Volatility forecasting with machine learning and intraday commonality

Chao Zhang*1, Yihuang Zhang*2,4, Mihai Cucuringu^{1,2,3}, and Zhongmin Qian^{1,4}

¹Department of Statistics, University of Oxford, Oxford, UK

²Mathematical Institute, University of Oxford, Oxford, UK

³The Alan Turing Institute, London, UK

⁴Oxford Suzhou Centre for Advanced Research

February 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 diurnal 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

^{*}Equal contribution. Correspondence to: Chao Zhang chao.zhang@stats.ox.ac.uk

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. Intraday volatility estimates are important for pricing derivatives, managing risk, and devising quantitative 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 that conventional parametric models, such as GARCH and stochastic volatility models, may be inadequate for modeling intraday returns [1]. In recent works [4, 20], 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 and daily 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. The data we explore spans the period from July 2011 to June 2021, and includes the top 100 most liquid components of S&P500 index. In our approach, the 10-min, 30-min, 65-min, and daily (without overnight information) forecasting horizons are studied.

Main contributions. 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 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 [58] 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 [20], SHAR [54], HARQ [12]). 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.

Paper outline. 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. [18] have recognized the existence of commonality in liquidity, and Karolyi et al. [43] have suggested that commonality in liquidity is related to market volatility, in particular, the presence of international investors and trading activity. Dang et al. [23] 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. [14]) all make use of the volatility spillover

effects. Herskovic et al. [38] have provided empirical evidence of the co-movement in volatility across the equity market. Bollerslev et al. [11] have observed strong similarities in daily realized volatility and have utilized them to forecast the daily realized volatility. Engle and Sokalska [29] have emphasised that pooled data is useful for intraday volatility forecasting, and Herskovic et al. [39] 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 [28], 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 [20] 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. [42] 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 [59] proposed an hourly volatility model based on an ARCH specification, and Engle and Sokalska [29] 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.

[31] have pointed out the superior performance of ML models for empirical asset pricing. Sirignano and Cont

[58] 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. [60] have applied LSTMs to forecast S&P 500 volatility, with Google domestic trends as predictors, and Bucci [13] 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 [55] 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 [48] have proposed a simple average ensemble model combining multiple machine learning algorithms for forecasting daily (and monthly) realized volatility, and Christensen et al. [19] 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 LOBSTER¹ to compute intraday returns via mid-prices. We select the top 100 components of S&P500 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 division².

Sector	Number	Tickers
Information Technology	20	AAPL ACN ADBE ADP AVGO CRM CSCO FIS FISV IBM
information Technology	20	INTC INTU MA MSFT MU NVDA ORCL QCOM TXN V
Health Care	19	ABT AMGN BDX BMY BSX CI CVS DHR GILD ISRG
Health Care	19	JNJ LLY MDT MRK PFE SYK TMO UNH VRTX
Financials	15	AXP BAC BLK BRK.B C CB CME GS JPM MMC
1 manerais	10	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

Table 1: Components in each sector.

3.2 Realized volatility

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

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

¹https://lobsterdata.com/

²The Global Industry Classification Standard (GICS) is an industry taxonomy developed in 1999 by MSCI and Standard & Poor's (S&P).

where μ_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

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

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

Throughout this paper, we consider the minutely logarithmic return for asset i during (t-1,t] as

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

Here, $P_{i,t}$ is the mid-price at time t, 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 square root of the unobserved IV (see Bollen and Inder [10], Hansen and Lunde [35], Andersen et al. [3]). To reduce the impact of extreme values, we consider the logarithm, in line with Andersen et al. [4], Bucci [13], Herskovic et al. [38]. Specifically, during a period (t - h, t], the realized volatility is defined as follows³

$$RV_{i,t}^{(h)} := \log \sqrt{\sum_{s=t-h+1}^{t} r_{i,s}^2} = \frac{1}{2} \log \left[\sum_{s=t-h+1}^{t} r_{i,s}^2 \right].$$
 (4)

As pointed out by Pascalau and Poirier [51], there are no conclusive methods to incorporate the overnight session's information content into the daily volatility. In line with Engle and Sokalska [29], 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

³Liu et al. [49] 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.

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.

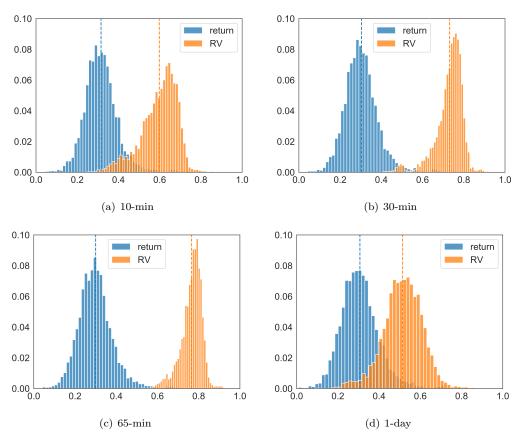


Figure 1: Histograms of pairwise correlations of realized volatilities and returns. (a)-(d) are based on observations in the frequency of 10-min, 30-min, 65-min, 1-day, respectively. The dashed vertical lines represent the average correlation values of RVs and returns.

Figure 2 plots the daily realized volatility over time. Stocks demonstrate similar time series patterns, consistent with Herskovic et al. [38], Bollerslev et al. [11]. 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 (Harris [36], Engle and Sokalska [29]).

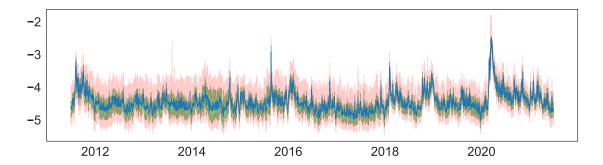


Figure 2: Daily realized volatility (in logs). The blue curve represents cross-sectional average of daily realized volatility across stocks, with the green area covering the 25-th percentile to the 75-th percentile, and the red area covering the 5-th percentile to the 95-th percentile.

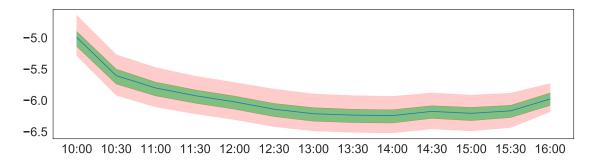


Figure 3: Diurnal realized volatility (in logs). The blue curve represents cross-sectional average of 30-min realized volatility across stocks and days, with the green area covering the 25-th percentile to the 75-th percentile and the red area covering the 5-th percentile to the 95-th percentile.

4 Commonality estimation

Inspired by prior studies (e.g., Morck et al. [50], Chordia et al. [18], Karolyi et al. [43], Dang et al. [23]), 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 $R_{(h)}^2$)⁴

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

 $^{^4}$ 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 [18, 43, 23]. The R-squared values are similar to the ones of Eqn (5).

where $RV_{M,t}^{(h)}$ is the contemporaneous market volatility during (t-h,t] for stock i, which is calculated as the equally weighted average of all individual stock volatilities during (t-h,t].

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. Table 2 also reports the results of the relation between the average commonality and the market volatility. As the horizon extends, the average commonality has a higher correlation with the market volatility⁵.

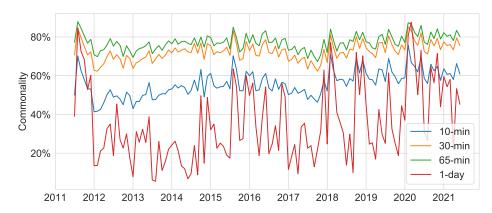


Figure 4: Commonality in realized volatility averaged across stocks for each month during the sample period of 2011-07 \sim 2021-06.

	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

Table 2: Statistics of the monthly average commonality in volatility. VIX represents the market volatility from the Chicago Board Options Exchange.

Figure 5 reports the averaged values and standard deviations (black vertical lines) of commonality for each half-hour in the trading session. To create this figure, we use the observations in a given interval, such as [09:30, 10:00], to fit 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.

 $^{^5\}mathrm{We}$ refer the reader to additional analysis on commonality in Appendix A.

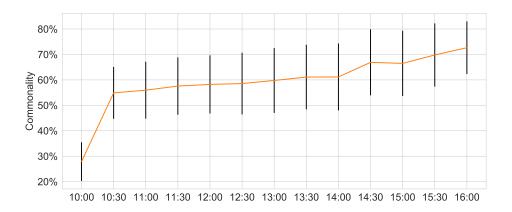


Figure 5: Commonality in realized volatility, averaged across stocks for each half-hour during the sample period of 2011-07 \sim 2021-06.

5 Methodology

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

$$RV_{i,t+h}^{(h)} = F_i\left(\mathbf{x}_{i,:t}^{(h)}, \mathbf{z}_{:t}^{(h)}; \theta\right) + \epsilon_{i,t},\tag{6}$$

where $\mathbf{x}_{i,:t}^{(h)}$ is a multi-dimensional vector of predictor variables for a specific stock i available up to time t, denoted as *individual features*. $\mathbf{z}_{:t}^{(h)}$ is a vector of features for all stocks in the studied universe up to t, denoted as *market features*, such as market volatility. θ 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 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 [20] 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 [54, 42]. For day t, the forecast of HAR is based on

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

$$(7)$$

where $\overline{RV}_{i,t}^{(d)}$ ($\overline{RV}_{i,t}^{(w)}$, $\overline{RV}_{i,t}^{(m)}$) denotes the daily (weekly, monthly) realized volatility in the past day (week, month), 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 [29], 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 [27] to learn the stochastic intraday volatility components.

Following the spirit of Engle and Sokalska [29], we extend the daily HAR model to intraday scenarios by adding diurnal-effect features, as follows⁶

$$RV_{i,t+h}^{(h)} = \alpha_i + \beta_i^{(\tau)} \overline{D}_{i,\tau_{t+h}} + \beta_i^{(d)} \overline{RV}_{i,t}^{(d)} + \beta_i^{(w)} \overline{RV}_{i,t}^{(w)} + \beta_i^{(m)} \overline{RV}_{i,t}^{(m)} + \epsilon_{i,t+h}, \tag{8}$$

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

When we consider the daily scenarios, Eqn (8) becomes the standard HAR model (Eqn (7)), 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. Let $\mathbf{u} = (u_1, \dots, u_p)'$ represent the vector of input features

$$RV_{i,t+h}^{(h)} = \alpha_i + \sum_{l=1}^p \beta_l u_l + \epsilon_{i,t+h}. \tag{9}$$

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. [37], 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 (10). Here, the hyperparameter λ controls the penalty weight.

⁶Since we use the log-version realized volatility, the multiplication of daily, diurnal, and stochastic intraday components in Engle and Sokalska [29] translates to addition in our model (8).

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.

$$\mathcal{L}_{i} = \sum_{t} \left[\text{RV}_{i,t+h}^{(h)} - \alpha_{i} - \sum_{l=1}^{p} \beta_{l} u_{l} \right]^{2} + \lambda \sum_{l=1}^{p} \|\beta_{l}\|_{1}.$$
 (10)

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. One way to add non-linearity and interactions is the decision tree, see more in Hastie et al. [37].

XGBoost is a decision-tree-based ensemble algorithm, implemented under a distributed gradient boosting framework by Chen and Guestrin [16]. 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 [16]. Let **u** represent the vector of input features

$$F_{i}(\mathbf{u}) = \sum_{l=1}^{B} f_{l}(\mathbf{u}), \quad f_{l} \in \mathcal{F},$$
(11)

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 (11) is trained sequentially. Boosting (see Friedman [30]) means that new models are added to minimize the errors made by existing models, until no further improvements are achieved.

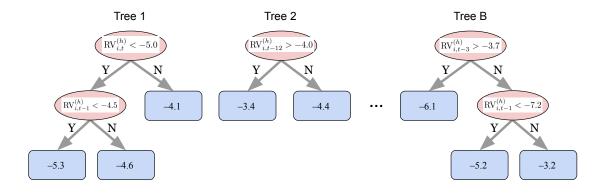


Figure 6: Illustration of a tree ensemble model. B is the number of trees and the final prediction is the sum of predictions from each tree, as shown in Eqn (11).

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 the flexibility to learn complex interactions. However, NNs also suffer from many problems, such as the lack of robustness, transparency, and interpretability, over-parameterization, etc.

MLP is a class of feedforward neural networks and a "universal approximator" that can learn any smooth functions (see Hornik et al. [41]). 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 [45], which is based on adaptive estimates of lower-order moments. Let $\mathbf{u} \in \mathbb{R}^p$ represent the input variables

$$F_i(\mathbf{u}; \theta) = \mathbf{W}_L \cdot \sigma(\mathbf{W}_{L-1} \dots \sigma(\mathbf{W}_1 \mathbf{u} + \mathbf{b}_1) \dots + \mathbf{b}_{L-1}) + \mathbf{b}_L, \tag{12}$$

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)

LSTM, proposed by Hochreiter and Schmidhuber [40], is an artificial recurrent neural network (RNN) architecture, which is well-suited to classifying, processing and making predictions based on time series data. 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 [40].

$$\mathbf{f}_{t} = \sigma_{g} \left(\mathbf{W}_{f} \mathbf{u}_{t} + \mathbf{U}_{f} \mathbf{h}_{t-1} + \mathbf{b}_{f} \right)$$

$$\mathbf{i}_{t} = \sigma_{g} \left(\mathbf{W}_{i} \mathbf{u}_{t} + \mathbf{U}_{i} \mathbf{h}_{t-1} + \mathbf{b}_{i} \right)$$

$$\mathbf{o}_{t} = \sigma_{g} \left(\mathbf{W}_{o} \mathbf{u}_{t} + \mathbf{U}_{o} \mathbf{h}_{t-1} + \mathbf{b}_{o} \right)$$

$$\tilde{\mathbf{c}}_{t} = \sigma_{c} \left(\mathbf{W}_{c} \mathbf{u}_{t} + U_{c} \mathbf{h}_{t-1} + \mathbf{b}_{c} \right)$$

$$\mathbf{c}_{t} = \mathbf{f}_{t} \circ \mathbf{c}_{t-1} + \mathbf{i}_{t} \circ \tilde{\mathbf{c}}_{t}$$

$$\mathbf{h}_{t} = \mathbf{o}_{t} \circ \sigma_{h} \left(\mathbf{c}_{t} \right),$$

$$(13)$$

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. σ_g is sigmoid function, and σ_c , σ_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 [13, 34]. We use a stock's own past RVs only as predictor features, namely $\mathbf{x}_{i,:t}^{(h)} = (\mathrm{RV}_{i,t}^{(h)}, \dots, \mathrm{RV}_{i,t-(p-1)h}^{(h)})'$ 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 (6). As in the Single scheme, we use a stock's own past RVs only as predictor features and no market features. Sirignano and Cont [58] 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. [11], Engle and Sokalska [29] 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. [11]) in order to leverage the commonality in volatility shown in Section 4. Namely, F_i is same for all stocks in Eqn (6). We use both individual features $\mathbf{x}_{i,:t}^{(h)} = (\mathrm{RV}_{i,t}^{(h)}, \ldots, \mathrm{RV}_{i,t-(p-1)h}^{(h)})'$ and market features $\mathbf{z}_{:t}^{(h)} = (\mathrm{RV}_{M,t}^{(h)}, \ldots, \mathrm{RV}_{M,t-(p-1)h}^{(h)})'$ as predictors. Note that for HAR-d models under the Augmented setting, we include aggregated market features as additional features, and use 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 metrics⁷ on the out-of-sample data (see [52, 29, 51, 11, 13, 55]).

- Mean squared error (MSE): $\frac{1}{N} \sum_{i=1}^{N} \frac{1}{\#\mathcal{T}_{test}} \sum_{t \in \mathcal{T}_{test}} \left(\text{RV}_{i,t}^{(h)} \widehat{\text{RV}}_{i,t}^{(h)} \right)^2$,
- Mean absolute percentage error (MAPE): $\frac{1}{N} \sum_{i=1}^{N} \frac{1}{\#\mathcal{T}_{test}} \sum_{t \in \mathcal{T}_{test}} \left| \frac{RV_{i,t}^{(h)} \widehat{RV}_{i,t}^{(h)}}{RV_{i,t}^{(h)}} \right|$,
- R^2 : $\frac{1}{N} \sum_{i=1}^{N} \left[1 \frac{\sum_{t} \left(\text{RV}_{i,t}^{(h)} \widehat{\text{RV}}_{i,t}^{(h)} \right)^2}{\sum_{t} \left(\text{RV}_{i,t}^{(h)} \overline{\text{RV}}_{i}^{(h)} \right)^2} \right].$

 $\widehat{RV}_{i,t}^{(h)}$ represents the predicted value of $RV_{i,t}^{(h)}$, the realized volatility for stock i during (t-h,t]. $\overline{RV}_{i}^{(h)}$ is the empirical mean of $RV_{i,t}^{(h)}$ 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 [26, 25]. 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

$$DM_{12} = \frac{\bar{d}^{(a-b)}}{\hat{\sigma}^{(a-b)}} \to N(0,1),$$
 (14)

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. [31], 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.

$$d_t^{(a-b)} = \frac{1}{N} \sum_{i=1}^{N} \left(L(e_{i,t}^{(a)}) - L(e_{i,t}^{(b)}) \right), \tag{15}$$

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.

⁷Another common measure quasi-likelihood (QLIKE) [53] is not employed in our analysis because we adopt the log volatility.

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 [31]. 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, 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 [33], Gu et al. [31]). Specifically, we train multiple neural networks with different random seeds for initialization, and construct final predictions by averaging the outputs of all networks.

In all of the models, the features are based on the observations in the last 21 days. For example, the number of features used to forecast 30-min RVs is $273 (= 13 \times 21)$. Prior to inputting variables in the models, at each rolling window estimation, we normalize them by removing the mean and scaling to unit variance.

MSE 1.030 1.040 1.013 1.000 1.003 1.004 1.003 1.005				HAR-d			OLS			$_{ m LASSO}$			XGBoost			MLP			$_{ m TSLM}$	
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$			SINGLE	UNIVERSAL	AUGMENTE	D SINGLE	UNIVERSAL	AUGMENTED	SINGLE U	NIVERSAL A	AUGMENTED	SINGLE 1	INIVERSAL	AUGMENTED	٠.		AUGMENTED	SINGLE [JNIVERSAL	AUGMENTED
$ \begin{array}{l l l l l l l l l l l l l l l l l l l $		MSE	1.030	1.040	1.013	1.009	1.008	0.962	1.053	1.012	0.961	1.047	0.968	0.968		0.947	0.945		0.950	0.934
$ \begin{array}{llllllllllllllllllllllllllllllllllll$	10-min	MAPE	_	0.045	0.044	0.044	0.044	0.042	0.045	0.044	0.042	0.045	0.042	0.042		0.042	0.042		0.042	0.041
$ \begin{array}{l l l l l l l l l l l l l l l l l l l $		R^2	0.618	0.612	0.624	0.630	0.626	0.649	0.615	0.624	0.650	0.616	0.646	0.646		0.656	0.657		0.651	0.659
$ \begin{array}{l l l l l l l l l l l l l l l l l l l $		MSE	0.332	0.333	0.328	0.307	0.307	0.293	0.325	0.309	0.293	0.345	0.290	0.297		0.284	0.280		0.287	0.279
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$	30-min	MAPE	$\overline{}$	0.037	0.036	0.035	0.035	0.034	0.036	0.035	0.034	0.037	0.034	0.034		0.034	0.033		0.034	0.033
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$		R^2	0.766	0.765	0.769	0.784	0.784	0.794	0.776	0.782	0.794	0.757	0.796	0.790		0.800	0.803		0.798	0.804
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$		MSE	0.270	0.270	0.265	0.251	0.250	0.241	0.252	0.251	0.242	0.297	0.242	0.249		0.232	0.229		0.232	0.229
0.791 0.794 0.875 0.887 0.881 0.881 0.870 0.813 0.704 0.881 0.881 0.870 0.881 0.881 0.870 0.881 0.881 0.870 0.883 - 0.820 0.206 0.261 0.264 0.264 0.261 0.265 0.261 0.785 0.286 - 0.689 0.048 - 0.048 - 0.045 0.044 - 0.045 0.044 - 0.045 0.045 0.048 - 0.045 0.044 - 0.045 0.044 - 0.045 0.045 0.048 - 0.045 0.044 - 0.045 0.045 0.048 - 0.045 0.045 0.589 0.587 0.612 0.457 0.579 0.579 0.639 0.609 - 0.597	65-min	MAPE		0.037	0.036	0.035	0.035	0.034	0.035	0.035	0.034	0.038	0.034	0.035		0.034	0.033		0.034	0.033
0.260 0.261 0.254 0.263 0.261 0.265 0.368 0.268 0.285 - 0.260 0.267 - 0.261 3.045 0.045 0.046 0.046 0.045 0.044 0.052 0.046 0.048 - 0.045 0.045 0.045 0.589 0.587 0.612 0.47 0.579 0.552 - 0.593 0.597 - 0.597		R^2	0.791	0.791	0.794	0.805	0.807	0.813	0.804	908.0	0.813	0.770	0.812	0.807		0.820	0.823		0.820	0.823
$\begin{array}{cccccccccccccccccccccccccccccccccccc$		MSE	0.260	0.261	0.254	0.263	0.260	0.254	0.263	0.261	0.255	0.358	0.268	0.285		0.260	0.257		0.261	0.258
0.587 0.613 0.589 0.613 0.586 0.587 0.612 0.457 0.579 0.552 - 0.593 0.609 - 0.597	1-day	MAPE	0.045	0.045	0.044	0.046	0.045	0.044	0.046	0.045	0.044	0.052	0.046	0.048		0.045	0.044		0.045	0.044
		R^2	0.589	0.587	0.613	0.583	0.589	0.613	0.586	0.587	0.612	0.457	0.579	0.552	,	0.593	609.0	,	0.597	0.605

Table 3: 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 MSE is highlighted in red (blue), respectively.

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 MSE. We draw the following conclusions from Table 3.

- Regarding HAR-d models, we observe that 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 MSE of
 AUGMENTED compared to SINGLE is 0.017, 0.004, 0.005, 0.006 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, MSEs are reduced from 1.013 (resp. 0.328, 0.265) with the best HAR-d model (i.e. under Augmented) to 0.962 (resp. 0.293, 0.241) with the best OLS model (i.e. under Augmented), across the three horizons, {10, 30, 65} minutes, respectively. Within the OLS models, conclusions are similar with HAR-d models, i.e. no benefits from Universal while significant benefits from Augmented.
- We observe similar findings in LASSO as in OLS, suggesting that regularization does not further aid performance.
- XGBoost slightly underperforms linear regressions, possibly due to overfitting. The best out-of-sample
 performance of XGBoost is achieved under the UNIVERSAL setting. Incorporating market volatility
 does not provide additional predictive power.
- MLPs and LSTMs achieve state-of-the-art accuracy across all measures and intraday horizons, implying the complex interactions between predictors. Further analysis is provided in Section 6.3.
- 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.

Let us now consider the OLS model as an illustrative example for understanding the relative reduction in error. We compare its mean squared errors under these three schemes, at a monthly level, as shown in Figure 7. For better readability, we report the reduction in error of UNIVERSAL relative to SINGLE (denoted as Univ-Single), the reduction of AUGMENTED relative to UNIVERSAL (denoted as Aug-Univ),

and the reduction of **Augmented** relative to **Single** (denoted as Aug-Single). Note that Aug-Single = (Aug-Univ) + (Univ-Single). Negative values of ΔMSE indicate an improvement on out-of-sample data, and positive values indicate degradation. To arrive at this figure, we average the ΔMSE 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.

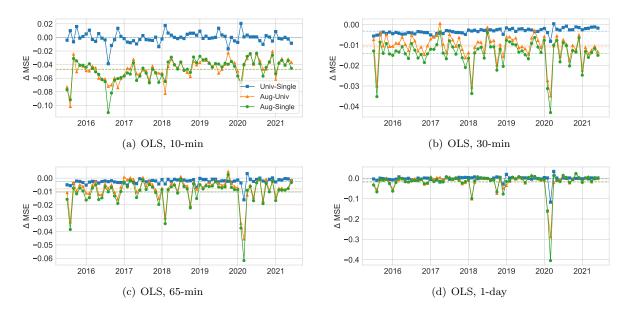


Figure 7: Pairwise Δ MSE of the OLS model, across three training schemes, averaged across stocks in each month during the testing period 2015-07 \sim 2021-06. The dashed horizontal lines represent the average reductions in MSE.

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 MSE relative to **Universal** is explained by commonality to a large extent. Generally, the out-of-sample MSE 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 MSE more for stocks (in the Q1 bucket) that are most loosely connected to the market in terms of 10-min volatility.

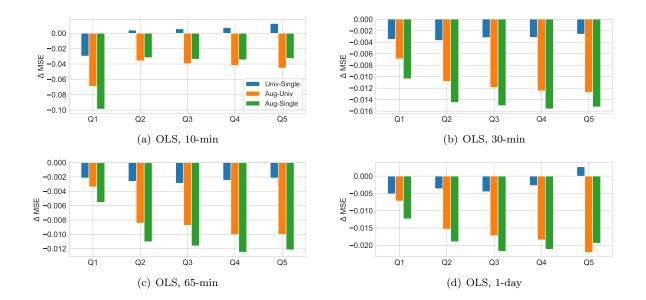


Figure 8: Results showing the pairwise Δ MSE of the OLS model between three training schemes, sorted by commonality. Q1, respectively Q5, denote the subset of stocks with the lowest, respectively highest, 20% values for the commonality.

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. [56]) 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

Sensitivity_k =
$$\sum_{i=1}^{N} \sum_{t \in \mathcal{T}_{train}} \left\| \frac{\partial F}{\partial u_k} \right|_{\mathbf{u} = \mathbf{u}_{i,t}} \right\|_{1}$$
. (16)

Here, F is the fitted model under the **Augmented** scheme, **u** represents the vector of predictors and u_k is the k-th element in **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 had, for OLS, the sensitivities of lag=1 individual (resp. market) RV are 0.081 (resp. 0.069).

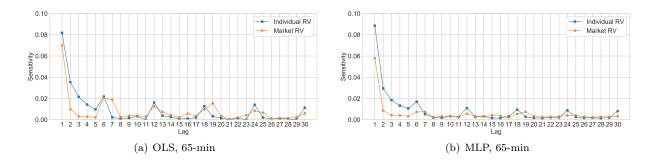


Figure 9: Relative importance of lagged individual and market RVs. For ease of readability, we only report the sensitivity values for the most recent 30 lagged RVs.

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. [31], Choi et al. [17]) that focuses on the partial relations between a pair of input variables and responses, while fixing other variables at their mean values

$$F_{i|i}(u_i \mid u_i = q) = F(u_i, u_i = q, u_k = \bar{u}_k, k \neq i, j),$$
 (17)

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 support⁸. 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 9(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 9(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 9(b), the distances between the curves become relatively smaller, conveying the message that, when an individual

⁸Recall that the variables are normalized by removing the mean and scaling to unit variance.

stock is very volatile, the market effect on it weakens.

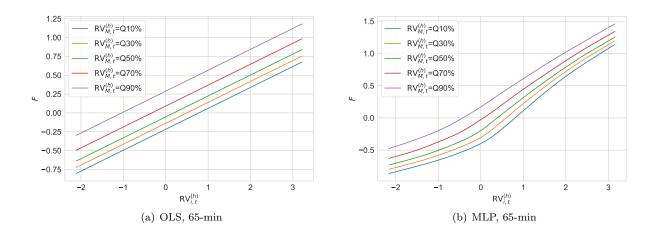


Figure 10: Predicted RV (y-axis) as a function of the lag=1 individual RV (x-axis) conditioned on various lag=1 market RV quantile values (keeping all other variables at their mean values).

6.4 Forecasting RVs of unseen stocks

To examine the ability to generalize and address concerns about overfitting, we perform a stringent outof-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 [58], Choi et al. [17]. 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**⁹. 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 (MSE=0.369) to the best non-neural network model (MSE=0.367, attained by LASSO).

⁹The 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.

		OLS	0	LS	LA	SSO	XGI	Boost	M	LP	LS	ТМ
		Unseen	UNIVERSAL	AUGMENTED								
	MSE	0.660	0.675	0.636	0.680	0.635	0.652	0.651	0.620	0.619	0.634	0.619
10-min	MAPE	0.046	0.046	0.045	0.047	0.045	0.045	0.045	0.044	0.044	0.045	0.044
	R^2	0.643	0.633	0.655	0.630	0.656	0.646	0.646	0.664	0.664	0.656	0.664
	MSE	0.335	0.333	0.322	0.335	0.322	0.318	0.325	0.311	0.306	0.316	0.308
30-min	MAPE	0.039	0.038	0.037	0.039	0.037	0.037	0.038	0.037	0.037	0.037	0.037
	R^2	0.768	0.769	0.777	0.768	0.777	0.780	0.775	0.785	0.788	0.781	0.787
	MSE	0.299	0.297	0.288	0.298	0.288	0.289	0.293	0.276	0.273	0.277	0.273
65-min	MAPE	0.039	0.039	0.038	0.039	0.038	0.038	0.039	0.038	0.037	0.038	0.037
	R^2	0.786	0.787	0.794	0.787	0.793	0.793	0.790	0.802	0.804	0.802	0.804
	MSE	0.391	0.386	0.367	0.387	0.367	0.402	0.415	0.382	0.369	0.379	0.371
1-day	MAPE	0.056	0.055	0.054	0.055	0.054	0.056	0.057	0.055	0.054	0.055	0.054
	R^2	0.467	0.474	0.499	0.473	0.500	0.457	0.439	0.478	0.496	0.480	0.492

Table 4: Results for predicting future realized volatility of unseen stocks over multiple horizons using different models under three training schemes. The column OLS Unseen represents the baseline results based on OLS models estimated for each unseen stock. Other columns 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 MSE 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 (e.g. ARFIMA [4], HAR [20], SHAR [54], HARQ [12]). 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 [4, 9, 42]). It is worth noting that in the traditional forecasting daily RV models, 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.

7.2 Benchmark models

In this section, we introduce a set of commonly used models, where the daily variables (such RV, semi-RV [54]) are employed as predictors. For simplicity, we refer to these models as **traditional** approaches.

OLS. The first benchmark explored is the OLS model with different lengths of lagged daily RVs, as shown in Eqn (9). Two specifications of OLS are considered: (1) OLS(1d) denotes only using the lag one RV as the predictor; (2) OLS(21d) denotes using RVs in the last month as predictors.

HAR. See 5.1.1.

SHAR. Recently, Patton and Sheppard [54] 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 [8] by decomposing the total RV of the first lag via signed intraday returns, as shown in Eqn (18). In other words, the lag one RV in SHAR (Eqn (19)) is split into the sum of squared positive returns and the sum of squared negative returns, as follows.

$$RV_{t}^{(d)+} = \sum_{i=0}^{M-1} r_{t-i\cdot\Delta}^{2} I_{\{r_{t-i\cdot\Delta}>0\}}$$

$$RV_{t}^{(d)-} = \sum_{i=0}^{M-1} r_{t-i\cdot\Delta}^{2} I_{\{r_{t-i\cdot\Delta}<0\}},$$
(18)

$$RV_{i,t+1}^{(d)} = \alpha_i + \beta_i^{(d)+} \overline{RV}_{i,t}^{(d)+} + \beta_i^{(d)-} \overline{RV}_{i,t}^{(d)-} + \beta_i^{(w)} \overline{RV}_{i,w}^{(w)} + \beta_i^{(m)} \overline{RV}_{i,m}^{(m)} + \epsilon_{i,t+1}.$$
(19)

Recall that $\overline{\mathrm{RV}}_{i,t}^{(w)}$ (resp. $\overline{\mathrm{RV}}_{i,t}^{(m)}$) denote the aggregated weekly (resp, monthly) realized volatility.

HARQ. Bollerslev et al. [12] pointed out that the beta coefficients in the HAR model can be affected by measurement errors in the realized volatilities. By exploiting the asymptotic theory for high-frequency realized volatility estimation, the authors propose an easy-to-implement model, termed as HARQ (Eqn (21)). The realized quarticity (RQ) is estimated according to Eqn (20), aiming to correct the measurement errors.

$$RQ_t^{(d)} = \frac{M}{3} \sum_{i=0}^{M-1} r_{t-i \cdot \Delta}^4$$
 (20)

$$RV_{t,t+b} = \alpha + \left(\beta^{(d)} + \beta^{(d)Q}\sqrt{RQ_t^{(d)}}\right)RV_t^{(d)} + \beta^{(w)}RV_t^{(w)} + \beta^{(m)}RV_t^{(m)} + \epsilon_{t+b}.$$
 (21)

7.3 Proposed approach

Previous sections 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.

$$RV_{i,t+1}^{(d)} = F_i\left(\mathbf{x}_{i,t}^{(h)}, \mathbf{z}_t^{(h)}, \mathbf{x}_{i,t-1}^{(d)}, \mathbf{z}_{t-1}^{(d)}; \theta\right) + \epsilon_{i,t}.$$
 (22)

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 intraday predictors for stock i computed at day t and $\mathbf{z}_{t}^{(h)}$ is a vector of intraday predictors for all stocks at day t.

 $\mathbf{x}_{i,:t}^{(d)}$ denotes daily features for a specific stock i available up to day t-1 and $\mathbf{z}_{:t}^{(h)}$ 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 to sub-sampled RVs, $\mathbf{x}_{i,t}^{(h)}$, computed for intervals of length h. Figure 11 illustrates the comparison between the traditional approaches and our **Intraday2Daily** approach.

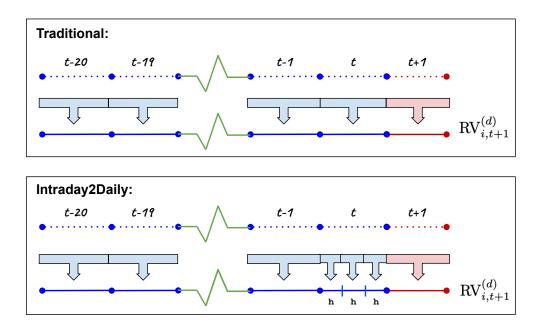


Figure 11: Illustration of two prediction approaches for future daily volatility (red segment). In each box, dots in the top line represent the intraday returns. The traditional approaches employ the aggregated daily (or weekly, or monthly) RVs (long blue segments) as predictors, while the Intraday2Daily 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 investigate the predictive power of intraday RVs on daily volatility.

7.4 Experiments

The forecasting performance of benchmark models with daily variables are summarized in Table 5. From this table, we find that the SHAR model generally performs as well as the standard HAR model, in line with Bollerslev et al. [12]. HARQ outperforms other commonly used models, including HAR and SHAR, when applied to individual stocks studied in the present paper.

Table 6 reports the results of models combined with the Intraday2Daily approach¹⁰. In other words, models in Table 6 use sub-sampled intraday RVs rather than the lag-one total RV. For example, regarding HARQ (Eqn (7)) in Panel A of Table 6, the lag-one total RV is replaced by non-overlapped intraday RVs. For other machine learning models in Panel B of Table 6, intraday RVs in the last day and daily RVs in the previous month are input as predictors.

- By comparing Table 5 with Panel A of Table 6, 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(1d) using daily RVs (MSE=0.284), 65-min RVs (MSE=0.267) improve the out-of-sample performance.
- OLS(21d) (resp. HARQ) is the best (resp. second best) benchmark model when combined with the Intraday2Daily approach.
- MLPs with intraday RVs again achieve the best out-of-sample performance. For example, the MSEs of MLPs under Universal are 0.243, 0.242, 0.246 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.

¹⁰We observe similar findings when applying the Single2Daily approach to forecast the raw volatilities (not in logs).

Feature			OLS(1d)			OLS(21d)			HAR			SHAR			HARQ	
frequency	S	INGLE U	NIVERSAL	SINGLE UNIVERSAL AUGMENTED	SINGLE L	INIVERSAL A	VUGMENTED	SINGLE U	JNIVERSAL .	AUGMENTED	SINGLE [JNIVERSAL	AUGMENTED	SINGLE U	'NIVERSAL A	UGMENTED
	MSE 0	0.284	0.283	0.280	0.263	0.260	0.254	0.261	0.261	0.254	0.261	0.260	0.254	0.256	0.257	0.253
1-day	MAPE 0.047		0.047	0.047	0.046	0.045	0.044	0.045	0.045	0.044	0.045	0.045	0.044	0.044	0.044	0.044
	R^2 0		0.559	0.565	0.598	0.601	0.613	0.600	0.601	0.613	0.600	0.601	0.613	0.609	809.0	0.617

Table 5: Results of forecasting future daily RVs based on various benchmark models with daily variables. Note the results of HAR and OLS(21d) are the same as those of HAR-d and OLS at 1-day horizon in Table 3, respectively. The benchmark model with the best out-of-sample performance in MSE is highlighted in green.

Panel A: Benchmark models combined with Intraday2Daily.

Feature			OLS(1d)			OLS(21d)			HAR			SHAR			HARQ	
frequency		SINGLE	UNIVERSAL	SINGLE UNIVERSAL AUGMENTED SINGLE	SINGLE 1	UNIVERSAL A	AUGMENTED	SINGLE L	JNIVERSAL A	UNIVERSAL AUGMENTED	SINGLE L	UNIVERSAL	L AUGMENTED	SINGLE [JNIVERSAL	JNIVERSAL AUGMENTED
	MSE	0.284	0.314	0.300	0.255	0.253	0.249	0.259	0.270	0.256	0.277	0.285	0.261	0.264	0.253	0.251
10-min	MAPE	0.047	0.051	0.049	0.044	0.044	0.043	0.044	0.045	0.044	0.046	0.047	0.045	0.044	0.044	0.044
	R^2	0.557	0.503	0.528	0.611	0.613	0.621	0.603	0.584	0.609	0.572	0.557	0.601	0.597	0.615	0.620
	MSE	0.268	0.272	0.271	0.252	0.252	0.248	0.252	0.255	0.249	0.257	0.263	0.253	0.254	0.253	0.248
30-min	MAPE	0.046	0.046	0.046	0.044	0.044	0.043	0.044	0.044	0.043	0.044	0.045	0.044	0.044	0.044	0.043
	R^2	0.587	0.579	0.580	0.616	0.616	0.624	0.616	0.611	0.621	0.609	0.598	0.612	0.616	0.615	0.623
	MSE	0.267	0.270	0.270	0.253	0.252	0.249	0.252	0.253	0.249	0.253	0.255	0.250	0.253	0.254	0.250
65-min	MAPE	0.046	0.046	0.046	0.044	0.044	0.043	0.044	0.044	0.043	0.044	0.044	0.043	0.044	0.044	0.043
	R^2	0.589	0.583	0.583	0.616	0.616	0.622	0.617	0.615	0.621	0.616	0.612	0.621	0.617	0.615	0.621

Panel B: Other machine learning models combined with Intraday2Daily.

	LED	İ								
	AUGMEN	0.258	0.044	0.607	0.249	0.044	0.622	0.250	0.044	0.621
$_{ m LSTM}$	SINGLE UNIVERSAL AUGMENTEI	0.247	0.043	0.625	0.244	0.043	0.629	0.244	0.043	0.628
	SINGLE		,							1
	AUGMENTEL	0.247	0.044	0.622	0.246	0.043	0.622	0.247	0.043	0.626
MLP	UNIVERSAL	0.243	0.043	0.627	0.242	0.043	0.628	0.246	0.043	0.627
	SINGLE		1	,		,	,		,	,
	SINGLE UNIVERSAL AUGMENTED SINGLE UNIVERSAL AUGMENTED	0.264	0.045	0.598	0.261	0.045	0.603	0.266	0.045	0.596
XGBoost	UNIVERSAL	0.261	0.045	0.604	0.257	0.044	0.610	0.257	0.044	0.610
	SINGLE	0.323	0.050	0.510	0.330	0.050	0.502	0.332	0.050	0.497
	SINGLE UNIVERSAL AUGMENTED	0.273	0.046	0.570	0.247	0.043	0.624	0.247	0.043	0.623
$_{ m LASSO}$	UNIVERSAL	0.261	0.045	0.589	0.248	0.043	0.622	0.248	0.043	0.623
	SINGLE	0.262	0.045	0.587	0.251	0.044	$R^2 = 0.614$	0.253	0.044	0.616
		MSE	MAPE	R^2	MSE	MAPE	R^2	MSE	MAPE	R^2
Feature	frequency		10-min			30-min			65-min	

Table 6: Results of forecasting daily RVs based on various machine learning models with intraday variables. Note the first column represents 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 MSE is highlighted in red (blue), respectively, for each panel.

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 day¹¹. 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.

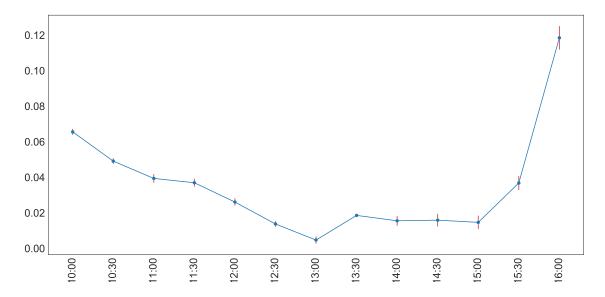


Figure 12: Coefficients of the OLS model using lagged individual 30-min RVs to forecast the next day's volatility. The x-axis represents the time of day. The y-axis represents the coefficients of lagged RVs.

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 [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.

¹¹We 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. The 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., B. Hood, J. Huss, and L. H. Pedersen (2018). Risk everywhere: Modeling and managing volatility. *The Review of Financial Studies* 31(7), 2729–2773.
- [12] 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.
- [13] Bucci, A. (2020). Realized volatility forecasting with neural networks. *Journal of Financial Econometrics* 18(3), 502–531.
- [14] Calvet, L. E., A. J. Fisher, and S. B. Thompson (2006). Volatility comovement: a multifrequency approach. *Journal of econometrics* 131(1-2), 179–215.
- [15] Carroll, C. D., J. C. Fuhrer, and D. W. Wilcox (1994). Does consumer sentiment forecast household spending? if so, why? The American Economic Review 84(5), 1397–1408.
- [16] 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.
- [17] Choi, D., W. Jiang, and C. Zhang (2021). Alpha go everywhere: Machine learning and international stock returns. Available at SSRN 3489679.
- [18] Chordia, T., R. Roll, and A. Subrahmanyam (2000). Commonality in liquidity. *Journal of financial economics* 56(1), 3–28.

- [19] 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.
- [20] Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics* 7(2), 174–196.
- [21] Da, Z., J. Engelberg, and P. Gao (2011). In search of attention. The journal of finance 66(5), 1461–1499.
- [22] Da, Z., J. Engelberg, and P. Gao (2015). The sum of all fears investor sentiment and asset prices. *The Review of Financial Studies* 28(1), 1–32.
- [23] Dang, T. L., F. Moshirian, and B. Zhang (2015). Commonality in news around the world. *Journal of Financial Economics* 116(1), 82–110.
- [24] 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.
- [25] 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.
- [26] Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics* 13(3), 253–263.
- [27] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econometrica: Journal of the econometric society, 987–1007.
- [28] 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.
- [29] 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.
- [30] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189–1232.
- [31] Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies* 33(5), 2223–2273.
- [32] Hameed, A., W. Kang, and S. Viswanathan (2010). Stock market declines and liquidity. *The Journal of finance* 65(1), 257–293.

- [33] Hansen, L. K. and P. Salamon (1990). Neural network ensembles. *IEEE transactions on pattern analysis* and machine intelligence 12(10), 993–1001.
- [34] 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.
- [35] Hansen, P. R. and A. Lunde (2006). Realized variance and market microstructure noise. *Journal of Business & Economic Statistics* 24(2), 127–161.
- [36] Harris, L. (1986). A transaction data study of weekly and intradaily patterns in stock returns. *Journal of financial economics* 16(1), 99–117.
- [37] Hastie, T., R. Tibshirani, and J. Friedman (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
- [38] 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.
- [39] Herskovic, B., B. Kelly, H. Lustig, and S. Van Nieuwerburgh (2020). Firm volatility in granular networks. *Journal of Political Economy* 128 (11), 4097–4162.
- [40] Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. Neural computation 9(8), 1735–1780.
- [41] Hornik, K., M. Stinchcombe, and H. White (1989). Multilayer feedforward networks are universal approximators. *Neural networks* 2(5), 359–366.
- [42] 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.
- [43] 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.
- [44] Keynes, J. M. (2018). The general theory of employment, interest, and money. Springer.
- [45] Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
- [46] Kogan, L., S. A. Ross, J. Wang, and M. M. Westerfield (2006). The price impact and survival of irrational traders. *The Journal of Finance* 61(1), 195–229.

- [47] Lemmon, M. and E. Portniaguina (2006). Consumer confidence and asset prices: Some empirical evidence. *The Review of Financial Studies* 19(4), 1499–1529.
- [48] Li, S. Z. and Y. Tang (2020). Forecasting realized volatility: An automatic system using many features and many machine learning algorithms. *Available at SSRN*.
- [49] 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.
- [50] 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.
- [51] Pascalau, R. and R. Poirier (2021). Increasing the information content of realized volatility forecasts. Journal of Financial Econometrics.
- [52] Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics* 160(1), 246–256.
- [53] Patton, A. J. and K. Sheppard (2009). Evaluating volatility and correlation forecasts. In *Handbook of financial time series*, pp. 801–838. Springer.
- [54] 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.
- [55] Rahimikia, E. and S.-H. Poon (2020). Machine learning for realised volatility forecasting. Available at SSRN 3707796.
- [56] Sadhwani, A., K. Giesecke, and J. Sirignano (2021). Deep learning for mortgage risk. Journal of Financial Econometrics 19(2), 313–368.
- [57] Shleifer, A. and L. H. Summers (1990). The noise trader approach to finance. Journal of Economic perspectives 4(2), 19–33.
- [58] Sirignano, J. and R. Cont (2019). Universal features of price formation in financial markets: perspectives from deep learning. *Quantitative Finance* 19(9), 1449–1459.
- [59] 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.
- [60] Xiong, R., E. P. Nichols, and Y. Shen (2015). Deep learning stock volatility with google domestic trends. arXiv preprint arXiv:1512.04916.

A What may drive commonality in volatility?

Previous studies, especially in the behavioural finance field, have shown that investor sentiments could affect stock prices [5, 46, 21, 11, 22, 43, 32]. Keynes [44] argued that animal spirits affect consumer confidence, thereby moving prices in times of high levels of uncertainty. De Long et al. [24], Shleifer and Summers [57], Kogan et al. [46] found that investor sentiments induce excess volatility. Karolyi et al. [43] considered the investor sentiment index as an important source of commonality in liquidity. Bollerslev et al. [11] 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 [22], 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)¹² by the University of Michigan's Survey Research Center as a proxy for survey-based indices (see Carroll et al. [15], Lemmon and Portniaguina [47]). 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. [22], we also include a news-based index EPU¹³ proposed by Baker et al. [6] to measure policy-related economic uncertainty.

As suggested by Morck et al. [50], the raw monthly commonality measures $R_{(h),m}^2$ (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 [50, 43, 23], we take the logistic transformation of $R_{(h),m}^2$, i.e. $\log \left[R_{(h),m}^2 / (1 - R_{(h),m}^2) \right]$, denoted by $(R_{(h),m}^2)_L$, in our following empirical analysis. To explain the commonality in volatility, we regress $(R_{(h),m}^2)_L$ against the aforementioned three indices, as shown in Eqn (23)

$$(R_{(h),m}^2)_L = \alpha + \beta_1 \text{CSI}_m + \beta_2 \text{VIX}_m + \beta_3 \text{EPU}_m + \epsilon_{i,t}.$$
(23)

Table 7 reports the estimation results¹⁴. 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

¹²http://www.sca.isr.umich.edu

 $^{^{13} {\}rm https://www.policyuncertainty.com}$

¹⁴To compare the effects of various investor sentiments, we normalize the three factors by removing the mean and scaling to unit variance.

changes in adjusted R^2 are subtle.

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

Table 7: Results of time series regressions of average commonality in volatility $(R_{(h),m}^2)_L$ against three sentiment measures, VIX, CSI, and EPU. Superscript * denotes the significance levels of 5%.

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 λ_1 . Hyperparameters for other models are summarized as follows.

	XGBoost	MLP	LSTM
Learning rate	0.1	0.001	0.001
Early stopping rounds	10	10	10
Ensemble	2000	10	10
Max depth	10	-	-
Batch size	-	1024	1024
Epoches	-	100	100
No. of hidden layers	-	3	2
Batch normalization	-	✓	X

Table 8: Hyperparameters in XGBoost, MLP, LSTM.

C Diebold-Mariano test

Table 9: Statistics of Diebold-Mariano tests comparing the out-of-sample performance of machine learning models. 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%.

Panel A: 10-min.

Univ Univ	HAR-d	OLS	LASSO	XGBoost	MLP	LSTM	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	HAR-d	OLS	LASSO	XGBoost	MLP	LSTM	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	HAR-d	OLS	LASSO	XGBoost	MLP	LSTM	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	HAR-d	OLS	LASSO	XGBoost	MLP	LSTM	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*