

# Volatility forecasting with machine learning and intraday commonality

Chao Zhang<sup>\*1</sup>, Yihuang Zhang<sup>\*2,4</sup>, Mihai Cucuringu<sup>1,2,3</sup>, and Zhongmin Qian<sup>1,4</sup>

<sup>1</sup>Department of Statistics, University of Oxford, Oxford, UK

<sup>2</sup>Mathematical Institute, University of Oxford, Oxford, UK

<sup>3</sup>The Alan Turing Institute, London, UK

<sup>4</sup>Oxford Suzhou Centre for Advanced Research

May 2022

## 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 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 one-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.

**Keywords:** Intraday volatility forecasting, Neural networks, Realized volatility, Commonality

**JEL Codes:** C45, C53, G17

---

<sup>\*</sup>Equal contribution. Correspondence to: Chao Zhang <chao.zhang@stats.ox.ac.uk>. The authors thank the Oxford Suzhou Centre for Advanced Research for providing the computational power. First draft: February 2022

# 1 Introduction

Forecasting and modeling stock return volatility has been of interest to both academics and practitioners over the past years. 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 the ten-second, one-minute and ten-minute volatility forecasts for the futures associated to DAX, EURO STOXX 50 and Euro-Bund.

As pointed out in Engle and Sokalska [30], intraday volatility forecasts are important for managing risk, pricing derivatives, and devising quantitative strategies, especially in high-frequency trading. Specifically, intraday volatility forecasts can support traders in assessing the likelihood of price changes, therefore better understanding the risk involved in certain automated trading strategies. On the other hand, screen traders can use these forecasts to support live trading and enhance the pre-trade transaction cost analysis from the risk assessment of price slippage. Moreover, intraday volatility forecasts are also helpful for practitioners screening for high-volatility opportunities and trading respective option strategies.

To the best of our knowledge, unlike daily volatility forecasting models, forecasting intraday volatility has received scant attention in the research literature. It is pointed out by Andersen and Bollerslev [1] that conventional parametric models, such as GARCH and stochastic volatility models, may be inadequate for modeling intraday returns. In recent works, such as Andersen et al. [4] and Corsi [21], high-frequency data are used to estimate daily realized volatility (RV) by summing squared intraday returns. However, these methods are potentially restrictive and are often difficult to apply when forecasting intraday volatility. Some related literature will be reviewed briefly in the next section.

In the present paper, we study and analyze various non-parametric machine learning models for forecasting *multi-asset intraday volatilities* by using high-frequency data from the U.S. equity market. We demonstrate that, by taking advantage of *commonality* in intraday volatility, the model's forecasting performance can significantly be improved. We also reveal the intraday volatility can afford additional useful information for forecasting daily volatility. The data we explore spans the period from July 2011 to June 2021, and includes the top 100 most liquid components of S&P 500 index. In our approach, the 10-min, 30-min, 65-min, and daily (without overnight information) forecasting horizons are studied.

The main contributions of our work can be summarized as follows. First, a measure for evaluating the commonality in intraday volatility is proposed, that is the adjusted R-squared value from linear regressions of a given stock's RVs against the market RVs. It is demonstrated that commonality over the daily horizon is turbulent over time, although commonality in intraday RVs is strong and stable. On the other hand,

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 the closing session, in contrast to the diurnal volatility pattern.

Second, in order to assess the benefits of incorporating commonality into models aimed at predicting intraday volatility, we train multiple machine learning algorithms (including HAR, OLS, LASSO, XGBoost, MLP, and LSTM) under three different schemes: (a) **SINGLE**: training specific models for each asset; (b) **UNIVERSAL**: training one model with pooled data for all assets; (c) **AUGMENTED**: training one model with pooled data and adding an additional predictor which takes account the impact of market realized volatility. We find that for most models, the incorporation of intraday commonality leads to better out-of-sample performance through pooling data together and adding market volatility as additional features.

In addition, the empirical results we present in the paper demonstrate that neural networks (NNs) are, in general, superior to other techniques. New empirical evidence is provided, in order to demonstrate the capability of NNs for handling complex interactions among predictors. Furthermore, to alleviate the concerns of overfitting, we conduct a stringent out-of-sample test, using the existent trained models to forecast the volatility of completely new stocks that are not included in the training sample. Our results reveal that NNs still outperform other approaches (including the OLS models trained for each new stock), thus presenting empirical evidence for a universal volatility mechanism among stocks, similar to the findings in Sirignano and Cont [61] concerning universal features of price formation in equity markets.

We conclude the paper 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 contributes to the improvement over traditional methods of modeling daily volatilities. In other words, the results presented in this paper demonstrate that machine learning models, where past intraday volatilities are used as predictors, generally outperform the traditional models with past daily volatilities (e.g. HAR of Corsi [21], SHAR of Patton and Sheppard [57], HARQ of Bollerslev et al. [13]). *To the best of our knowledge, this is the first line of work that studies the effectiveness of past intraday volatilities in forecasting future daily volatility.*

The remainder of this paper is structured as follows. We begin in Section 2 by reviewing the related literature. Section 3 describes the data and the definition of realized volatility. Section 4 discusses the commonality in intraday volatility. Section 5 introduces various machine learning models and three training schemes for predicting future intraday volatility. Section 6 provides the forecasting results and discusses the empirical findings. In Section 7, we introduce a new approach to forecasting daily volatility using past intraday volatility as predictors. Finally, we conclude our analysis in Section 8.

## 2 Related literature

Our study is built on several research streams by various authors over the recent years. The first stream is related to the research on the commonality in financial markets. Chordia et al. [19] have recognized the existence of commonality in liquidity, and Karolyi et al. [46] have suggested that commonality in liquidity is related to market volatility, in particular, the presence of international investors and trading activity. Dang et al. [24] 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 stochastic volatility models (e.g. Andersen et al. [2], Calvet et al. [15]) all make use of the volatility spillover effects. Herskovic et al. [40] have provided empirical evidence of the co-movement in volatility across the equity market. Bollerslev et al. [12] have observed strong similarities in daily realized volatility and have utilized them to forecast the daily realized volatility. Engle and Sokalska [30] have emphasised that pooled data is useful for intraday volatility forecasting, and Herskovic et al. [41] 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 made by many researchers 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 stochastic volatility models, which provide forecasts of daily volatility from daily return. As pointed out by Andersen et al. [4, 2], Engle and Patton [29], 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, realized volatility (RV), computed from summing squared intraday returns, has gained popularity in recent years. Andersen et al. [4] have proposed an ARFIMA model for forecasting daily RVs, which outperforms conventional GARCH and related approaches. Corsi [21] has put forward a parsimonious AR-type model, termed Heterogeneous Autoregressive (HAR), for predicting daily RVs using various realized volatility components over different time horizons. Recently Izzeldin et al. [45] have made a comparison investigation for the forecasting performance of ARFIMA and HAR, and have concluded their performance is essentially indistinguishable. See Section 7 for more models to predict daily volatility.

On the other hand, little attention has been paid to the role of forecasting intraday volatility. Taylor and Xu [62] proposed an hourly volatility model based on an ARCH specification, and Engle and Sokalska [30] 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 stochastic

volatility, 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, machine learning (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 et al. [33] have pointed out the superior performance of ML models for empirical asset pricing. Sirignano and Cont [61] have used LSTMs to forecast high-frequency price movements, and have provided empirical evidence for the existence of a universal and stationary price formation mechanism.

Recently, Xiong et al. [63] have applied LSTMs to forecast S&P 500 volatility, with Google domestic trends as predictors, and Bucci [14] has demonstrated that RNNs are able to outperform all the traditional econometric methods in forecasting monthly volatility of the S&P index. More recently, Rahimikia and Poon [58] have compared machine learning models with HAR models for forecasting daily realized volatility by using variables extracted from limit order books and news. Li and Tang [51] have proposed a simple average ensemble model combining multiple machine learning algorithms for forecasting daily (and monthly) realized volatility, and Christensen et al. [20] have examined the performance of machine learning models in forecasting one-day-ahead realized volatility with firm-specific characteristics and macroeconomic indicators.

The main goal of the present paper is to assess the usefulness of non-parametric ML models through the lens of forecasting multi-asset intraday volatilities.

## 3 Data and realized volatility

### 3.1 Data

We use the Nasdaq ITCH data from LOBSTER<sup>1</sup> to compute intraday returns via mid-prices. We select the top 100 components of S&P 500 index, for the period between 2011-07-01 and 2021-06-30. 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 GICS sector division<sup>2</sup>.

---

<sup>1</sup><https://lobsterdata.com/>

<sup>2</sup>The Global Industry Classification Standard (GICS) is an industry taxonomy developed in 1999 by MSCI and Standard & Poor's (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

### 3.2 Realized volatility

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

$$d \log P_{i,t} = \mu_i dt + \sigma_{i,t} dW_t, \quad (1)$$

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

$$IV_{i,t}(h) = \int_{t-h}^t \sigma_{i,s}^2 ds, \quad (2)$$

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

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

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

Here,  $P_{i,t}$  is the mid-price at time  $t$ , i.e.  $P_{i,t} = \frac{P_{i,t}^b + P_{i,t}^s}{2}$ , and  $P_{i,t}^b$  (respectively,  $P_{i,t}^a$ ) represents the best bid (respectively, ask) price.

Andersen et al. [3], Barndorff-Nielsen and Shephard [9] 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 realized volatility as a proxy for the unobserved IV (see Bollen and Inder [10], Hansen and Lunde [37], Andersen et al. [3]). To reduce the impact of extreme values, we consider the logarithm, in line with Andersen et al. [4], Bucci [14], Herskovic et al. [40]. Specifically, during a period  $(t-h, t]$ , the

realized volatility is defined as follows<sup>3</sup>

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

As pointed out by Pascalau and Poirier [54], there are no conclusive methods to incorporate the overnight session’s information content into the daily volatility. In line with Engle and Sokalska [30], 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 paper.

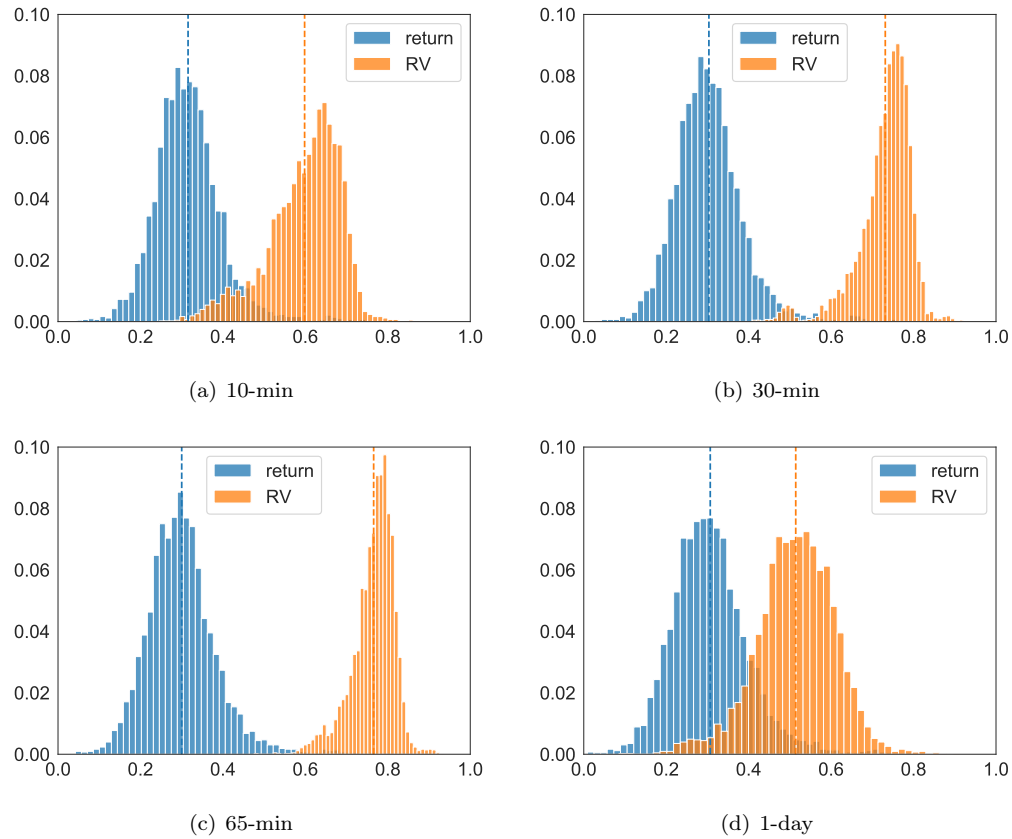
### 3.3 Summary statistics

To mitigate the effect of possibly spurious data errors, for each stock, we set the data of return/volatility below the 0.5% percentile set to the 0.5% percentile, and data above the 99.5% percentile set 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 observe generally higher correlations in realized volatility than the counterparts in return. Figure 1 also reveals that, on average, as the horizon gets longer, realized volatility’s correlations increase from 0.598 (10-min) to 0.731 (30-min) to 0.766 (65-min). However, when turning to daily realized volatility, 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 realized volatility over time. Stocks demonstrate similar time series patterns, consistent with Herskovic et al. [40], Bollerslev et al. [12]. Additionally, the width shrinks during the periods of higher volatility, such as, stock market crashes in August 2011 (European sovereign debt crisis), between June 2015 to June 2016 (Chinese stock market turbulence and Brexit), in March 2018 (China–United States trade war), in March 2020 (COVID-19). Figure 3 shows that the diurnal volatility forms a so-called reverse-J-shape, namely larger fluctuations near the open and close (see Harris [38], Engle and Sokalska [30]).

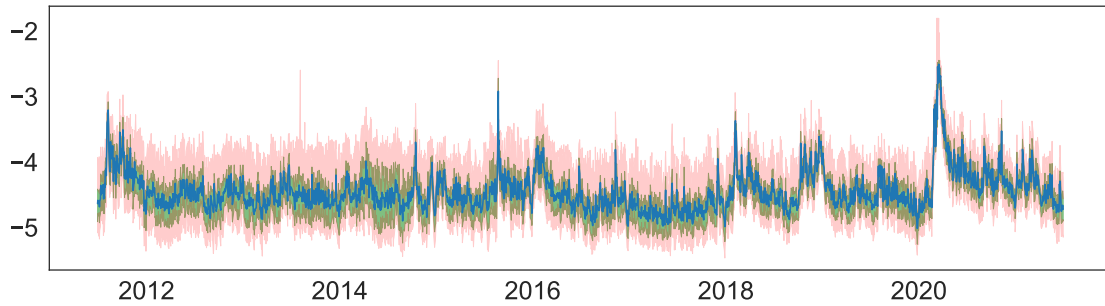
---

<sup>3</sup>Liu et al. [52] 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 the present paper, 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.

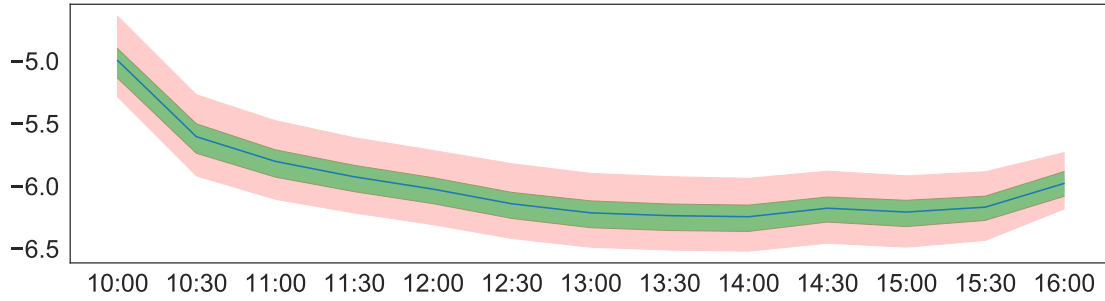
*Notes:* (a)-(d) are based on observations in the frequency of 10-min, 30-min, 65-min, 1-day, respectively. The dashed vertical lines represent the average correlation values of RVs and returns.



**Figure 2:** Daily realized volatility (in logs).

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





**Figure 3:** Diurnal realized volatility (in logs).

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

## 4 Commonality estimation

Inspired by prior studies (e.g., Morck et al. [53], Chordia et al. [19], Karolyi et al. [46], Dang et al. [24]), we follow an analogous procedure to estimate the commonality in volatility. Specifically, we use the average adjusted R-squared value from the following regressions across stocks, as a measure of commonality in volatility (denoted as  $R_{(h)}^2$ )<sup>4</sup>

$$RV_{i,t}^{(h)} = \alpha_i + \beta_i RV_{M,t}^{(h)} + \epsilon_{i,t}, \quad (5)$$

where  $RV_{M,t}^{(h)}$  (see Bollerslev et al. [12]) is the contemporaneous market volatility during  $(t - h, t]$  for stock  $i$ , which is calculated as the equally weighted average<sup>5</sup> of all individual stock volatilities during  $(t - h, t]$ , i.e.

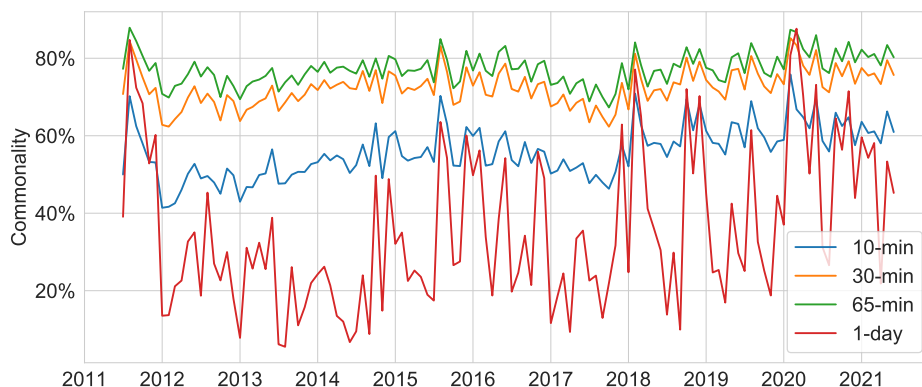
$$RV_{M,t}^{(h)} = \frac{1}{N} \sum_{i=1}^N RV_{i,t}^{(h)}. \quad (6)$$

Figure 4 presents the commonality in realized volatility, averaged across stocks for each month. To create this figure, we use the observations in each month, to obtain the R-squared value from Eqn (5). We notice that commonality effects in intraday scenarios (especially 30-min, 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,  $R_{(h)}^2$  is much more turbulent at the daily frequency. The last column in Table 2 also reports the results of the relation between the average commonality and the

<sup>4</sup>We also perform another regression, where except for contemporaneous market volatility, the lag one (thus  $t - 1$  in (5)) and lead one (thus  $t + 1$  in (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 et al. [19], Karolyi et al. [46], Dang et al. [24]. The R-squared values are similar to the ones of Eqn (5).

<sup>5</sup>We also implemented the value weighted market volatility and the results are similar to the equally weighted market volatility.

market volatility. As the horizon extends, the average commonality has a higher correlation with the market volatility<sup>6</sup>.



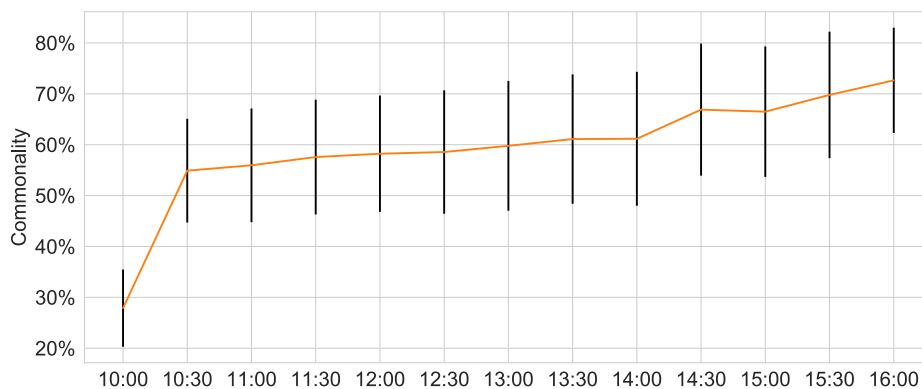
**Figure 4:** Commonality in realized volatility.

*Notes:* The commonality is averaged across stocks for each month during the sample period of 2011-07 ~ 2021-06.

**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

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



**Figure 5:** Commonality in realized volatility.

*Notes:* The commonality is averaged across stocks for each half-hour during the sample period of 2011-07 ~ 2021-06.

Figure 5 reports the averaged values and standard deviations (black vertical lines) of commonality for

<sup>6</sup>We refer the reader to additional analysis on commonality in Appendix A.

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 Eqn (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.

## 5 Methodology

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

$$\begin{aligned} \text{RV}_{i,t+h}^{(h)} &= F_i(\mathbf{u}; \theta) + \epsilon_{i,t+h} \\ &= F_i(\mathbf{x}_{i,:t}^{(h)}, \mathbf{z}_{:,t}^{(h)}; \theta) + \epsilon_{i,t+h} \end{aligned} \quad (7)$$

where  $\text{RV}_{i,t+h}^{(h)}$  is the volatility of asset  $i$  during  $(t, t+h]$ .  $\mathbf{u}$  represents the input features, which can be further separated into 2 categories: (1)  $\mathbf{x}_{i,:t}^{(h)}$  is a multi-dimensional vector of predictor variables for a specific stock  $i$  available up to time  $t$ , denoted as *individual features*; (2)  $\mathbf{z}_{:,t}^{(h)}$  is a vector of features for all stocks in the studied universe up to  $t$ , denoted as *market features*, such as market volatility.  $\theta$  refers to the parameters that need to be estimated. We are aiming to find a function of variables which minimizes the out-of-sample errors for future realized volatility.

### 5.1 Models

This section summarizes the six models employed in our numerical experiments.

#### 5.1.1 Heterogeneous autoregressive with diurnal effects (HAR-d)

Corsi [21] proposed a volatility model, named as Heterogeneous autoregressive (HAR), which considers realized volatilities over different interval sizes. HAR has shown remarkably good forecasting performance on daily data Patton and Sheppard [57], Izzeldin et al. [45]. For day  $t$ , the forecast of HAR is based on

$$\text{RV}_{i,t+1}^{(d)} = \alpha_i + \beta_i^{(d)} \text{RV}_{i,t}^{(d)} + \beta_i^{(w)} \text{RV}_{i,t}^{(w)} + \beta_i^{(m)} \text{RV}_{i,t}^{(m)} + \epsilon_{i,t+1}, \quad (8)$$

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

However, very little attention has been paid to forecasting intraday volatility with HAR. One closely

connected model is that of Engle and Sokalska [30], 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 [28] to learn the stochastic intraday volatility components.

Following the spirit of Engle and Sokalska [30], we extend the daily HAR model to intraday scenarios by adding diurnal-effect features, as follows<sup>7</sup>

$$RV_{i,t+h}^{(h)} = \alpha_i + \beta_i^{(\tau)} D_{i,\tau_{t+h}} + \beta_i^{(d)} RV_{i,t}^{(d)} + \beta_i^{(w)} RV_{i,t}^{(w)} + \beta_i^{(m)} RV_{i,t}^{(m)} + \epsilon_{i,t+h}, \quad (9)$$

where  $D_{i,\tau_{t+h}}$  denote the average diurnal realized volatility in the bucket-of-the-day  $\tau_{t+h}$  computed from the last 21 days. For example, when  $t = 10:30$  and  $h = 30$  minutes, then  $\tau_{t+h}$  corresponds to the bucket 10:30-11:00.  $RV_{i,t}^{(d)}$  ( $RV_{i,t}^{(w)}$ ,  $RV_{i,t}^{(m)}$ ) denotes the aggregated daily (weekly, monthly) realized volatility.

When we consider the daily scenarios, Eqn (9) becomes the standard HAR model (Eqn (8)), by removing the diurnal term. For simplicity, we denote this model as HAR-d in the following experiments.

### 5.1.2 Ordinary least squares (OLS)

Instead of using aggregated realized volatility, we apply OLS to original features, as follows, with its loss function being the sum of squared errors. Recall  $\mathbf{u} = (u_1, \dots, u_p)'$  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.

$$RV_{i,t+h}^{(h)} = \alpha_i + \sum_{l=1}^p \beta_l u_l + \epsilon_{i,t+h}. \quad (10)$$

### 5.1.3 Least absolute shrinkage and selection operator (LASSO)

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 et al. [39], 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 Eqn (11). Here, the hyperparameter  $\lambda$  controls the penalty weight.

---

<sup>7</sup>Since we use the log-version realized volatility, the multiplication of daily, diurnal, and stochastic intraday components in Engle and Sokalska [30] translates to the addition in our model (9).

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.

$$\mathcal{L}_i = \sum_t \left[ \text{RV}_{i,t+h}^{(h)} - \alpha_i - \sum_{l=1}^p \beta_l u_l \right]^2 + \lambda \sum_{l=1}^p \|\beta_l\|_1. \quad (11)$$

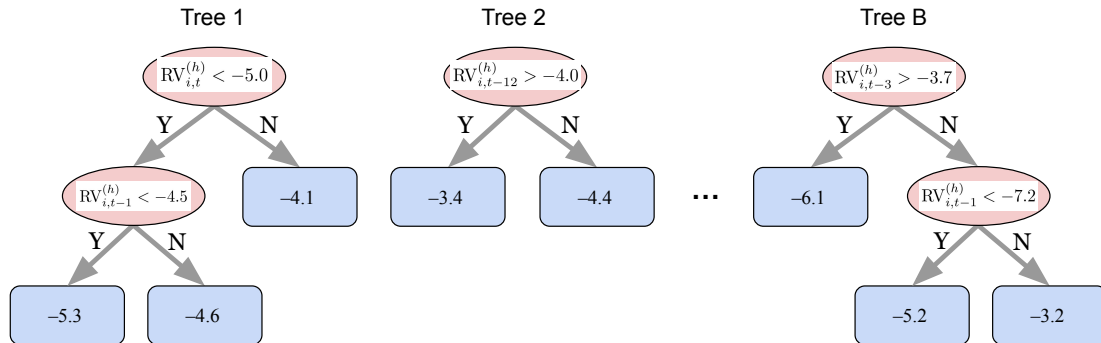
#### 5.1.4 XGBoost

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 [14], 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 et al. [39].

XGBoost is a decision-tree-based ensemble algorithm, implemented under a distributed gradient boosting framework by Chen and Guestrin [17]. 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 [17]. Let  $\mathbf{u}$  represent the vector of input features

$$F_i(\mathbf{u}) = \sum_{l=1}^B f_l(\mathbf{u}), \quad f_l \in \mathcal{F}, \quad (12)$$

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 Eqn (12) is trained sequentially. Boosting (see Friedman [31]) means that new models are added to minimize the errors made by existing models, until no further improvements are achieved.



**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 Eqn (12).

### 5.1.5 Multilayer perceptron (MLP)

Another non-linear method is the neural network (NN), which has become increasingly popular for machine learning problems, e.g. in computer vision and natural language processing, due to its flexibility to learn complex interactions. Hill et al. [42] and Zhang [64] suggest that NN is a promising non-linear alternative to the traditional linear methods in time series forecasting. We will introduce two commonly-implemented NNs in the following sections.

MLP is a class of feedforward neural networks and a “universal approximator” that can learn any smooth functions (see Hornik et al. [44]). MLP has been applied to many fields, e.g. computer vision, natural language processing, etc. 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. The parameters in MLPs can be updated via stochastic gradient descent (SGD). In this work, we use Adam (see Kingma and Ba [48]), which is based on adaptive estimates of lower-order moments. Let  $\mathbf{u} \in \mathbb{R}^p$  represent the input variables

$$F_i(\mathbf{u}; \theta) = \mathbf{W}_L \cdot \sigma(\mathbf{W}_{L-1} \dots \sigma(\mathbf{W}_1 \mathbf{u} + \mathbf{b}_1) \dots + \mathbf{b}_{L-1}) + \mathbf{b}_L, \quad (13)$$

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

### 5.1.6 Long short-term memory (LSTM)

A recurrent neural network (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. LSTM, proposed by Hochreiter and Schmidhuber [43], is an extension of the RNN architecture by replacing each hidden unit in RNNs with a memory block to capture the long-term effect.

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 [43].

$$\begin{aligned}
\mathbf{f}_t &= \sigma_g (\mathbf{W}_f \mathbf{u}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{b}_f) \\
\mathbf{i}_t &= \sigma_g (\mathbf{W}_i \mathbf{u}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{b}_i) \\
\mathbf{o}_t &= \sigma_g (\mathbf{W}_o \mathbf{u}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{b}_o) \\
\tilde{\mathbf{c}}_t &= \sigma_c (\mathbf{W}_c \mathbf{u}_t + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{b}_c) \\
\mathbf{c}_t &= \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \tilde{\mathbf{c}}_t \\
\mathbf{h}_t &= \mathbf{o}_t \circ \sigma_h (\mathbf{c}_t),
\end{aligned} \tag{14}$$

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

## 5.2 Training scheme

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

- **SINGLE** denotes that we train customized models  $F_i$  for each stock  $i$ , as in Bucci [14] and Hansen and Lunde [36]. We use a stock's own past RVs only as predictor features, namely  $\mathbf{u} = \{\mathbf{x}_{i,t}^{(h)}\} = (\text{RV}_{i,t}^{(h)}, \dots, \text{RV}_{i,t-(p-1)h}^{(h)})'$  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 Eqn (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 [61] 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. [12], Engle and Sokalska [30] suggested that models estimated under the **UNIVERSAL** setting yield superior out-of-sample risk forecasts, compared to models under the **SINGLE** setting, when forecasting daily realized volatility.
- **AUGMENTED** denotes that we train models with the pooled data of all stocks in our universe, but in addition, we also incorporate a predictor which takes into account the impact of the market realized volatility (e.g. Bollerslev et al. [12]) in order to leverage the commonality in volatility shown in Section 4. Namely,  $F_i$  is same for all stocks in Eqn (7). We use both individual features  $\mathbf{x}_{i,t}^{(h)} = (\text{RV}_{i,t}^{(h)}, \dots, \text{RV}_{i,t-(p-1)h}^{(h)})'$  and market features  $\mathbf{z}_{i,t}^{(h)} = (\text{RV}_{M,t}^{(h)}, \dots, \text{RV}_{M,t-(p-1)h}^{(h)})'$  as predictors. Note

that for 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 to 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.

### 5.3 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 [56], Patton [55], Engle and Sokalska [30], Pascualau and Poirier [54], Bollerslev et al. [12], Bucci [14], Rahimikia and Poon [58]). Patton and Sheppard [56] demonstrate that QLIKE has the highest power in the Diebold–Mariano (DM) test. Therefore, we focus more on the QLIKE over MSE.

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

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

**Diebold-Mariano (DM) test.** This test is used to discriminate the significant differences of forecasting accuracy between different time series models (for example Diebold and Mariano [27], Diebold [26]). Denote the loss associated with forecast error  $e_t$  by  $L(e_t)$ , e.g.  $L(e_t) = e_t^2$ . 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 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

$$DM_{12} = \frac{\bar{d}^{(a-b)}}{\hat{\sigma}^{(a-b)}} \rightarrow N(0, 1), \quad (15)$$

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

Following Gu et al. [33], we apply a modified DM test, to make pairwise comparisons of models' performance when forecasting multi-asset volatility. In other words, the modified DM test compares the



cross-sectional average of prediction errors from each model, rather than comparing errors for each individual asset, i.e.

$$d_t^{(a-b)} = \frac{1}{N} \sum_{i=1}^N \left( L(e_{i,t}^{(a)}) - L(e_{i,t}^{(b)}) \right), \quad (16)$$

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.

## 5.4 Utility benefits

We can assess the out-of-sample statistical performance for each model via the above metrics and tests. However, this says nothing about the economic magnitude of the gain from complex risk models. Bollerslev et al. [12] 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. As such, we implement this framework to measure the volatility forecasts. For a more detailed description, we refer the reader to Bollerslev et al. [12].

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

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

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 his 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 his wealth at  $t+1$  becomes  $W_{t+1} = W_t \left( 1 + r_t^f + x_t r_{t+1}^e \right)$ , where  $r_{t+1}^e \equiv r_{t+1} - r_t^f$ . After dropping constant terms, the expected utility in Eqn (17) can be expressed as

$$\mathbb{E}_t(u(W_{t+1})) := U(x_t) = W_t \left( x_t \mathbb{E}_t(r_{t+1}^e) - \frac{\gamma}{2} x_t^2 \mathbb{E}_t(\text{RV}_{t+1}) \right), \quad (18)$$

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

Assuming a constant conditional Sharpe ratio  $SR \equiv \mathbb{E}_t(r_{t+1}^e) / \sqrt{\mathbb{E}_t(\text{RV}_{t+1})}$ , the expected utility is

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

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

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

To gauge the utility gains from different risk models, we denote the expectations from model  $\theta$  as  $\mathbb{E}_t^\theta(\cdot)$ . Assuming that the investor uses model  $\theta$ , he will choose the position  $x_t^\theta = \frac{SR/\gamma}{\sqrt{\mathbb{E}_t^\theta(RV_{t+1})}}$ . By plugging  $x_t^\theta$  into Eqn (19) and replacing  $\mathbb{E}_t(RV_{t+1})$  with the realized volatility  $RV_{t+1}$ , the expected utility per unit of his wealth (denoted as realized utility or RU) is

$$RU_t = \frac{SR^2}{\gamma} \times \frac{\sqrt{RV_{t+1}}}{\sqrt{\mathbb{E}_t^\theta(RV_{t+1})}} - \frac{SR^2}{2\gamma} \times \frac{RV_{t+1}}{\mathbb{E}_t^\theta(RV_{t+1})}. \quad (21)$$

If a risk model that perfectly predicts the realized volatilities, i.e.  $\mathbb{E}_t^\theta(RV_{t+1}) = RV_{t+1}$ , its realized utility is  $\frac{SR^2}{2\gamma}$ . Alternatively, the investor is willing to give up  $\frac{SR^2}{2\gamma}$  of his wealth to have access to a perfect risk model instead of investing only in the risk-free asset. In this paper, we choose the same Sharpe ratio ( $SR = 0.4$ ) and the same coefficient of risk aversion ( $\gamma = 2$ ) as in Bollerslev et al. [12], Li and Tang [51].

The previous comparisons are based on a frictionless setting, ignoring the trading cost. We now consider the case of incorporating the effect of transaction costs. Consistent with Bollerslev et al. [12], 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 as in Li and Tang [51]. The realized utility after trading costs (denoted as RU-TC) is simply subtracting the simulated costs from realized utility. 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 as in Section 5.3.

## 6 Experiments

### 6.1 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  $[2011-07-01, t - 251]$ , are used to estimate models subject to a given architecture. Validation data, including the recent one-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 expands by one year, whereas the validation samples are rolled forward to include the samples in the most recent one-year period, following Gu et al. [33]. 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 2015-07-01 until 2016-

06-30, and the corresponding training and validation samples are [2011-07-01, 2014-06-30] and [2014-07-01, 2015-06-30], respectively. When we predict the realized volatility in [2016-07-01, 2017-06-30], the training and validation samples are [2011-07-01, 2015-06-30] and [2015-07-01, 2016-06-30], 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 neural networks, we apply an ensemble approach to MLPs and LSTMs for improving their robustness (see Hansen and Salamon [35], Gu et al. [33]). Specifically, we train multiple neural networks 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.

## 6.2 Main results

Table 3 presents the results of each model under three training schemes. Due to limited computation power, MLPs and LSTMs are only performed under the **UNIVERSAL** and **AUGMENTED** settings. To more formally assess the statistical significance of the differences in out-of-sample volatility forecasts, Table 9 in Appendix C reports the results of all DM tests in terms of QLIKE. We draw the following conclusions from Table 3.

Regarding linear models, we first observe that for HAR-d, **UNIVERSAL** shows no improvement in forecasting, compared to **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 to **SINGLE** is 0.031, -0.005, 0.004, 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 are reduced from 0.453 (resp. 0.227, 0.186) with the best HAR-d model (i.e. under **AUGMENTED**) to 0.430 (resp. 0.204, 0.171) with the best OLS model (i.e. under **AUGMENTED**), across the three horizons (i.e. 10, 30, 65-min), respectively. Within the OLS models, conclusions are similar with HAR-d models, i.e. 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, 65-min), implying the complex interactions between predictors. Further analysis is

provided in Section 6.3.

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 neural networks to underperform due to lack of training data.

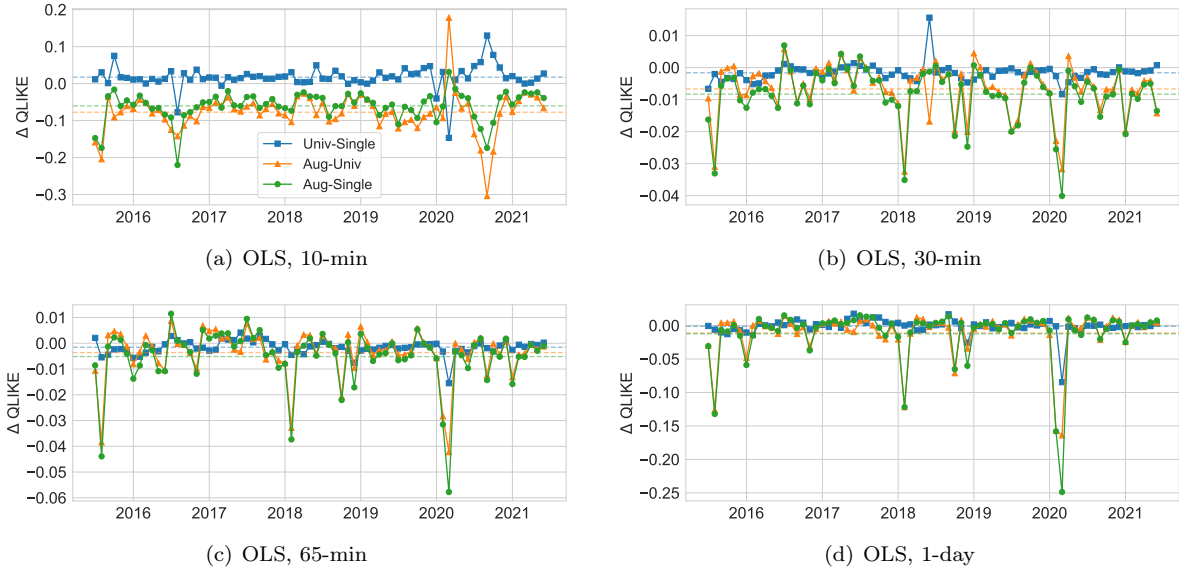
Echoing the findings from Panel A, OLS based on the 21-day rolling daily RVs deliver the higher utility than the HAR type models, consistent with Bollerslev et al. [12]. Neural networks still perform the best, with the highest realized utility achieved by LSTMs.

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  $\text{Aug-Single} = (\text{Aug-Univ}) + (\text{Univ-Single})$ . Negative values of  $\Delta\text{QLIKE}$  indicate an improvement on out-of-sample data, and positive values indicate degradation. To arrive at this figure, we average the  $\Delta\text{QLIKE}$  values in each month, across stocks. Figure 7 reveals that the improvement of **UNIVERSAL** compared to **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.

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

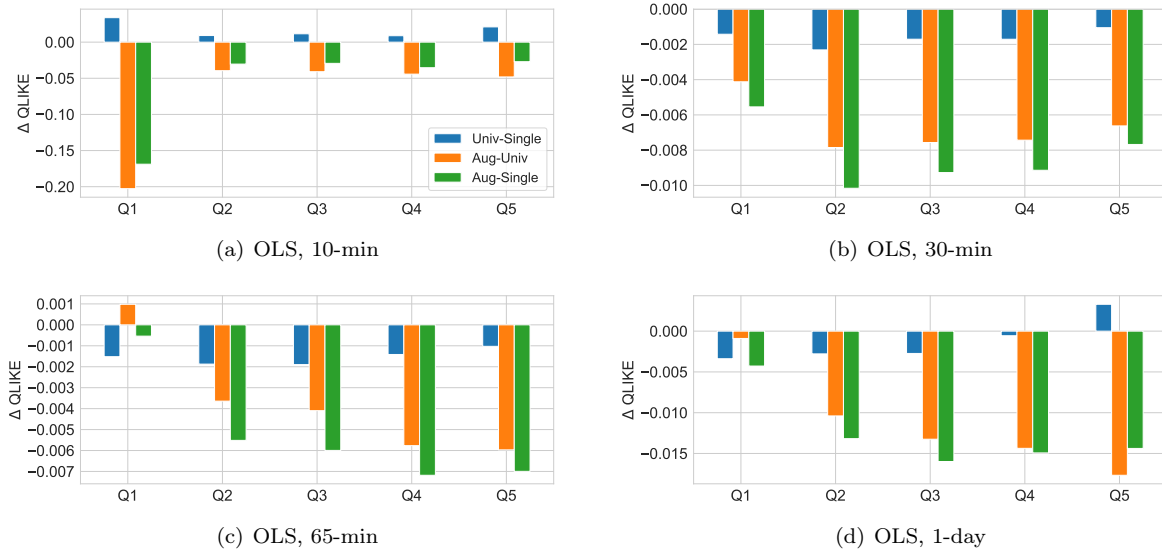
<b>Panel A:</b>		10-min		30-min		65-min		1-day	
<b>Statistical performance</b>		MSE	QLIKE	MSE	QLIKE	MSE	QLIKE	MSE	QLIKE
HAR-d	<b>SINGLE</b>	1.013	0.484	0.332	0.222	0.270	0.190	0.269	0.190
	<b>UNIVERSAL</b>	1.021	0.518	0.333	0.230	0.270	0.191	0.269	0.190
	<b>AUGMENTED</b>	0.995	0.453	0.323	0.227	0.262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



**Figure 7:** Pairwise  $\Delta\text{QLIKE}$  of the OLS model across three training schemes.

*Notes:*  $\Delta\text{QLIKE}$  is averaged across stocks in each month during the testing period 2015-07 ~ 2021-06. The dashed horizontal lines represent the average reductions in MSE.



**Figure 8:** Pairwise  $\Delta\text{QLIKE}$  of the OLS model sorted by commonality.

*Notes:* Q1, respectively Q5, denote 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 Eqn (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.

### 6.3 Variable importance & interaction effects

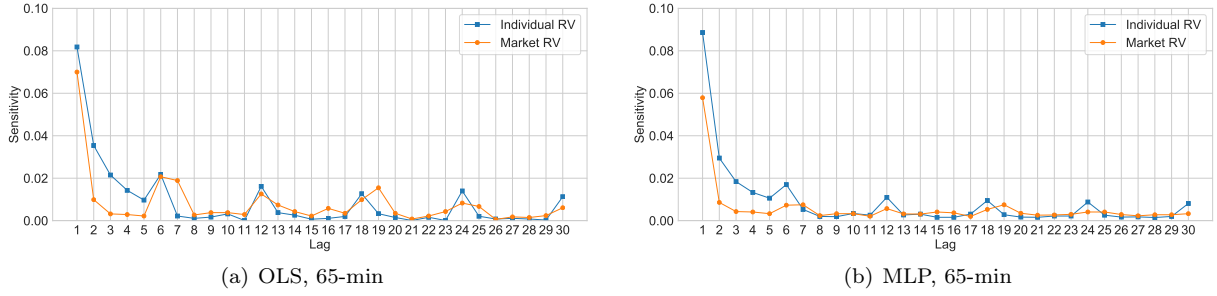
This section provides intuition for why neural networks perform as strong as they do, with an eye towards explainability. Due to the use of non-linear activation functions and multiple hidden layers, neural networks enjoy the benefit of allowing for potentially complex interactions among predictors, 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 neural networks and explain their competitive performance.

**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 et al. [59]) 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

$$\text{Sensitivity}_k = \sum_{i=1}^N \sum_{t \in \mathcal{T}_{train}} \left\| \frac{\partial F}{\partial u_k} \Big|_{\mathbf{u}=\mathbf{u}_{i,t}} \right\|_1. \quad (22)$$

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 all variables’ sensitivity, such that they sum up to one. In a special case of a 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 6 lags, corresponding to 1 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 (resp. market) RV are 0.081 (resp. 0.069).



**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 5 days for 65-min horizon).

**Interaction effects.** To analyze the interactions between the two most significant features implied by neural networks, we adopt an approach (e.g. Gu et al. [33], Choi et al. [18]) that focuses on the partial relations between a pair of input variables and responses, while fixing other variables at their mean values

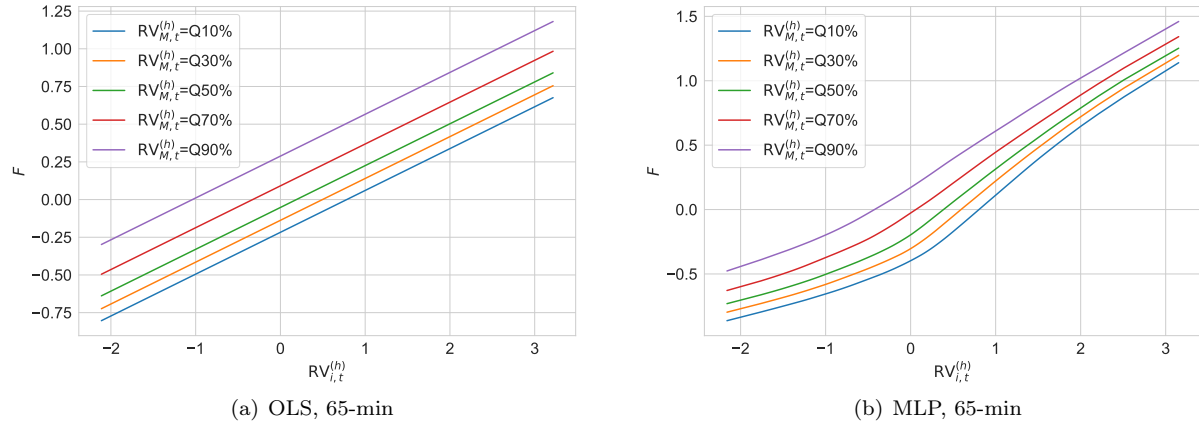
$$F_{j|i}(u_j | u_i = q) = F(u_j, u_i = q, u_k = \bar{u}_k, k \neq i, j), \quad (23)$$

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  $RV_{i,t}^{(h)}$  and  $RV_{M,t}^{(h)}$ , over their support<sup>8</sup>. In particular, we analyze the interaction of the lag=1 individual RV ( $RV_{i,t}^{(h)}$ ), with the lag=1 market RV ( $RV_{M,t}^{(h)}$ ). As shown in Figure 10(a), a parallel shift between the different curves occurs if there are no interaction effects between  $RV_{i,t}^{(h)}$  and  $RV_{M,t}^{(h)}$ . Figure 10(b) first reveals that the predicted volatility is non-linear in  $RV_{i,t}^{(h)}$ , and the slope of that relationship becomes higher after  $RV_{i,t}^{(h)}$  exceeds a certain threshold (around 0.5). Furthermore, it demonstrates clear interaction effects between  $RV_{i,t}^{(h)}$  and  $RV_{M,t}^{(h)}$ . 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.

<sup>8</sup>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).

## 6.4 Forecasting RVs of unseen stocks

To examine the ability to generalize and address concerns about overfitting, we perform a stringent out-of-sample test, i.e. 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 [61], Choi et al. [18]. For better distinction, we denote the stocks used for estimating machine learning models as **raw stocks**, and those new stocks not in the training sample as **unseen stocks**<sup>9</sup>. We follow the procedure of training, validation, and testing periods described in Section 6.1. 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 models trained on pooled data of raw stocks have better forecasting performance compared to baselines, across all horizons. This presents new empirical evidence for a *universal volatility mechanism* among stocks. Furthermore, neural networks 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, neural networks obtain comparable results (QLIKE=0.252) to the best non-neural network model (MSE=0.249, attained by LASSO). The realized utility of the different risk models echoes that of the out-of-sample QLIKE.

<sup>9</sup>The set of unseen stocks includes the following 17 tickers: AMAT, AON, APD, BIIB, COF, DE, EQIX, EW, GPN, HUM, ICE, ILMN, ITW, NOC, NSC, PLD, SLB.

**Table 4:** Performance on unseen stocks.

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

## 7 Forecasting daily RVs with intraday RVs

Given the fact that intraday volatility exhibits a high and stable commonality (see Sections 3 and 4), we are interested in the potential benefits of using past intraday RVs to forecast daily RVs.

## 7.1 Related literature

Generally speaking, there are two broad families of models used to forecast daily volatility: (i) GARCH and stochastic volatility (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 realized volatility is a superior proxy for the unobserved daily volatility, when compared to the parametric volatility measures generated from the GARCH and SV models of daily returns (see Andersen et al. [4], Barndorff-Nielsen and Shephard [9], Izzeldin et al. [45]). It is worth noting that in these traditional forecasting daily RV models (e.g. ARFIMA of Andersen et al. [4], HAR of Corsi [21], SHAR of Patton and Sheppard [57], HARQ of Bollerslev et al. [13]), 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 [11] also pointed out that “semivariation measured over shorter interday time intervals may afford additional useful information.”

## 7.2 Proposed approach

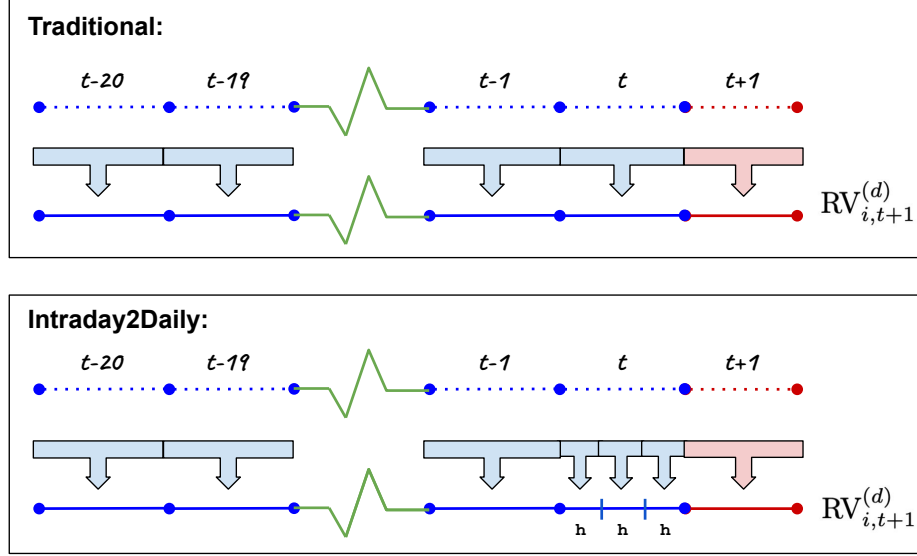
Two new paragraphs In Section 5.1, we introduced a set of commonly used models, where the daily variables (lagged daily RVs) are employed as predictors when forecasting 1-day RVs. For simplicity, we refer to these models as **traditional** approaches. Previous sections also concluded that the most recent 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, denoted by **Intraday2Daily** approach.

The formulation of our approach is also motivated by the time-of-day effect in volatility and draws on insights from the mixed data sampling (MIDAS) approach of Ghysels et al. [32] and the “Rolling” approach proposed by Pascalau and Poirier [54]. However, the classic MIDAS approach uses smooth distributed lag polynomials of daily RVs for representing the dynamic dependencies, failing to collect the information of the intraday RVs. Pascalau and Poirier [54] enrich the training samples by simply summing trailing M returns at each timestamp, regardless if all M returns belong to the same calendar date. In contrast to Ghysels et al. [32], Pascalau and Poirier [54], our **Intraday2Daily** approach takes the time-of-day effect into account in an explicit way

$$RV_{i,t+1}^{(d)} = F_i \left( \mathbf{x}_{i,t}^{(h)}, \mathbf{z}_t^{(h)}, \mathbf{x}_{i,:t-1}^{(d)}, \mathbf{z}_{:t-1}^{(d)}; \theta \right) + \epsilon_{i,t+1}. \quad (24)$$

Recall that  $RV_{i,t+1}^{(d)}$  is the RV of stock  $i$  at day  $t+1$ .  $\mathbf{x}_{i,t}^{(h)}$  represents a multi-dimensional vector of predictors for stock  $i$  computed over shorter intraday time horizons at day  $t$  and  $\mathbf{z}_t^{(h)}$  is a vector of predictors for all stocks computed over shorter intraday time horizons at day  $t$ .  $\mathbf{x}_{i,:t}^{(d)}$  denotes daily features for a specific stock

$i$  available up to day  $t-1$  and  $\mathbf{z}_{i,t}^{(h)}$  denoted daily features for all stocks up to  $t-1$ . Departing from traditional models where all the variables are computed in the daily frequency, we decompose the lag-one total RV ( $\mathbf{x}_{i,t}^{(d)}$ ) to sub-sampled RVs ( $\mathbf{x}_{i,t}^{(h)}$ ), computed for intervals of length  $h$ . Figure 11 illustrates the comparison between the traditional approach and our **Intraday2Daily** approach.



**Figure 11:** Illustration of two prediction approaches for future daily volatility (red segment).

*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 (long blue segments) as predictors, while the **Intraday2Daily** approach employs intraday RVs (short blue segments between two adjacent vertical ticks).  $h$  represents the horizon of intraday RVs. In this example,  $h = 130$  minutes.

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 (or monthly) risk measures, such as value-at-risk (VaR), etc. For example, one may use half-hour VaRs to forecast the one-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.*

### 7.3 Experiments

The forecasting performance of traditional approaches with daily variables are already summarized in the column '1-day' of Table 3. Table 5 reports the results of models combined with the **Intraday2Daily**

approach<sup>10</sup>. In other words, models in Table 5 use sub-sampled intraday 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 (Eqn (8)) 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 to 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. 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. We reach to the same conclusions in terms of realized utilities.

## 7.4 Robustness check

In this section, we present the empirical analysis of examining the robustness of **Intraday2Daily** approach when incorporating new type of predictors (including semi-RV of Patton and Sheppard [57], realized quarticity of Bollerslev et al. [13]).

**SHAR.** Patton and Sheppard [57] proposed the Semi-variance-HAR (SHAR) model as an extension of the standard HAR model (see further details in 5.1.1), 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 Eqn (25) (see Barndorff-Nielsen et al. [8]). In other words, the lag one RV in SHAR (Eqn (26)) is split into the sum of squared positive returns and the sum of squared negative returns, as follows.

$$\begin{aligned} RV_{i,t}^{(d)+} &= \sum_{l=0}^{M-1} r_{i,t-l\Delta}^2 I_{\{r_{t-l\Delta} > 0\}} \\ RV_{i,t}^{(d)-} &= \sum_{l=0}^{M-1} r_{i,t-l\Delta}^2 I_{\{r_{t-l\Delta} < 0\}} \end{aligned} \quad (25)$$

$$RV_{i,t+1}^{(d)} = \alpha_i + \beta_i^{(d)+} RV_{i,t}^{(d)+} + \beta_i^{(d)-} RV_{i,t}^{(d)-} + \beta_i^{(w)} RV_{i,w}^{(w)} + \beta_i^{(m)} RV_{i,m}^{(m)} + \epsilon_{i,t+1}. \quad (26)$$

Here  $\Delta$  is the interval for computing the intraday returns.

**HARQ.** Bollerslev et al. [13] 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

<sup>10</sup>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.

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

*Notes:* The table reports the out-of-sample results for predicting future realized volatility over multiple horizons 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.

realized volatility estimation, the authors propose an easy-to-implement model, termed as HARQ (Eqn (28)). The realized quarticity (RQ) is estimated according to Eqn (27), aiming to correct the measurement errors.

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

$$RV_{i,t+1}^{(d)} = \alpha_i + \left( \beta_i^{(d)} + \beta_i^{(d)Q} \sqrt{RQ_{i,t}^{(d)}} \right) RV_{i,t}^{(d)} + \beta_i^{(w)} RV_{i,t}^{(w)} + \beta_i^{(m)} RV_{i,t}^{(m)} + \epsilon_{i,t+1}. \quad (28)$$

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 et al. [13]. 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.

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

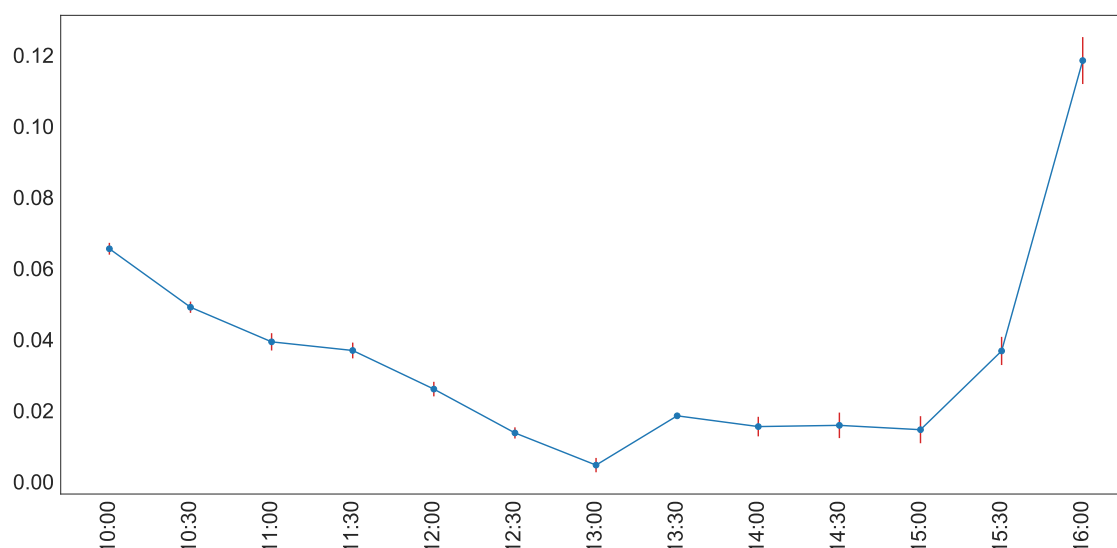
<b>Panel A:</b>		10-min		30-min		65-min		Traditional	
<b>Statistical performance</b>		MSE	QLIKE	MSE	QLIKE	MSE	QLIKE	MSE	QLIKE
SHAR	<b>SINGLE</b>	0.277	0.191	0.257	0.178	0.253	0.176	0.261	0.183
	<b>UNIVERSAL</b>	0.285	0.198	0.263	0.183	0.255	0.178	0.261	0.182
	<b>AUGMENTED</b>	0.261	0.181	0.253	0.175	0.250	0.174	0.254	0.178
HARQ	<b>SINGLE</b>	0.264	0.204	0.254	0.178	0.253	0.176	0.256	0.179
	<b>UNIVERSAL</b>	0.253	0.176	0.253	0.176	0.254	0.176	0.257	0.179
	<b>AUGMENTED</b>	0.251	0.174	0.248	0.172	0.250	0.174	0.253	0.176
<b>Panel B:</b>		10-min		30-min		65-min		Traditional	
<b>Realized utility</b>		RU	RU-TC	RU	RU-TC	RU	RU-TC	RU	RU-TC
SHAR	<b>SINGLE</b>	3.528	3.497	3.559	3.525	3.563	3.529	3.548	3.515
	<b>UNIVERSAL</b>	3.510	3.499	3.548	3.525	3.560	3.529	3.550	3.516
	<b>AUGMENTED</b>	3.563	3.533	3.576	3.545	3.578	3.545	3.571	3.537
HARQ	<b>SINGLE</b>	3.467	3.425	3.556	3.520	3.564	3.528	3.557	3.525
	<b>UNIVERSAL</b>	3.564	3.530	3.564	3.530	3.564	3.530	3.558	3.525
	<b>AUGMENTED</b>	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 realized volatility over multiple horizons 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.

## 7.5 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 zero and a standard deviation of one. 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<sup>11</sup>. 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.



**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.

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 Banerji [7], 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 US 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.

<sup>11</sup>We attain similar results for models using intraday RVs based on other frequencies.



## 8 Conclusion

In this paper, 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 machine learning models in our analysis, pooling stock data together (**UNIVERSAL**) and adding the market volatility as an additional predictor (**AUGMENTED**) generally improves the out-of-sample performance, in comparison with asset-specific models (**SINGLE**).

We show that neural networks 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 neural networks 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.

**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 realized volatility, 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, e.g. Value-at-Risk, that could potentially benefit from time-of-day dependent features.

## References

- [1] Andersen, T. G. and T. Bollerslev (1997). Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance* 4(2-3), 115–158.
- [2] Andersen, T. G., T. Bollerslev, P. F. Christoffersen, and F. X. Diebold (2006). Volatility and correlation forecasting. *Handbook of Economic Forecasting* 1, 777–878.
- [3] Andersen, T. G., T. Bollerslev, F. X. Diebold, and H. Ebens (2001). The distribution of realized stock return volatility. *Journal of Financial Economics* 61(1), 43–76.
- [4] Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003). Modeling and forecasting realized volatility. *Econometrica* 71(2), 579–625.

- [5] Baker, M. and J. Wurgler (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives* 21(2), 129–152.
- [6] Baker, S. R., N. Bloom, and S. J. Davis (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics* 131(4), 1593–1636.
- [7] Banerji, G. (2020). The 30 minutes that can make or break the trading day.
- [8] Barndorff-Nielsen, O. E., S. Kinnebrock, and N. Shephard (2008). Measuring downside risk-realised semivariance. *CREATES Research Paper* (2008-42).
- [9] Barndorff-Nielsen, O. E. and N. 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(2), 253–280.
- [10] Bollen, B. and B. Inder (2002). Estimating daily volatility in financial markets utilizing intraday data. *Journal of Empirical Finance* 9(5), 551–562.
- [11] Bollerslev, T. (2022). Realized semi (co) variation: Signs that all volatilities are not created equal. *Journal of Financial Econometrics* 20(2), 219–252.
- [12] Bollerslev, T., B. Hood, J. Huss, and L. H. Pedersen (2018). Risk everywhere: Modeling and managing volatility. *Review of Financial Studies* 31(7), 2729–2773.
- [13] Bollerslev, T., A. J. Patton, and R. Quaedvlieg (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics* 192(1), 1–18.
- [14] Bucci, A. (2020). Realized volatility forecasting with neural networks. *Journal of Financial Econometrics* 18(3), 502–531.
- [15] Calvet, L. E., A. J. Fisher, and S. B. Thompson (2006). Volatility comovement: a multifrequency approach. *Journal of Econometrics* 131(1-2), 179–215.
- [16] Carroll, C. D., J. C. Fuhrer, and D. W. Wilcox (1994). Does consumer sentiment forecast household spending? if so, why? *American Economic Review* 84(5), 1397–1408.
- [17] Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.
- [18] Choi, D., W. Jiang, and C. Zhang (2021). Alpha go everywhere: Machine learning and international stock returns. *Working paper*.

- [19] Chordia, T., R. Roll, and A. Subrahmanyam (2000). Commonality in liquidity. *Journal of Financial Economics* 56(1), 3–28.
- [20] Christensen, K., M. Siggaard, B. Veliyev, et al. (2021). *A machine learning approach to volatility forecasting*, Volume 3. Department of Economics and Business Economics, Aarhus University.
- [21] Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics* 7(2), 174–196.
- [22] Da, Z., J. Engelberg, and P. Gao (2011). In search of attention. *Journal of Finance* 66(5), 1461–1499.
- [23] Da, Z., J. Engelberg, and P. Gao (2015). The sum of all fears investor sentiment and asset prices. *Review of Financial Studies* 28(1), 1–32.
- [24] Dang, T. L., F. Moshirian, and B. Zhang (2015). Commonality in news around the world. *Journal of Financial Economics* 116(1), 82–110.
- [25] De Long, J. B., A. Shleifer, L. H. Summers, and R. J. Waldmann (1990). Noise trader risk in financial markets. *Journal of Political Economy* 98(4), 703–738.
- [26] Diebold, F. 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), 1–1.
- [27] Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics* 13(3), 253–263.
- [28] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, 987–1007.
- [29] Engle, R. F. and A. J. Patton (2007). What good is a volatility model? In *Forecasting volatility in the financial markets*, pp. 47–63. Elsevier.
- [30] Engle, R. F. and M. E. Sokalska (2012). Forecasting intraday volatility in the us equity market. multiplicative component garch. *Journal of Financial Econometrics* 10(1), 54–83.
- [31] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.
- [32] Ghysels, E., P. Santa-Clara, and R. Valkanov (2006). Predicting volatility: getting the most out of return data sampled at different frequencies. *Journal of Econometrics* 131(1-2), 59–95.

- [33] Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *Review of Financial Studies* 33(5), 2223–2273.
- [34] Hameed, A., W. Kang, and S. Viswanathan (2010). Stock market declines and liquidity. *Journal of Finance* 65(1), 257–293.
- [35] Hansen, L. K. and P. Salamon (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12(10), 993–1001.
- [36] Hansen, P. R. and A. Lunde (2005). A forecast comparison of volatility models: does anything beat a garch (1, 1)? *Journal of Applied Econometrics* 20(7), 873–889.
- [37] Hansen, P. R. and A. Lunde (2006). Realized variance and market microstructure noise. *Journal of Business & Economic Statistics* 24(2), 127–161.
- [38] Harris, L. (1986). A transaction data study of weekly and intradaily patterns in stock returns. *Journal of Financial Economics* 16(1), 99–117.
- [39] Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
- [40] Herskovic, B., B. Kelly, H. Lustig, and S. Van Nieuwerburgh (2016). The common factor in idiosyncratic volatility: Quantitative asset pricing implications. *Journal of Financial Economics* 119(2), 249–283.
- [41] Herskovic, B., B. Kelly, H. Lustig, and S. Van Nieuwerburgh (2020). Firm volatility in granular networks. *Journal of Political Economy* 128(11), 4097–4162.
- [42] Hill, T., M. O'Connor, and W. Remus (1996). Neural network models for time series forecasts. *Management Science* 42(7), 1082–1092.
- [43] Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural Computation* 9(8), 1735–1780.
- [44] Hornik, K., M. Stinchcombe, and H. White (1989). Multilayer feedforward networks are universal approximators. *Neural Networks* 2(5), 359–366.
- [45] Izzeldin, M., M. K. Hassan, V. Pappas, and M. Tsionas (2019). Forecasting realised volatility using arfima and har models. *Quantitative Finance* 19(10), 1627–1638.
- [46] Karolyi, G. A., K.-H. Lee, and M. A. Van Dijk (2012). Understanding commonality in liquidity around the world. *Journal of Financial Economics* 105(1), 82–112.

- [47] Keynes, J. M. (2018). *The general theory of employment, interest, and money*. Springer.
- [48] Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization. *Working paper*.
- [49] Kogan, L., S. A. Ross, J. Wang, and M. M. Westerfield (2006). The price impact and survival of irrational traders. *Journal of Finance* 61(1), 195–229.
- [50] Lemmon, M. and E. Portniaguina (2006). Consumer confidence and asset prices: Some empirical evidence. *Review of Financial Studies* 19(4), 1499–1529.
- [51] Li, S. Z. and Y. Tang (2020). Forecasting realized volatility: An automatic system using many features and many machine learning algorithms. *Working paper*.
- [52] Liu, L. Y., A. J. Patton, and K. Sheppard (2015). Does anything beat 5-minute rv? a comparison of realized measures across multiple asset classes. *Journal of Econometrics* 187(1), 293–311.
- [53] Morck, R., B. Yeung, and W. Yu (2000). The information content of stock markets: why do emerging markets have synchronous stock price movements? *Journal of Financial Economics* 58(1-2), 215–260.
- [54] Pascalau, R. and R. Poirier (2021). Increasing the information content of realized volatility forecasts. *Journal of Financial Econometrics*.
- [55] Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics* 160(1), 246–256.
- [56] Patton, A. J. and K. Sheppard (2009). Evaluating volatility and correlation forecasts. In *Handbook of Financial Time Series*, pp. 801–838. Springer.
- [57] Patton, A. J. and K. Sheppard (2015). Good volatility, bad volatility: Signed jumps and the persistence of volatility. *Review of Economics and Statistics* 97(3), 683–697.
- [58] Rahimikia, E. and S.-H. Poon (2020). Machine learning for realised volatility forecasting. *Working paper*.
- [59] Sadhwani, A., K. Giesecke, and J. Sirignano (2021). Deep learning for mortgage risk. *Journal of Financial Econometrics* 19(2), 313–368.
- [60] Shleifer, A. and L. H. Summers (1990). The noise trader approach to finance. *Journal of Economic Perspectives* 4(2), 19–33.
- [61] Sirignano, J. and R. Cont (2019). Universal features of price formation in financial markets: perspectives from deep learning. *Quantitative Finance* 19(9), 1449–1459.

- [62] Taylor, S. J. and X. Xu (1997). The incremental volatility information in one million foreign exchange quotations. *Journal of Empirical Finance* 4(4), 317–340.
- [63] Xiong, R., E. P. Nichols, and Y. Shen (2015). Deep learning stock volatility with google domestic trends. *Working paper*.
- [64] Zhang, G. P. (2003). Time series forecasting using a hybrid arima and neural network model. *Neurocomputing* 50, 159–175.

## A What may drive commonality in volatility?

Previous studies, especially in the behavioural finance field, have shown that investor sentiments could affect stock prices (e.g. Baker and Wurgler [5], Kogan et al. [49], Da et al. [22], Bollerslev et al. [12], Da et al. [23], Karolyi et al. [46], Hameed et al. [34]). Keynes [47] argued that animal spirits affect consumer confidence, thereby moving prices in times of high levels of uncertainty. De Long et al. [25], Shleifer and Summers [60], Kogan et al. [49] found that investor sentiments induce excess volatility. Karolyi et al. [46] considered the investor sentiment index as an important source of commonality in liquidity. Bollerslev et al. [12] 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 measure investor sentiments (see Da et al. [23]), i.e. market-based measures and survey-based indices. Following Baker and Wurgler [5], 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)<sup>12</sup> by the University of Michigan’s Survey Research Center as a proxy for survey-based indices (see Carroll et al. [16], Lemmon and Portniaguina [50]). 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 et al. [23], we also include a news-based index EPU<sup>13</sup> proposed by Baker et al. [6] to measure policy-related economic uncertainty.

As suggested by Morck et al. [53], the raw monthly commonality measures  $R^2_{(h),m}$  (computed based on Eqn (5)) are inappropriate to use as the dependent variable in regressions, because they are bounded by 0 and 1. Consistent with Morck et al. [53], Karolyi et al. [46] and Dang et al. [24], we take the logistic transformation of  $R^2_{(h),m}$ , i.e.  $\log \left[ R^2_{(h),m} / (1 - R^2_{(h),m}) \right]$ , 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 Eqn (29)

<sup>12</sup><http://www.sca.isr.umich.edu>

<sup>13</sup><https://www.policyuncertainty.com>

$$(R_{(h),m}^2)_L = \alpha + \beta_1 \text{CSI}_m + \beta_2 \text{VIX}_m + \beta_3 \text{EPU}_m + \epsilon_{i,t}. \quad (29)$$

Table 7 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-squared 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-squared 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 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.

**Table 7:** Time series regression of commonality.

	10-min	30-min	65-min	1-day
VIX	0.233* (0.030)	0.196* (0.024)	0.192* (0.023)	0.714* (0.084)
CSI	0.214* (0.025)	0.097* (0.020)	0.066* (0.019)	0.237* (0.070)
EPU	0.079* (0.029)	0.025 (0.023)	0.022 (0.022)	0.114 (0.080)
Constant	0.161* (0.023)	0.982* (0.018)	1.267* (0.018)	-0.689* (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  $(R_{(h),m}^2)_L$  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.

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, 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 5-fold cross-validation method to determine  $\lambda_1$ . Hyperparameters for other models are summarized as follows.

**Table 8:** 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
Epoques	-	100	100
No. of hidden layers	-	3	2
Batch normalization	-	✓	✗

## C Diebold-Mariano test

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 9:** Statistics of Diebold-Mariano tests.

Panel A: 10-min.

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

Panel C: 65-min.

Univ Univ	HAR-d	OLS	LASSO	XGBoost	MLP	LSTM	Aug Aug	HAR-d	OLS	LASSO	XGBoost	MLP	LSTM
HAR-d		28.27*	27.75*	21.56*	30.06*	29.00*	HAR-d		22.11*	22.67*	9.87*	25.92*	26.01*
OLS			-11.73*	7.78*	20.22*	18.67*	OLS			-4.89*	-5.56*	12.84*	11.79*
Lasso				8.81*	20.91*	19.35*	Lasso				-5.11*	13.72*	12.67*
XGBoost					19.83*	18.17*	XGBoost					17.71*	18.54*
MLP						0.68*	MLP						0.94*
LSTM							LSTM						
Single vs	-1.87	8.17*	8.53*	41.26*			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	Aug Aug	HAR-d	OLS	LASSO	XGBoost	MLP	LSTM
HAR-d		0.50	-0.12	-4.29*	2.69*	3.86*	HAR-d		-0.12	5.39*	-8.97*	3.20*	1.87
OLS			-4.94*	-5.50*	1.91	3.36*	OLS			-1.10	-12.63*	-3.77*	-3.99*
Lasso				-4.29*	2.76*	3.88*	Lasso				-12.68*	-3.34*	-3.91*
XGBoost					9.49*	10.90*	XGBoost					12.30*	11.61*
MLP						3.03*	MLP						-2.20*
LSTM							LSTM						
Single vs	-1.32	3.41*	5.42*	20.53*			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%.

## D Model update frequency

In the present paper, 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, i.e. weekly, monthly, and yearly, and results are summarized as follows.

**Table 10:** 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