR (SHAR) model as an extension of the standard HAR model (see further details in Section 4.1.2), in order to exploit the well-documented leverage effect by decomposing the total RV of the first lag via signed intraday returns, as shown in Equation (25) (see Barndorff-Nielsen, Kinnebrock, and Shephard 2008). In other words, the lag-one RV in SHAR (Equation 26) is split into the sum of squared positive returns and the sum of squared negative returns, as follows:

$$\begin{aligned} \mathrm{RV}_{i,t}^{(d)+} &= \textstyle\sum_{l=0}^{M-1} r_{i,t-l\cdot\Delta}^2 I_{\{r_{t-l\cdot\Delta}>0\}}, \\ \mathrm{RV}_{i,t}^{(d)-} &= \textstyle\sum_{l=0}^{M-1} r_{i,t-l\cdot\Delta}^2 I_{\{r_{t-l\cdot\Delta}<0\}}, \end{aligned} \tag{25}$$

$$\mathrm{RV}_{i,t+1}^{(d)} = \alpha_i + \beta_i^{(d)+}\mathrm{RV}_{i,t}^{(d)+} + \beta_i^{(d)-}\mathrm{RV}_{i,t}^{(d)-} + \beta_i^{(w)}\mathrm{RV}_{i,w}^{(w)} + \beta_i^{(m)}\mathrm{RV}_{i,m}^{(m)} + \epsilon_{i,t+1}. \tag{26}$$

In the above, $\Delta$ denotes the interval for computing the intraday returns.

### 6.4.2 HARQ

Bollerslev, Patton, and Quaedvlieg (2016) pointed out that the beta coefficients in the HAR model may be affected by measurement errors in the realized volatilities. By exploiting the asymptotic theory for high-frequency RV estimation, the authors propose an easy-to-implement model, termed as HARQ (Equation 28). The RQ is estimated according to Equation (27), aiming to correct the measurement errors:

$$\mathrm{RQ}_{i,t}^{(d)} = \frac{M}{3}\sum_{l=0}^{M-1} r_{t-l\Delta}^{4} \tag{27}$$

$$\mathrm{RV}_{i,t+1}^{(d)} = \alpha_i + \left(\beta_i^{(d)} + \beta_i^{(d)\mathrm{Q}}\sqrt{\mathrm{RQ}_{i,t}^{(d)}}\right)\mathrm{RV}_{i,t}^{(d)} + \beta_i^{(w)}\mathrm{RV}_{i,t}^{(w)} + \beta_i^{(m)}\mathrm{RV}_{i,t}^{(m)} + \epsilon_{i,t+1}. \tag{28}$$

We compute the corresponding intraday variables of semi-RVs and RQs and then include them as new predictors in the Intraday2Daily approach. From Table 6, we first observe that the SHAR model generally performs as well as the standard HAR model (in Table 3), in line with Bollerslev, Patton, and Quaedvlieg (2016). HARQ outperforms HAR and SHAR, when applied to individual stocks studied in the present paper. Comparing the "Traditional" column with others, we conclude that in general, replacing the daily RVs with intraday RVs as predictors helps improve the out-of-sample performance of benchmark models.

## 6.5. Analysis of the Time-of-Day Dependent RV

To offer a more comprehensive understanding of the performance of *time-of-day dependent* RVs, we examine the coefficients of the Intraday2Daily OLS model trained under Augmented. Recall that before we input features into the model, we rescale them to have a mean of 0 and a standard deviation of 1. Hence, we can compare the coefficients of different lagged variables.

For better readability, we only report the first 13 = (390/30) coefficients of the OLS model using 30-min features in Figure 12, corresponding to the observations of RV in the most recent day.[13] We observe that the contributions of time-of-day dependent RVs are not even. Interestingly, *volatility near the close* (15:30–16:00) is the most important predictor, in contrast to the diurnal volatility pattern. These results shed new light on the modeling of volatility.

To explain why the most recent half-hour RV is the most important predictor for forecasting the next day's volatility, we provide a handful of perspectives. According to the *Wall Street Journal*,[14] there is a significant fraction of the total daily trading volume in the last half-hour of the trading day. For example, for the first few months of 2020 in the U.S. equity market, about 23% of trading volume in the 3,000 largest stocks by market value has taken place after 15:30. We also conclude from Figure 5 that the market achieves the highest level of consensus near the close. Therefore, volatility near the close in the previous trading day might contain more useful information for predicting the next day's volatility.

13 We attain similar results for models using intraday RVs based on other frequencies.

14 The 30-min that can make or break the trading day. https://www.wsj.com/articles/the-30-minutes-that-can-make-or-break-the-trading-day-11583886131 (accessed on February 28, 2022).

**Table 6.** Out-of-sample performance of the Intraday2Daily approach

| Panel A: | | 10-min | | 30-min | | 65-min | | Traditional | |
|---|---|---|---|---|---|---|---|---|---|
| Statistical performance | | MSE | QLIKE | MSE | QLIKE | MSE | QLIKE | MSE | QLIKE |
| SHAR | Single | 0.277 | 0.191 | 0.257 | 0.178 | 0.253 | 0.176 | 0.261 | 0.183 |
| | Universal | 0.285 | 0.198 | 0.263 | 0.183 | 0.255 | 0.178 | 0.261 | 0.182 |
| | Augmented | 0.261 | 0.181 | 0.253 | 0.175 | 0.250 | 0.174 | 0.254 | 0.178 |
| HARQ | Single | 0.264 | 0.204 | 0.254 | 0.178 | 0.253 | 0.176 | 0.256 | 0.179 |
| | Universal | 0.253 | 0.176 | 0.253 | 0.176 | 0.254 | 0.176 | 0.257 | 0.179 |
| | Augmented | 0.251 | 0.174 | 0.248* | 0.172* | 0.250 | 0.174 | 0.253 | 0.176 |
| Panel B: | | 10-min | | 30-min | | 65-min | | Traditional | |
| Realized utility | | RU | RU-TC | RU | RU-TC | RU | RU-TC | RU | RU-TC |
| SHAR | Single | 3.528 | 3.497 | 3.559 | 3.525 | 3.563 | 3.529 | 3.548 | 3.515 |
| | Universal | 3.510 | 3.499 | 3.548 | 3.525 | 3.560 | 3.529 | 3.550 | 3.516 |
| | Augmented | 3.563 | 3.533 | 3.576 | 3.545 | 3.578 | 3.545 | 3.571 | 3.537 |
| HARQ | Single | 3.467 | 3.425 | 3.556 | 3.520 | 3.564 | 3.528 | 3.557 | 3.525 |
| | Universal | 3.564 | 3.530 | 3.564 | 3.530 | 3.564 | 3.530 | 3.558 | 3.525 |
| | Augmented | 3.580 | 3.544 | 3.583 | 3.546 | 3.578 | 3.541 | 3.575 | 3.538 |

*Notes:* The table reports the out-of-sample results of SHAR and HARQ for predicting future daily RV under three training schemes. Columns "10-min," "30-min," and "65-min" represent the Intraday2Daily approach with different frequencies of predictors while the column "Traditional" represents that lagged daily RVs are used as predictors. The dependent variable in this table always corresponds to future daily volatility. The model with the best (second best) out-of-sample performance in QLIKE (in Panel A)/RU (in Panel B) is highlighted in red (blue), respectively. An asterisk (*) indicates models that are included in the MCS at the 5% significance level.

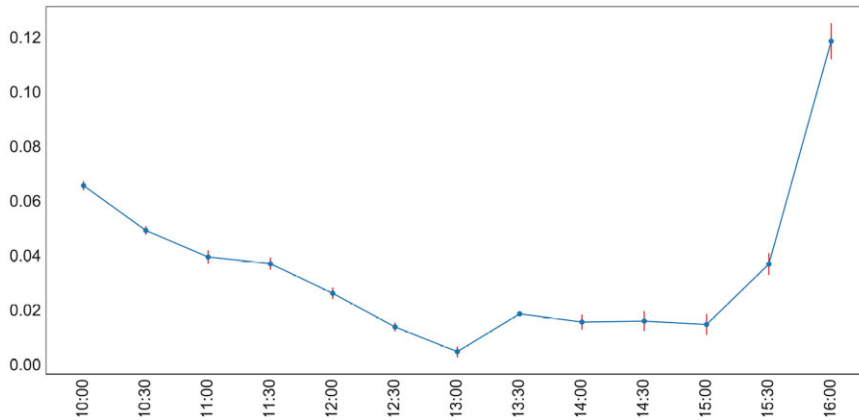


**Figure 12.** Coefficients of the Intraday2Daily OLS model under Augmented.

*Notes:* The Intraday2Daily OLS model uses lagged individual 30-min RVs to forecast the next day's volatility. The *x*-axis represents the time of day. The *y*-axis represents the coefficients of lagged RVs.

## 7 Conclusion

In this article, the commonality in intraday volatility over multiple horizons across the U.S. equity market is studied. By leveraging the information content of commonality, we have demonstrated that for most ML models in our analysis, pooling stock data together (Universal) and adding the market volatility as an additional predictor (Augmented) generally improve the out-of-sample performance, in comparison with asset-specific models (Single).

We show that NNs achieve superior performance, possibly due to their ability to uncover and model complex interactions among predictors. To alleviate concerns of overfitting, we perform a stringent out-of-sample test, applying the existent trained models to unseen stocks, and conclude that NNs still outperform traditional models.

Lastly and perhaps most importantly, motivated by the high commonality in intraday volatility, we propose a new approach (Intraday2Daily) to forecast daily RVs using past intraday RVs. The empirical findings suggest that the proposed Intraday2Daily approach generally yields superior out-of-sample forecasts. We further examine the coefficients in Intraday2Daily OLS models, and the results suggest that volatility near the close (15:30–16:00) in the previous day (lag = 1) is the most important predictor.

### 7.1. Future Research Directions

There are a number of interesting avenues to explore in future research. One direction pertains to the assessment of whether other characteristics, such as sector RVs, can improve the forecast of future RV, since in the present work, we have only considered the individual and market RVs. Another interesting direction is to apply the underlying idea of Intraday2Daily approach to other risk metrics, for example, VaR, that could potentially benefit from time-of-day dependent features.

## Appendix

## A: What May Drive Commonality in Volatility?

Previous studies, especially in the behavioral finance field, have shown that investor sentiments could affect stock prices (e.g., Kogan et al. 2006; Baker and Wurgler 2007; Hameed, Kang, and Viswanathan 2010; Da, Engelberg, and Gao 2011, 2015; Karolyi, Lee, and Van Dijk 2012; Bollerslev et al. 2018). Keynes (2018) argued that animal spirits affect consumer confidence, thereby moving prices in times of high levels of uncertainty. De Long et al. (1990), Shleifer and Summers (1990), and Kogan et al. (2006) found that investor sentiments induce excess volatility. Karolyi, Lee, and Van Dijk (2012) considered the investor sentiment index as an important source of commonality in liquidity. Bollerslev et al. (2018) found a monotonic relationship between volatility and sentiment, possibly driven by correlated trading. In this section, we are interested in the relation between investor sentiments and commonality in volatility.

Traditionally, there are two approaches to measuring investor sentiments (see Da, Engelberg, and Gao 2015), that is, market-based measures and survey-based indices. Following Baker and Wurgler (2007), we consider the daily market volatility index (VIX) from Chicago Board Options Exchange to be the market sentiment measure. We use the

Consumer Sentiment Index (CSI)[15] by the University of Michigan's Survey Research Center as a proxy for survey-based indices (see Carroll, Fuhrer, and Wilcox 1994; Lemmon and Portniaguina 2006). Generally speaking, CSI is a consumer confidence index, calculated by subtracting the percentage of unfavorable consumer replies from the percentage of favorable ones. Following Da, Engelberg, and Gao (2015), we also include a news-based index EPU[16] proposed by Baker, Bloom, and Davis (2016) to measure policy-related economic uncertainty.

As suggested by Morck, Yeung, and Yu (2000), the raw monthly commonality measures $R^2_{(h),m}$ (computed based on Equation 5) are inappropriate to use as the dependent variable in regressions, because they are bounded by 0 and 1. Consistent with Morck, Yeung, and Yu (2000), Karolyi, Lee, and Van Dijk (2012), and Dang, Moshirian, and Zhang (2015), we take the logistic transformation of $R^2_{(h),m}$, that is, $\log[R^2_{(h),m}/(1-R^2_{(h),m})]$, denoted by $(R^2_{(h),m})_L$, in our following empirical analysis. To explain the commonality in volatility, we regress $(R^2_{(h),m})_L$ against the aforementioned three indices, as shown in Equation (A1)

$$(R^2_{(h),m})_L = \alpha + \beta_1 \mathrm{CSI}_m + \beta_2 \mathrm{VIX}_m + \beta_3 \mathrm{EPU}_m + \epsilon_{i,t}. \tag{29}$$

Table A.1 reports the estimation results. First, we notice that a large proportion of the variance for the commonality is explained by these three sentiment factors. For example, the commonality for the 1-day scenario is 51.6%. In terms of intraday scenarios, the $R^2$ values for 30-min and 65-min horizons are slightly small, 48.6% and 48.1%, respectively. The results on 10-min data are somewhat surprising, where the $R^2$ reaches to 55.6%. One possible reason is that economic policy uncertainty is significant in the 10-min scenario. In another unreported robustness test, we estimate the regressions without the EPU factor. The

**Table A.1.** Time-series regression of commonality

| | 10-min | 30-min | 65-min | 1-day |
|---|---|---|---|---|
| VIX | 0.233* | 0.196* | 0.192* | 0.714* |
| | (0.030) | (0.024) | (0.023) | (0.084) |
| CSI | 0.214* | 0.097* | 0.066* | 0.237* |
| | (0.025) | (0.020) | (0.019) | (0.070) |
| EPU | 0.079* | 0.025 | 0.022 | 0.114 |
| | (0.029) | (0.023) | (0.022) | (0.080) |
| Constant | 0.161* | 0.982* | 1.267* | −0.689* |
| | (0.023) | (0.018) | (0.018) | (0.063) |
| Adjusted $R^2$ (%) | 55.6 | 48.6 | 48.1 | 51.6 |

*Notes:* The table reports the results of time-series regressions of average commonality in volatility $(R^2_{(h),m})_L$ over different horizons against three sentiment measures, VIX, CSI, and EPU. Superscript * denotes the significance levels of 5%. To compare the effects of various investor sentiments, we normalize each explanatory variable by removing its mean and scaling to the unit variance.

15 http://www.sca.isr.umich.edu (accessed on February 28, 2022).

16 https://www.policyuncertainty.com (accessed on February 28, 2022).

adjusted $R^2$ value in the regression of 10-min data declines 2.5% while for other regressions, the changes in adjusted $R^2$ are subtle.

Besides the market volatility (VIX), we also find a significant effect of consumer sentiment (CSI) on the commonality of volatility over every studied horizon. The level of commonality is higher in times of higher market volatility and consumer sentiments. In addition, we observe that the coefficients of VIX and CSI for commonality in intraday volatility (especially for 30-min and 65-min) are substantially smaller than those in the daily case.

## B: Hyperparameter Tuning

There is no hyperparameter to tune in HAR-D and OLS. For LASSO, we use the standard five-fold cross-validation method to determine $\lambda_1$. Hyperparameters for other models in the main analysis are summarized as follows.

To assess the robustness of NNs to different architectures, we repeat the main analysis using one, two, and three hidden layers.[17] The results reported in Table B.2 are generally consistent with those reported in Table 3.

## C: DM Test

DM test is used to discriminate the significant differences of forecasting accuracy between different time-series models (e.g., Diebold and Mariano 1995; Diebold 2015). Denote the loss associated with forecast error $e_t$ by $L(e_t)$, for example, $L(e_t) = e_t^2$. Then the loss difference between the forecasts of models $a$ and $b$ is given by $d_t^{(a-b)} = L(e_t^{(a)}) - L(e_t^{(b)})$, where $e_t^{(a)}$ ($e_t^{(b)}$) represents the forecast error from model $a$ ($b$), respectively. The DM test makes

**Table B.1**. Hyperparameters in XGBoost, MLP, LSTM

| | XGBoost | MLP | LSTM |
|---|---|---|---|
| Learning rate | 0.1 | 0.001 | 0.001 |
| Early stopping rounds | 10 | 10 | 10 |
| Ensemble | 2000 | 10 | 10 |
| Max depth | 10 | – | – |
| Batch size | – | 1024 | 1024 |
| Epoches | – | 100 | 100 |
| Number of hidden layers | – | 3 | 2 |
| Batch normalization | – | ✓ | ✗ |

17 The number of neurons is chosen based on the geometric pyramid rule, following Gu, Kelly, and Xiu (2020).

**Table B.2.** Out-of-sample performance of alternative hyperparameters in NNs

| | | 10-min | | 30-min | | 65-min | | 1-day | |
|---|---|---|---|---|---|---|---|---|---|
| | | MSE | QLIKE | MSE | QLIKE | MSE | QLIKE | MSE | QLIKE |
| MLP1 | Universal | 0.949 | 0.398 | 0.286 | 0.182 | 0.234 | 0.164 | 0.261 | 0.191 |
| | Augmented | 0.947 | 0.388 | 0.281 | 0.181 | 0.229 | 0.162 | 0.257 | 0.181 |
| MLP2 | Universal | 0.948 | 0.398 | 0.284 | 0.182 | 0.232 | 0.164 | 0.260 | 0.190 |
| | Augmented | 0.947 | 0.387 | 0.281 | 0.180 | 0.229 | 0.163 | 0.256 | 0.180 |
| MLP3 | Universal | 0.947 | 0.397 | 0.284 | 0.181 | 0.232 | 0.163 | 0.260 | 0.191 |
| | Augmented | 0.945 | 0.386 | 0.280 | 0.179 | 0.229 | 0.162 | 0.257 | 0.180 |
| LSTM1 | Universal | 0.956 | 0.398 | 0.293 | 0.190 | 0.232 | 0.163 | 0.262 | 0.189 |
| | Augmented | 0.938 | 0.383 | 0.286 | 0.181 | 0.230 | 0.161 | 0.259 | 0.182 |
| LSTM2 | Universal | 0.950 | 0.393 | 0.287 | 0.179 | 0.232 | 0.162 | 0.261 | 0.188 |
| | Augmented | 0.934 | 0.376 | 0.279 | 0.171 | 0.229 | 0.160 | 0.258 | 0.182 |
| LSTM3 | Universal | 0.949 | 0.392 | 0.286 | 0.178 | 0.232 | 0.163 | 0.260 | 0.187 |
| | Augmented | 0.933 | 0.376 | 0.280 | 0.171 | 0.229 | 0.161 | 0.256 | 0.181 |

*Notes:* MLP1 has Single hidden layer with 128 neurons. MLP2 has two hidden layers of 128 and 64 neurons, respectively. MLP3 has three hidden layers of 128, 64, and 32 neurons, respectively. LSTM variants have similar meanings.

one assumption that $d_t^{(a-b)}$ is covariance stationary. The null hypothesis is that $\mathbb{E}(d_t^{(a-b)}) = 0$. Under the covariance stationary assumption, we have the test statistic

$$\mathrm{DM}_{12} = \frac{\overline{d}^{(a-b)}}{\widehat{\sigma}^{(a-b)}} \to N(0,1), \tag{30}$$

where $\overline{d}^{(a-b)} = \frac{1}{T}\sum_{t=1}^{T} d_t^{(a-b)}$ is the sample mean of $d_t^{(a-b)}$ and $\widehat{\sigma}^{(a-b)}$ is a consistent estimate of the standard deviation of $\overline{d}^{(a-b)}$.

Following Gu, Kelly, and Xiu (2020), we apply a modified DM test, to make pairwise comparisons of models' performance when forecasting multi-asset volatility. Specifically, the modified DM test compares the cross-sectional average of prediction errors from each model, rather than comparing errors for each individual asset, that is,

$$d_t^{(a-b)} = \frac{1}{N}\sum_{i=1}^{N}\left(L(e_{i,t}^{(a)}) - L(e_{i,t}^{(b)})\right), \tag{31}$$

where $e_{i,t}^{(a)}$ ($e_{i,t}^{(b)}$) refers to the forecast error for stock $i$ at time $t$ from model $a$ ($b$), respectively.

To assess the statistical significance of the differences in out-of-sample volatility forecasts as shown in Table 3, we report the results of all DM tests in terms of QLIKE for each horizon.

**Table C.1.** Statistics of DM tests

| Panel A: 10-min. | | | | | | |
|---|---|---|---|---|---|---|
| Univ / Univ | LASSO | OLS | LASSO | XGBoost | MLP | LSTM |
| LASSO | | 42.33* | 36.17* | 56.55* | 83.26* | 72.43* |
| OLS | | | −32.84* | 33.30* | 62.29* | 52.90* |
| Lasso | | | | 35.00* | 62.15* | 54.17* |
| XGBoost | | | | | 25.86* | 20.31* |
| MLP | | | | | | −3.39* |
| LSTM | | | | | | |
| Single vs | −30.07* | −1.02 | 31.27* | 59.38* | | |
| Aug / Aug | LASSO | OLS | LASSO | XGBoost | MLP | LSTM |
| LASSO | | 56.60* | 58.95* | 33.75* | 68.81* | 66.94* |
| OLS | | | 5.69* | −6.02* | 31.82* | 31.71* |
| Lasso | | | | −6.52* | 30.99* | 31.43* |
| XGBoost | | | | | 21.54* | 28.72* |
| MLP | | | | | | 14.51* |
| LSTM | | | | | | |
| Univ vs | 46.23* | 51.28* | 53.53* | −0.32 | 6.24* | 29.63* |
| Panel B: 30-min. | | | | | | |
| Univ / Univ | HAR-D | OLS | LASSO | XGBoost | MLP | LSTM |
| HAR-D | | 44.74* | 44.00* | 42.96* | 52.99* | 46.17* |
| OLS | | | −23.57* | 22.63* | 35.25* | 26.19* |
| LASSO | | | | 24.55* | 37.07* | 28.04* |
| XGBoost | | | | | 16.46* | 8.35* |
| MLP | | | | | | −8.72* |
| LSTM | | | | | | |
| Single vs | −7.47* | −0.48 | 23.14* | 48.57* | | |
| Aug / Aug | HAR-D | OLS | LASSO | XGBoost | MLP | LSTM |
| HAR-D | | 36.36* | 38.00* | 24.16* | 45.12* | 44.43* |
| OLS | | | −2.11* | −4.50* | 20.55* | 18.26* |
| LASSO | | | | −4.34* | 21.21* | 19.05* |
| XGBoost | | | | | 21.04* | 23.02* |
| MLP | | | | | | 2.24* |
| LSTM | | | | | | |
| Univ vs | 17.70* | 22.51* | 25.24* | −9.59* | 9.56* | 23.23* |

(continued)

Panel C: 65-min.

| Univ \ Univ | HAR-D | OLS | LASSO | XGBoost | MLP | LSTM |
|---|---|---|---|---|---|---|
| HAR-D | | 28.27* | 27.75* | 21.56* | 30.06* | 29.00* |
| OLS | | | −11.73* | 7.78* | 20.22* | 18.67* |
| LASSO | | | | 8.81* | 20.91* | 19.35* |
| XGBoost | | | | | 19.83* | 18.17* |
| MLP | | | | | | 0.68* |
| LSTM | | | | | | |
| Single vs | −1.87 | 8.17* | 8.53* | 41.26* | | |
| Aug \ Aug | HAR-D | OLS | LASSO | XGBoost | MLP | LSTM. |
| HAR-D | | 22.11* | 22.67* | 9.87* | 25.92* | 26.01* |
| OLS | | | −4.89* | −5.56* | 12.84* | 11.79* |
| LASSO | | | | −5.11* | 13.72* | 12.67* |
| XGBoost | | | | | 17.71* | 18.54* |
| MLP | | | | | | 0.94* |
| LSTM | | | | | | |
| Univ vs | 10.92* | 12.47* | 13.35* | −7.52* | 7.15* | 7.94* |

Panel D: 1-day

| Univ \ Univ | HAR-D | OLS | LASSO | XGBoost | MLP | LSTM |
|---|---|---|---|---|---|---|
| HAR-D | | 0.50 | −0.12 | −4.29* | 2.69* | 3.86* |
| OLS | | | −4.94* | −5.50* | 1.91 | 3.36* |
| LASSO | | | | −4.29* | 2.76* | 3.88* |
| XGBoost | | | | | 9.49* | 10.90* |
| MLP | | | | | | 3.03* |
| LSTM | | | | | | |
| Single vs | −1.32 | 3.41* | 5.42* | 20.53* | | |
| Aug \ Aug | HAR-D | OLS | LASSO | XGBoost | MLP | LSTM. |
| HAR-D | | −0.12 | 5.39* | −8.97* | 3.20* | 1.87 |
| OLS | | | −1.10 | −12.63* | −3.77* | −3.99* |
| LASSO | | | | −12.68* | −3.34* | −3.91* |
| XGBoost | | | | | 12.30* | 11.61* |
| MLP | | | | | | −2.20* |
| LSTM | | | | | | |
| Univ vs | 4.50* | 5.81* | 6.30* | −5.01* | 4.63* | 2.78* |

*Notes:* In each panel, the left sub-table represents the pairwise comparison of forecasting performance of six models trained under Universal and the right one represents the pairwise comparison of forecasting performance of six models trained under Augmented. The bottom row in each sub-table represents the comparison of forecasting performance of the same model under two different training schemes. Positive numbers indicate the column model outperforms the row model. Superscript * denotes the significance levels of 5%.

**Table D.1.** Frequency of updating HAR-D for predicting intraday RVs

| Panel A: | | 10-min | | 30-min | | 65-min | | 1-day | |
|---|---|---|---|---|---|---|---|---|---|
| Statistical performance | | MSE | QLIKE | MSE | QLIKE | MSE | QLIKE | MSE | QLIKE |
| Weekly | Single | 1.013 | 0.483 | 0.332 | 0.221 | 0.270 | 0.190 | 0.267 | 0.188 |
| | Universal | 1.021 | 0.517 | 0.333 | 0.230 | 0.270 | 0.190 | 0.268 | 0.189 |
| | Augmented | 0.995 | 0.453 | 0.323 | 0.228 | 0.262 | 0.185 | 0.256 | 0.180 |
| Monthly | Single | 1.013 | 0.483 | 0.332 | 0.222 | 0.270 | 0.190 | 0.267 | 0.189 |
| | Universal | 1.021 | 0.517 | 0.333 | 0.230 | 0.270 | 0.191 | 0.268 | 0.190 |
| | Augmented | 0.995 | 0.453 | 0.323 | 0.227 | 0.262 | 0.185 | 0.256 | 0.180 |
| Yearly | Single | 1.013 | 0.484 | 0.332 | 0.222 | 0.270 | 0.190 | 0.269 | 0.190 |
| | Universal | 1.021 | 0.518 | 0.333 | 0.230 | 0.270 | 0.191 | 0.269 | 0.190 |
| | Augmented | 0.995 | 0.453 | 0.323 | 0.227 | 0.262 | 0.186 | 0.257 | 0.180 |
| Panel B: | | 10-min | | 30-min | | 65-min | | 1-day | |
| Realized utility | | RU | RU-TC | RU | RU-TC | RU | RU-TC | RU | RU-TC |
| Weekly | Single | 2.694 | 2.069 | 3.459 | 3.042 | 3.543 | 3.096 | 3.551 | 3.518 |
| | Universal | 2.575 | 1.972 | 3.427 | 3.014 | 3.541 | 3.095 | 3.548 | 3.516 |
| | Augmented | 2.790 | 2.280 | 3.427 | 3.020 | 3.553 | 3.108 | 3.571 | 3.536 |
| Monthly | Single | 2.693 | 2.068 | 3.458 | 3.042 | 3.542 | 3.096 | 3.549 | 3.516 |
| | Universal | 2.574 | 1.972 | 3.427 | 3.014 | 3.541 | 3.095 | 3.547 | 3.514 |
| | Augmented | 2.789 | 2.279 | 3.426 | 3.020 | 3.553 | 3.107 | 3.571 | 3.536 |
| Yearly | Single | 2.690 | 2.065 | 3.457 | 3.040 | 3.542 | 3.095 | 3.548 | 3.515 |
| | Universal | 2.574 | 1.975 | 3.429 | 3.016 | 3.541 | 3.095 | 3.547 | 3.514 |
| | Augmented | 2.790 | 2.280 | 3.428 | 3.022 | 3.552 | 3.107 | 3.571 | 3.536 |

## D: Model Update Frequency

In this article, we choose to update each risk model annually due to the limited computation resources. To understand whether the model's performance might change with respect to the update frequency, we update the HAR-D model with different frequencies, that is, weekly, monthly, and yearly, and results are summarized as follows. The conclusions are generally consistent with those from our main analysis.

## References

Andersen, Torben G., and Tim Bollerslev. 1997. Intraday Periodicity and Volatility Persistence in Financial Markets. *Journal of Empirical Finance* 4: 115–158.

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Heiko Ebens. 2001. The Distribution of Realized Stock Return Volatility. *Journal of Financial Economics* 61: 43–76.

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Paul Labys. 2003. Modeling and Forecasting Realized Volatility. *Econometrica* 71: 579–625.

Andersen, Torben G., Tim Bollerslev, Peter F. Christoffersen, and Francis X. Diebold. 2006. "Volatility and Correlation Forecasting." In G. Elliott, C. Granger, and A. Timmermann (eds.), *Handbook of Economic Forecasting*. Elsevier, 1 edition, Vol. 1, pp. 777–878.

Baker, Malcolm, and Jeffrey Wurgler. 2007. Investor Sentiment in the Stock Market. *Journal of Economic Perspectives* 21: 129–151.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis. 2016. Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics* 131: 1593–1636.

Barndorff-Nielsen, Ole E., and Neil Shephard. 2002. Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64: 253–280.

Barndorff-Nielsen, Ole E., Silja Kinnebrock, and Neil Shephard. 2008. "Measuring Downside Risk-Realised Semivariance." CREATES research paper (2008-42).

Bates, David S. 2019. How Crashes Develop: Intradaily Volatility and Crash Evolution. *The Journal of Finance* 74: 193–238.

Bollen, Bernard, and Brett Inder. 2002. Estimating Daily Volatility in Financial Markets Utilizing Intraday Data. *Journal of Empirical Finance* 9: 551–562.

Bollerslev, Tim. 2022. Realized Semi (Co)Variation: Signs That All Volatilities Are Not Created Equal. *Journal of Financial Econometrics* 20: 219–252.

Bollerslev, Tim, Andrew J. Patton, and Rogier Quaedvlieg. 2016. Exploiting the Errors: A Simple Approach for Improved Volatility Forecasting. *Journal of Econometrics* 192: 1–18.

Bollerslev, Tim, Benjamin Hood, John Huss, and Lasse Heje Pedersen. 2018. Risk Everywhere: Modeling and Managing Volatility. *The Review of Financial Studies* 31: 2729–2773.

Bucci, Andrea. 2020. Realized Volatility Forecasting with Neural Networks. *Journal of Financial Econometrics* 18: 502–531.

Calvet, Laurent E., Adlai J. Fisher, and Samuel B. Thompson. 2006. Volatility Comovement: A Multifrequency Approach. *Journal of Econometrics* 131: 179–215.

Carroll, Christopher D., Jeffrey C. Fuhrer, and David W. Wilcox. 1994. Does Consumer Sentiment Forecast Household Spending? If So, Why? *American Economic Review* 84: 1397–1408.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'16, ACM, pp. 785–794.

Choi, Darwin, Wenxi Jiang, and Chao Zhang. 2021. "Alpha Go Everywhere: Machine Learning and International Stock Returns," Working paper.

Chordia, Tarun, Richard Roll, and Avanidhar Subrahmanyam. 2000. Commonality in Liquidity. *Journal of Financial Economics* 56: 3–28.

Christensen, Bent J., and Nagpurnanand R. Prabhala. 1998. The Relation between Implied and Realized Volatility. *Journal of Financial Economics* 50: 125–150.

Christensen, Kim, Mathias Siggaard, and Bezirgen Veliyev. 2021. "A Machine Learning Approach to Volatility Forecasting." Working paper.

Corsi, Fulvio. 2009. A Simple Approximate Long-Memory Model of Realized Volatility. *Journal of Financial Econometrics* 7: 174–196.

Da, Zhi, Joseph Engelberg, and Pengjie Gao. 2011. In Search of Attention. *The Journal of Finance* 66: 1461–1499.

Da, Zhi, Joseph Engelberg, and Pengjie Gao. 2015. The Sum of All FEARS Investor Sentiment and Asset Prices. *Review of Financial Studies* 28: 1–32.

Dang, Tung Lam, Fariborz Moshirian, and Bohui Zhang. 2015. Commonality in News around the World. *Journal of Financial Economics* 116: 82–110.

De Long, J. Bradford, Andrei Shleifer, Lawrence H. Summers, and Robert J. Waldmann. 1990. Noise Trader Risk in Financial Markets. *Journal of Political Economy* 98: 703–738.

Diebold, Francis X. 2015. Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests. *Journal of Business & Economic Statistics* 33: 1–1.

Diebold, Francis X., and Roberto S. Mariano. 1995. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics* 13: 253–263.

Engle, Robert F. 1982. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* 50: 987–1007.

Engle, Robert F., and Andrew J. Patton. 2007. "What Good Is a Volatility Model?" In J. Knight and S. Satchell (eds.), *Forecasting Volatility in the Financial Markets*. Elsevier, pp. 47–63.

Engle, Robert F., and Magdalena E. Sokalska. 2012. Forecasting Intraday Volatility in the US Equity Market: Multiplicative Component GARCH. *Journal of Financial Econometrics* 10: 54–83.

Friedman, Jerome H. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29: 1189–1232.

Ghysels, Eric, Pedro Santa-Clara, and Rossen Valkanov. 2005. There Is a Risk-Return Trade-Off after All. *Journal of Financial Economics* 76: 509–548.

Ghysels, Eric, Pedro Santa-Clara, and Rossen Valkanov. 2006. Predicting Volatility: Getting the Most Out of Return Data Sampled at Different Frequencies. *Journal of Econometrics* 131: 59–95.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu. 2020. Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies* 33: 2223–2273.

Hameed, Allaudeen, Wenjin Kang, and Shivesh Viswanathan. 2010. Stock Market Declines and Liquidity. *The Journal of Finance* 65: 257–293.

Hansen, Lars Kai, and Peter Salamon. 1990. Neural Network Ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12: 993–1001.

Hansen, Peter R., and Asger Lunde. 2005. A Forecast Comparison of Volatility Models: Does anything Beat a GARCH(1,1)? *Journal of Applied Econometrics* 20: 873–889.

Hansen, Peter R., and Asger Lunde. 2006. Realized Variance and Market Microstructure Noise. *Journal of Business & Economic Statistics* 24: 127–161.

Hansen, Peter R., Asger Lunde, and James M. Nason. 2011. The Model Confidence Set. *Econometrica* 79: 453–497.

Harris, Lawrence. 1986. A Transaction Data Study of Weekly and Intradaily Patterns in Stock Returns. *Journal of Financial Economics* 16: 99–117.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media,

Herskovic, Bernard, Bryan Kelly, Hanno Lustig, and Stijn Van Nieuwerburgh. 2016. The Common Factor in Idiosyncratic Volatility: Quantitative Asset Pricing Implications. *Journal of Financial Economics* 119: 249–283.

Herskovic, Bernard, Bryan Kelly, Hanno Lustig, and Stijn Van Nieuwerburgh. 2020. Firm Volatility in Granular Networks. *Journal of Political Economy* 128: 4097–4162.

Hill, Tim, Marcus O'Connor, and William Remus. 1996. Neural Network Models for Time Series Forecasts. *Management Science* 42: 1082–1092.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9: 1735–1780.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer Feedforward Networks Are Universal Approximators. *Neural Networks* 2: 359–366.

Izzeldin, Marwan, M. Kabir Hassan, Vasileios Pappas, and Mike Tsionas. 2019. Forecasting Realised Volatility Using ARFIMA and HAR Models. *Quantitative Finance* 19: 1627–1638.

Karolyi, G. Andrew, Kuan-Hui Lee, and Mathijs A. Van Dijk. 2012. Understanding Commonality in Liquidity around the World. *Journal of Financial Economics* 105: 82–112.

Keynes, John Maynard. 2018. *The General Theory of Employment, Interest, and Money*. Springer.

Kingma, Diederik P., and Jimmy Ba. 2014. "Adam: A Method for Stochastic Optimization." Working paper.

Kogan, Leonid, Stephen A. Ross, Jiang Wang, and Mark M. Westerfield. 2006. The Price Impact and Survival of Irrational Traders. *The Journal of Finance* 61: 195–229.

Lemmon, Michael, and Evgenia Portniaguina. 2006. Consumer Confidence and Asset Prices: Some Empirical Evidence. *Review of Financial Studies* 19: 1499–1529.

Li, Sophia Zhengzi, and Yushan Tang. 2020. "Forecasting Realized Volatility: An Automatic System Using Many Features and Many Machine Learning Algorithms." Working paper.

Liu, Lily Y., Andrew J. Patton, and Kevin Sheppard. 2015. Does Anything Beat 5-Minute RV? A Comparison of Realized Measures across Multiple Asset Classes. *Journal of Econometrics* 187: 293–311.

Morck, Randall, Bernard Yeung, and Wayne Yu. 2000. The Information Content of Stock Markets: Why Do Emerging Markets Have Synchronous Stock Price Movements? *Journal of Financial Economics* 58: 215–260.

Ni, Sophie X., Jun Pan, and Allen M. Poteshman. 2008. Volatility Information Trading in the Option Market. *The Journal of Finance* 63: 1059–1091.

Pascalau, Razvan, and Ryan Poirier. 2021. Increasing the Information Content of Realized Volatility Forecasts. *Journal of Financial Econometrics*, 1–35. https://academic.oup.com/jfec/advance-article/doi/10.1093/jjfinec/nbab028/6459606?searchresult=1

Patton, Andrew J. 2011. Volatility Forecast Comparison Using Imperfect Volatility Proxies. *Journal of Econometrics* 160: 246–256.

Patton, Andrew J., and Kevin Sheppard. 2009. "Evaluating Volatility and Correlation Forecasts." In T. Mikosch, J.-P. Kreiß, R. A. Davis, and T. G. Andersen (eds.), *Handbook of Financial Time Series*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 801–838.

Patton, Andrew J. and Kevin Sheppard. 2015. Good Volatility, Bad Volatility: Signed Jumps and the Persistence of Volatility. *Review of Economics and Statistics* 97: 683–697.

Rahimikia, Eghbal, and Ser-Huang Poon. 2020. "Machine Learning for Realised Volatility Forecasting." Working paper.

Ribeiro, Gabriel Trierweiler, André Alves Portela Santos, Viviana Cocco Mariani, and Leandro dos Santos Coelho. 2021. Novel Hybrid Model Based on Echo State Neural Network Applied to the Prediction of Stock Price Return Volatility. *Expert Systems with Applications* 184: 115490.

Sadhwani, Apaar, Kay Giesecke, and Justin Sirignano. 2021. Deep Learning for Mortgage Risk. *Journal of Financial Econometrics* 19: 313–368.

Sheppard, Kevin. 2010. *Financial Econometrics Notes*, pp. 333–426. University of Oxford.

Shleifer, Andrei, and Lawrence H. Summers. 1990. The Noise Trader Approach to Finance. *Journal of Economic Perspectives* 4: 19–33.

Sirignano, Justin, and Rama Cont. 2019. Universal Features of Price Formation in Financial Markets: Perspectives from Deep Learning. *Quantitative Finance* 19: 1449–1459.

Stroud, Jonathan R., and Michael S. Johannes. 2014. Bayesian Modeling and Forecasting of 24-Hour High-Frequency Volatility. *Journal of the American Statistical Association* 109: 1368–1384.

Taylor, Stephen J., and Xinzhong Xu. 1997. The Incremental Volatility Information in One Million Foreign Exchange Quotations. *Journal of Empirical Finance* 4: 317–340.

Xiong, Ruoxuan, Eric P. Nichols, and Yuan Shen. 2015. "Deep Learning Stock Volatility with Google Domestic Trends." Working paper.

Zhang, G. Peter. 2003. Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model. *Neurocomputing* 50: 159–175.