Journal of Intelligent Decision Making and Information Science

www.jidmis.org
eISSN: 3079-0875



The Role of VIX In Neural Network–Based Short-Horizon Volatility Forecasting: Evidence from Global Equity Markets

Orhan Özaydın[1,*], Anton Abdulbasah Kamil[2]

1 Department of Economy and Finance, Business Faculty, Istanbul Medipol University, Istanbul, Türkiye; Email: orhan.ozaydin@medipol.edu.tr; ORCID: https://orcid.org/0000-0003-2585-1437

2 Department of Business, Business and Administration Faculty, Istanbul Gelisim University, Istanbul, Türkiye; Email: akamil@gelisim.edu.tr; ORCID: https://orcid.org/0000-0001-5410-812X

ARTICLE INFO

Article history:
Received 15 April 2026
Received in revised form 25 May 2026
Accepted 15 June 2026

Keywords: Volatility forecasting; VIX; Neural networks; Rolling out-of-sample evaluation; Global equity markets.

ABSTRACT

Volatility forecasts play an important role in financial markets. They are used for setting position limits, determining margin requirements, and guiding short-term trading decisions. Financial volatility is known to exhibit persistence and regime dependence, which makes short-horizon forecasting challenging. In addition to historical volatility measures, implied volatility indicators such as the Volatility Index (VIX) are often used as forward-looking proxies for market risk expectations. However, it is not clear whether the predictive usefulness of the VIX is consistent across different neural network architectures. This study investigates the role of the VIX in one-day-ahead realized volatility forecasting and examines whether its contribution varies across alternative neural network models. The empirical analysis focuses on major global equity markets, including the S&P 500, DAX, FTSE 100, Nikkei 225, and Hang Seng Index. Using market data from these indices, the study evaluates the forecasting performance of artificial neural networks (ANN), long short-term memory networks (LSTM), and gated recurrent unit (GRU) models. All models are estimated within a rolling out-of-sample framework and are tested with and without the VIX variable. Forecast accuracy is evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Quasi-Likelihood (QLIKE) loss functions, and Diebold–Mariano tests. The results show that memory-based models outperform feedforward ANN models, while the contribution of the VIX varies across models and markets.

1. Introduction

Short-term volatility in financial markets exhibits distinct clustering, regime dependence, and nonlinear structures. This aspect of volatility has been extensively studied in literature and poses significant challenges for traditional forecasting models (Engle 1982; Bollerslev 1986; Cont 2001). Accurate short-term volatility forecasts are of central importance for risk management, derivative pricing, and capital allocation decisions. Volatility measures derived from options, such as the VIX index, incorporate both expected variance and risk premium components. With these characteristics, VIX indices provide forward-looking information about market expectations regarding future risk (Carr and Wu 2009). In this case, forward-looking forecasts can be made using the interaction between realized volatility dynamics and predefined risk expectations. This raises important

questions about the extent to which the information provided by forecasting models meets expectations.

There is sufficient evidence suggesting that short-run volatility clustering, regime dependence and nonlinearity are key features of many financial time series (Engle 1982; Bollerslev 1986; Cont 2001). Although long-term modeling of volatility is important for theoretical purposes, short-term ahead (days) accurate modeling of volatility is important for a number of applications, ranging from risk management to derivatives pricing and capital allocation. The volatility of a derivative extracted from options data is given by the volatility implied from options prices and has two components: the expected volatility and the risk premium or volatility risk premium (Carr and Wu 2009). So VIX type indices provide the user with a snapshot of the market's expectations regarding future risk. Therefore, for this type of data, we can contemplate forecasts that are the intersection of the realized volatility evolution and the predefined risk expectations that the model is designed to meet.

Neural network (NN) architectures have been increasingly used in volatility forecasting in recent years due to their ability to approximate complex nonlinear functions and uncover hidden data structures. Despite growing interest in deep learning applications, empirical evidence on how different ANN architectures process forward-looking risk information and whether this information leads to statistically significant improvements in out-of-sample prediction performance remains limited. Existing studies often support volatility models with external risk indicators such as the VIX. However, it is rarely distinguished whether the observed performance gains stem from the information content of the risk indicator or from the model architecture's capacity to internalize and utilize this information. Consequently, in this context, the role of model architecture design, alongside the risk indicator (VIX), in extracting forward-looking risk signals remains insufficiently understood.

In this study, we compare the short-term (one-day-ahead) volatility prediction performance of different neural network architectures. Additionally, in this research, we examine the extent to which these architectures can integrate forward-looking risk information (VIX) and how they can benefit from it. To understand the importance of implied volatility in forecasting, we create models with and without VIX. Furthermore, we investigate how different neural network models incorporate information about risk regimes associated with architecture when making forecasts. The VIX index is used as an indicator of market expectations and risk regimes. However, in addition to the predictive power of the VIX alone, the way in which architecture-dependent information extraction mechanisms shape prediction performance is also important. Unlike existing studies that use implied volatility as an additional prediction input, this study distinguishes the interaction between forward-looking risk signals and neural network architecture design and also examines intertemporal volatility dynamics through neural network models. It provides architectural-level evidence of how the model structure mediates the transformation of implied risk expectations into realized volatility predictions.

We investigate whether estimating daily volatility in the components of five developed country high-frequency equity indices—S&P 500, DAX, FTSE 100, Nikkei 225 and Hang Seng Index—is a local or a global activity. We use daily data from these indices, selecting them because they represent the largest and most liquid indices worldwide. We assume that, in such large indices, the local effects are

unlikely to be significant. We focus on the post-COVID-19 data, because the COVID-19 has imposed several large structural changes on the economy and the markets, and it is not feasible to capture these in any reasonable volatility model. Excluding this period provides a homogeneous sample of data that is not affected by large external shocks, which is a better sample for our analysis. In addition, risk management, where volatility estimates are used, is typically a daily activity. Another advantage of using daily data is that it allows us to investigate short-run clustering of volatility in a finer level of detail. While models, such as GARCH, that are designed to model the clustering of volatility require longer windows for their volatility estimates to converge, the fact that we have more observations when we use a smaller window for calculating realised volatility means that the loss of information due to using a shorter window is smaller. We also use the VIX index as an alternative predictor, which is widely used in the literature, because the one-step-ahead forecasts of volatility models are based on the realisations of the variables that occur after the forecast date, while the realisations of realised volatility are based on past realisations, while the VIX is a forward-looking measure of the market's view of volatility.

We examine six neural network specifications in this paper: feedforward ANN, LSTM and GRU and including and excluding the VIX. There are several reasons for doing this. First, comparisons of the performance of different neural network architectures can provide valuable insights into the relative merits of different models. Second, it is interesting to consider the role of the VIX in models of future volatility. Models that do not only include past values of the VIX but also learn to exclude this information in favour of past realizations of the forecasted volatility seem to perform best. Therefore, it is interesting to consider models that include the VIX as well as models that exclude it and see if the former or the latter can learn to do the same thing. Third, including the VIX in conjunction with other neural network architectures allows us to assess the extent to which the VIX and neural network architectures interact in the modelling process. We use a rolling forecasting framework to assess the out-of-sample performance of the six models considered in this paper.

We test three hypotheses. H1 examines whether the variation in the forecast accuracy of different neural network models is statistically and economically significant. H2 examines the role of the VIX and determines if the VIX (which is an index of the market's expectation of future volatility) is significant to one-day-ahead point forecasts, and if so, whether the impact of VIX is different for different network architectures. H3 is based on the fact that the main advantage of using LSTMs and GRUs is to preserve the relevant information from one time period to the next, and we suspect from our model specification that there should be a high degree of significance in modeling the past volatility since current volatility is highly influenced by past period's values.

The remainder of the article is organized as follows. The next section summarizes the relevant literature. The third section introduces the methods and model structures used and explains the dataset and variables. The fourth section presents and discusses empirical findings. The final section evaluates the main results of the study and discusses implications for future research.

2. Literature

The volatility estimation literature provides an important framework for evaluating model performance, particularly in short-term forecasts. The literature section first briefly discusses the fundamental non-linear characteristics of financial volatility. It then explains the distinction between realized volatility and implied volatility and discusses the information content of the CBOE Volatility Index (VIX) that reflects market expectations. Finally, the use of neural network models in volatility forecasting is reviewed, and the limited findings in the literature regarding architectural differences in incorporating forward-looking risk information into the model are noted.

2.1 Stylized Facts of Financial Volatility

The most common feature observed in financial volatility modelling is volatility clustering, where large price movements tend to be followed by large movements, and small movements tend to be followed by small movements. This behaviour implies that conditional variance changes over time and is not constant. Engle (1982) revealed this feature of volatility with the ARCH framework, while Bollerslev (1986) extended it with the GARCH model, demonstrating that volatility persistence is a fundamental feature of financial time series.

Volatility exhibits long-term memory behaviour and strong persistence, in addition to short-term clustering. It shows that the autocorrelation of squared returns gradually decreases over time, and shocks to volatility can have long-term effects (Ding, Granger, and Engle 1993). Estimation models that do not take time dependence into account may carry the risk of structural misspecification. This indicates that volatility is not merely an instantaneous response to individual shocks, but rather a dynamic process that persists over time.

A comprehensive analysis of financial time series reveals some distinct characteristics of volatility. These include the fact that volatility is non-linear, has heavy tails, and is also sensitive to regime shifts (Cont 2001). These characteristics indicate that volatility dynamics are shaped by changing market conditions and information flows. Short-term forecasting models should not rely solely on static functional mappings. Instead, they should consider the temporal structure and nonlinear dynamics. Fundamental empirical findings in the literature indicate that volatility is a persistent and state-dependent process.

2.2 Implied Versus Realized Volatility and the Information Content of VIX

In volatility modelling literature, two key measures stand out. Realized volatility is calculated retrospectively from return data, while implied volatility is derived from option prices. Unlike realized measures, implied volatility is derived under a risk-free measure. Therefore, it reflects not only expectations about future variance but also a variance risk premium embedded in option markets (Carr and Wu 2009). The difference between implied volatility and realized volatility is considered a conceptual indicator reflecting markets' perception of future risk and macro-financial uncertainty. Empirical evidence shows that this difference captures systematic risk compensation (Bollerslev, Tauchen, and Zhou 2009). Implied volatility should be interpreted as a forward-looking risk signal rather than a simple forecast of future volatility.

The VIX index provides a measure of implied volatility based on S&P 500 options and is also referred to as the market's “fear gauge” (Whaley 2000). It is true that the VIX contains information based on predictions about future volatility; however, its informational value is neither fixed nor always superior to historical volatility measures. The effectiveness of the VIX depends on risk conditions and market regimes (Bekaert and Hoerova 2014). This distinction implies that including the VIX in volatility forecasting frameworks does not automatically increase accuracy. Its contribution depends solely on the forecasting model's ability to effectively integrate forward-looking risk expectations with realized volatility dynamics.

2.3 Short-Horizon Volatility Forecasting and Risk Management Applications

In practice, volatility forecasting within institutional trading environments serves operational rather than purely theoretical purposes. It is a critical tool to assist in the management of a portfolio and in the trading of derivatives. Banks need to maintain a measure of risk in their portfolios, and this is usually measured by some form of volatility metric. This measure is forecast in real time and is updated daily. The morning is a key period for the management of risk, where the risk manager will adjust the risk positions of the bank including the allocation of assets within a portfolio, the levels of trading limits and the levels of hedging. The risk manager will be looking to optimize the management of risk for tomorrow based on their forecast of the volatility of assets for tomorrow. Achieving short horizon forecast accuracy is a key requirement in this environment. Under the Market Risk Regulations (Basel Committee on Banking Supervision 2016) the capital that banks require to hold is related to the quality of their short horizon forecasts. Even small improvements to the accuracy of the forecasts of future daily volatilities can have a material impact on the required capital.

In forecasting exercises, the choice of the dependent variable is problem specific. Instead of focusing on price movements, in risk forecasting we focus on the volatility of these movements. We then model the realized volatility using the definition of realized measures as the volatility over a fixed length rolling window of the actual movements in the prices. In line with Andersen et al. (2003) we also believe that realized measures are a better measure of the underlying volatility dynamics. Realized measures of volatility have their own errors which are also an important aspect of the criteria that are used to evaluate the performance of the different volatility forecasting models. Following Patton (2011) — the consistency of the model rankings can then vary depending on the loss function that is used. In this study we compare the performance of six different neural network specifications and hence the choice of criteria is an important part of the analysis. In Bollerslev, Gibson and Zhou (2011) the authors argue that the options-derived variance expectations can contain some information about the risk premium that is not contained in the measures of past realizations of volatility. Hence, we also include the VIX in some of the specifications of the models.

2.4 Neural Network Architectures in Volatility Forecasting

In cases where financial time series do not exhibit linearity, neural network models are emerging as a preferred architecture for forecasting due to their flexibility in modelling non-linear relationships. Feedforward artificial neural networks (ANNs) provide a non-linear mapping between input variables and target outcomes. This mapping allows for approximating a broad range of functional forms

(Hornik 1989). In volatility prediction models, ANN models, which do not impose predefined structural constraints, allow the capture of interactions between lagged volatility measures and external risk indicators.

In sequential data environments, recurrent neural network (RNN) architectures and their gated variants, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), are designed to process temporal information by preserving internal state representations across time steps (Hochreiter and Schmidhuber 1997; Cho et al. 2014). These architectures differ from static feedforward networks in how they handle temporal dependencies because they incorporate mechanisms that dynamically update and preserve information. Financial volatility has a time-dependent and persistent structure. In this context, the application of static and memory-based neural network architectures provides a systematic framework for comparing volatility dynamics and forward-looking risk indicators across different models.

3. Methodology

3.1 Data

The sample period begins on January 1, 2020, in order to capture the post–COVID-19 market regime, during which volatility dynamics are widely acknowledged to have undergone structural changes. The post–COVID-19 period is characterized by heightened volatility clustering, structural breaks, and rapid regime shifts, providing a demanding and informative environment for short-horizon volatility forecasting evaluation.

The equity index data used in this study consist of daily closing prices obtained from the Stooq (https://stooq.com/) database. The selected indices are chosen to reflect the geographical and economic diversity of global financial markets: the S&P 500 Index for the United States, the DAX Index for Germany, the FTSE 100 Index for the United Kingdom, the Nikkei 225 Index for Japan, and the HSI Index for Hong Kong. By covering both developed and emerging markets, this set of indices provides a broad representation of global equity market dynamics.

Forward-looking risk information is proxied by the CBOE Volatility Index (VIX), obtained from the official data releases of the Chicago Board Options Exchange (CBOE) (https://www.cboe.com/tradable-products/vix/vix-historical-data/). The VIX is computed from S&P 500 index option prices and represents a risk-neutral measure of market-implied volatility expectations. In this study, the VIX is incorporated into the forecasting models as an external risk proxy that captures market expectations and risk regimes, rather than as a standalone predictor whose intrinsic forecasting power is of primary interest.

To model volatility dynamics, the analysis is conducted using logarithmic return series constructed from daily closing prices. Log returns provide a scale-free representation of price movements and exhibit more favourable statistical properties than price levels, making them the standard choice in empirical finance. The daily return series is defined as:

$$r_t = \ln(P_t) - \ln(P_{t-1}) \quad \text{(Eq.1)}$$

where P_t denotes the closing price of the financial asset at time t.

3.2 Volatility Measurement

In this study, realized volatility is computed using a 22-day rolling window. This choice is motivated by the convention in financial markets that approximately 22 trading days correspond to one calendar month, making it a standard proxy for monthly volatility in the literature (Barndorff-Nielsen and Shephard 2002; Andersen et al. 2007). As such, the selected window length provides a natural time scale for capturing short-term risk dynamics.

Volatility is an inherently latent process and cannot be directly observed. In empirical applications, it is therefore represented using proxy measures derived from return data. Estimation and forecasting based on return-based volatility measures are consistent with the established volatility forecasting literature (Patton 2011). In financial economics, volatility is typically defined as the conditional variance of asset returns (Andersen et al. 2003).

Given that conditional variance is unobservable, realized volatility (RV) is employed as the target variable in the forecasting exercise. Realized volatility is widely used as an observable proxy for latent volatility and has become a benchmark measure in forecast evaluation and comparison studies (Patton 2011). In this study, 22-day realized volatility is defined as the square root of the corresponding 22-day realized variance:

$$RV_t(22) = \sqrt{\sum_{i=0}^{21} r_{t-i}^2} \qquad \text{(Eq.2)}$$

This definition is consistent with the standard realized volatility framework established in the literature (Andersen et al. 2003). The 22-day window length serves to attenuate high-frequency noise in daily returns while preserving information relevant for short-horizon risk assessment. By smoothing daily volatility over an approximate one-month horizon, this measure provides a more stable representation of short-term risk conditions. The forecasting target is specified as one-day-ahead realized volatility:

$$y_t = RV_{t+1}(22) \qquad \text{(Eq.3)}$$

This setup is designed to generate forward-looking risk signals relevant for practical volatility forecasting applications, particularly in real-time risk management contexts.

3.3 Forecasting Design and Rolling Out-of-Sample Framework

Model performance is evaluated within a rolling out-of-sample (OOS) framework that closely mimics real-time forecasting conditions. Unlike a single fixed training–testing split, the rolling design treats each forecast origin as an independent prediction for an unseen period, thereby systematically eliminating look-ahead bias. At each forecast origin, models are estimated using only the information available up to the prediction date, ensuring that forecasts are genuinely ex ante in nature (Diebold and Mariano 1995).

In practice, a fixed-length estimation window of size W=1000 observations is employed. At each forecast origin, the model is re-estimated using observations contained within this window, after

which a one-step-ahead volatility forecast is generated (Fig.1). The estimation window is then rolled forward by one trading day by discarding the oldest observation and incorporating the most recent one. Importantly, models are re-estimated from scratch at each forecast origin, and no incremental or online learning procedures are used. This design choice ensures that forecast performance reflects the model's ability to adapt to evolving data environments rather than relying on parameter persistence across estimation windows. Rolling OOS evaluation is therefore particularly well suited to financial time series characterized by regime shifts and parameter instability and is widely regarded as a benchmark framework in forecast comparison studies (Diebold and Mariano 1995; Diebold 2015).

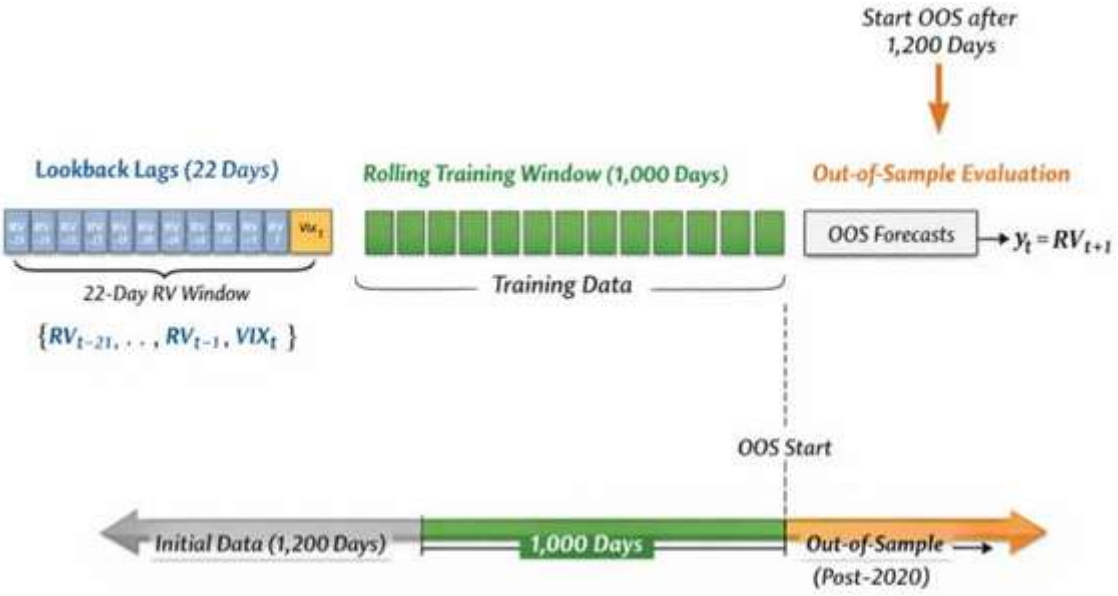


Fig. 1. Forecasting Design and Rolling Out-of-Sample

The empirical analysis begins on January 1, 2020. This starting date is selected to focus on the post–COVID-19 period, during which volatility dynamics are widely acknowledged to have undergone structural changes. Within the rolling forecasting framework, the training window length is fixed at W=1000 observations. To ensure that the first out-of-sample (OOS) forecast is generated using a fully populated estimation window and to mitigate potential instability arising from early-sample dynamics, an additional buffer of 200 observations is excluded. Accordingly, the initial 1200 observations are omitted from OOS evaluation. From this point onward, all feasible forecast steps are retained in the analysis.

All neural network models are trained for a maximum of 10 epochs. To mitigate overfitting, an early stopping procedure is implemented whereby training is terminated if the validation loss fails to improve over two consecutive epochs. The mini-batch size is set to 128 for all model specifications. Forecast accuracy is assessed using the root mean squared error (RMSE) and the QLIKE loss function, which has been shown to deliver more reliable model rankings in volatility forecasting contexts. The statistical significance of differences in forecast performance is evaluated using Diebold–Mariano tests.

All computations and model estimations are conducted in the Google Colab environment using the Python programming language. Neural network models are implemented via the TensorFlow–Keras library, while data processing and analysis rely on NumPy, pandas, and related scientific Python packages. Feedforward artificial neural networks (ANN), long short-term memory (LSTM) networks, and gated recurrent unit (GRU) architectures are estimated using the Keras API. The early stopping mechanism is applied uniformly across all models, and all architectures are compared under identical computational infrastructure and hyperparameter settings.

3.4 Model Architectures and Information Sets

The primary objective of this study is to compare the short-horizon volatility forecasting performance of three neural network architectures—ANN, LSTM, and GRU—and to assess whether the marginal contribution of forward-looking risk information, proxied by the VIX, is conditional on model architecture. All models are estimated using identical information sets, except for the inclusion or exclusion of the VIX variable. This design allows performance differences to be attributed to architectural characteristics rather than differences in data availability.

The 22-day realized volatility window (Equation 3) is used both as the lag length in ANN models and as the lookback window in LSTM and GRU models. In the rolling forecasting procedure, the fixed-length training window contains Wobservations and is re-estimated at each iteration.

3.4.1 Feedforward Artificial Neural Networks (ANN)

Artificial Neural Networks (ANNs), originally conceptualized by Warren McCulloch and Walter Pitts (1943:116), are powerful computational frameworks inspired by the functional mechanisms of biological neural systems. Nevertheless, early ANN structures represent relatively primitive forms of neural network modelling. In particular, standard feedforward architectures lack an internal memory mechanism. Consequently, temporal dependencies cannot be captured intrinsically and must instead be incorporated explicitly through lagged input variables. Within this framework, the specification of the lag structure effectively defines the information set accessible to the model. Although ANNs provide considerable flexibility via nonlinear function approximation, dynamic relationships do not emerge endogenously; rather, they must be imposed through deliberate feature engineering (Bishop 1995).

In this study, the baseline ANN specification includes lagged realized volatility terms as inputs. In the VIX-augmented specification, the contemporaneous VIX value is added to the same set of realized volatility lags. This ablation-style design isolates the effect of expanding the information set while holding the model architecture fixed.

i. Feedforward ANN Model 1: ANN_noVIX

– Target: $y_t = RV_{t+1}(22)$

– Inputs: $x_t = [RV_t, RV_{t-1}, \dots, RV_{t-L+1}]$

– Lag structure: ANN lags tested 22

ii. Feedforward ANN Model 2: ANN_VIX

– Target: $y_t = RV_{t+1}(22)$

– Inputs: $x_t = [RV_t, RV_{t-1}, \dots, RV_{t-L+1}, VIX_t]$

– Lag structure: Same ANN lag candidates
– VIX: VIX_t

3.4.2 Long Short-Term Memory Networks (LSTM)

LSTM networks are recurrent neural network architectures designed to selectively retain and discard past information through gated memory mechanisms, including input, forget, and output gates (Hochreiter and Schmidhuber 1997). These features make LSTM models particularly suitable for learning long-range dependencies in time series data. Given the persistence and clustering commonly observed in volatility series, memory-based architectures may offer empirical advantages in volatility forecasting.

In this study, LSTM models receive sequences of past realized volatility values over a fixed lookback window. The VIX variable is incorporated as a contemporaneous external signal that is held constant across the sequence. This design allows the model to learn the relative importance of different time steps in the realized volatility history while conditioning on forward-looking risk information.

i. LSTM Model 3: LSTM_noVIX
– Target: $y_t = RV_{t+1}(22)$
– Daily feature vector :$z_t = [RV_t]$
– Sequence input: $X_t = [z_{t-L+1}, \dots, z_t]$
– Lookback window: $L = 22$
ii. LSTM Model 4: LSTM_VIX
– Target: $y_t = RV_{t+1}(22)$
– Daily feature vector: $z_t = [RV_t, VIX_t]$
– Sequence input: $X_t = [z_{t-L+1}, \dots, z_t]$
– Lookback window: $L = 22$
– VIX: Included within the sequence (no separate lag choice)

3.4.3 Gated Recurrent Unit Networks (GRU)

GRU networks are closely related to LSTM architectures but employ a more parsimonious gating structure. By combining certain gating mechanisms, GRU models often achieve competitive forecasting performance with lower computational complexity (Cho et al. 2014; Chung et al. 2014). In this study, GRU models serve as a key comparison benchmark to evaluate whether a simpler memory architecture can deliver advantages in short-horizon volatility forecasting.

The GRU specifications mirror those of the LSTM models in terms of input structure and lookback window. As with LSTM, the VIX variable is included as a contemporaneous external signal repeated across the sequence. This setup ensures a fair comparison across architectures and isolates the role of internal memory design in exploiting forward-looking risk information.

i. GRU Model 5: GRU_noVIX
– Target: $y_t = RV_{t+1}(22)$
– Daily feature vector: $z_t = [RV_t]$

– Sequence input: $X_t = [z_{t-L+1}, \dots, z_t]$
– Lookback window: $L = 22$

ii. GRU Model 6: GRU_VIX

– Target:$y_t = RV_{t+1}(22)$
– Daily feature vector: $z_t = [RV_t, VIX_t]$
– Sequence input: $X_t = [z_{t-L+1}, \dots, z_t]$
– Lookback window: $L = 22$
– VIX: Included within the sequence (no separate lag choice)

3.5 Performance Measures and Forecast Evaluation

Forecast performance is evaluated using both scale-dependent and likelihood-based loss functions in order to obtain a comprehensive and robust assessment of volatility prediction accuracy. Relying on a single evaluation criterion may lead to misleading conclusions, particularly in volatility forecasting contexts where loss functions differ in their sensitivity to forecast errors and measurement noise. Accordingly, model comparisons are conducted using the root mean squared error (RMSE) and the QLIKE loss function, both of which are well established in the volatility forecasting literature.

The root mean squared error (RMSE) provides an intuitive, scale-dependent measure of forecast accuracy and captures the average magnitude of deviations between realized and predicted volatility. It is defined as

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2} \qquad \text{(Eq. 4)}$$

where y_t denotes the realized volatility proxy at time t, $\hat{y}_t$ represents the corresponding model forecast, and T is the total number of out-of-sample observations. Lower RMSE values indicate superior predictive performance. While RMSE is straightforward to interpret, it may overweight large forecast errors and is sensitive to the scale of the volatility series.

To complement RMSE, forecast accuracy is also evaluated using the QLIKE loss function, derived from the Gaussian quasi-likelihood framework. The QLIKE specification is particularly suitable for volatility forecasting because it remains robust in the presence of measurement error in realized volatility proxies. Moreover, alternative normalized representations of QLIKE yield identical model rankings, ensuring that results are not sensitive to the specific functional normalization adopted. The QLIKE loss at time t is defined as

$$\text{QLIKE}_t = \log(\hat{y}_t) + \frac{y_t}{\hat{y}_t} \qquad \text{(Eq. 5)}$$

where y_t denotes the realized volatility proxy and $\hat{y}_t$ represents the corresponding model forecast. Lower QLIKE values indicate superior predictive performance. The QLIKE criterion has been shown to produce more reliable rankings of competing volatility models relative to purely error-based measures when the volatility proxy is noisy.

To assess whether differences in forecast performance across models are statistically significant, Diebold–Mariano (DM) tests are employed. The DM test evaluates the null hypothesis of equal predictive accuracy between two competing models based on the time series of loss differentials. In line with the rolling out-of-sample framework described in Section 3.3, DM statistics are computed using the full sequence of available out-of-sample forecasts. This approach allows for formal inference on whether observed performance differences reflect systematic improvements rather than sampling variability.

Model comparisons are conducted on a pairwise basis across all competing specifications using identical evaluation samples and loss definitions. By jointly considering RMSE, QLIKE, and Diebold–Mariano test results, the forecast evaluation framework provides a balanced assessment of predictive accuracy, ranking stability, and statistical significance. This multi-criteria approach mitigates the risk of model selection being driven by a single performance measure and is consistent with best practices in the volatility forecasting literature.

The marginal contribution of incorporating the VIX index is summarized in terms of economic magnitude using differences in forecast loss measures. Specifically, changes in RMSE and QLIKE are computed as

$$\Delta\text{RMSE} = \text{RMSE}_{\text{VIX}} - \text{RMSE}_{\text{noVIX}} \quad \text{(Eq. 6)}$$

$$\Delta\text{QLIKE} = \text{QLIKE}_{\text{VIX}} - \text{QLIKE}_{\text{noVIX}} \quad \text{(Eq. 7)}$$

where RMSE_{VIX} and $\text{QLIKE}_{\text{VIX}}$ denote forecast losses obtained from models that include the VIX index as an explanatory input, while $\text{RMSE}_{\text{noVIX}}$ and $\text{QLIKE}_{\text{noVIX}}$ correspond to otherwise identical model specifications that exclude VIX information. Negative values of ΔRMSEand ΔQLIKEindicate that the inclusion of VIX improves forecast performance, whereas positive values imply a deterioration in predictive accuracy.

4. Results

The main findings of the study indicate that model architecture plays a decisive role in short-horizon volatility forecasting and that the contribution of the VIX index is not universal, but conditional on both model architecture and market characteristics. When out-of-sample (OOS) forecast results are jointly evaluated using error-based loss measures (RMSE and QLIKE) and Diebold–Mariano (DM) tests, the results highlight the critical importance of architecture–data compatibility in modeling financial volatility dynamics (Tables 1, 2 and 3).

4.1 Overall Model Performance

Errors for each model of six and market can be seen from Table 1. As expected, the feedforward ANN model has higher average errors than the LSTM and GRU models for all markets. The ANN's poor performance suggests that the empirically known stabilizing effects of volatility, such as volatility persistence and clustering, are not preserved in the ANN's static functional form. The non-linear nature of the ANN allows it to capture the complex relationships between the variables, but it treats the volatility as a static contemporaneous mapping. Hence, the ANN can be viewed more as a maintained hypothesis for short-run volatility forecasting rather than a serious contender.

Table 1 Out-of-sample forecast accuracy across markets

Market	Model	RMSE	QLIKE
HSI	LSTM_noVIX	0.000851*	-7.973514**
HSI	GRU_noVIX	0.000863	-7.972964
HSI	GRU_VIX	0.000910	-7.971928
HSI	LSTM_VIX	0.000923	-7.971647
HSI	ANN_noVIX	0.001043	-7.965944
HSI	ANN_VIX	0.001081	-7.965029
NIKKEI	GRU_noVIX	0.001090*	-7.708016**
NIKKEI	GRU_VIX	0.001068	-7.706432
NIKKEI	LSTM_noVIX	0.001154	-7.704035
NIKKEI	LSTM_VIX	0.001139	-7.703665
NIKKEI	ANN_noVIX	0.001510	-7.684550
NIKKEI	ANN_VIX	0.001668	-7.674502
FTSE	GRU_VIX	0.000553*	-9.096941**
FTSE	GRU_noVIX	0.000556	-9.096352
FTSE	LSTM_noVIX	0.000634	-9.092369
FTSE	LSTM_VIX	0.000648	-9.091937
FTSE	ANN_VIX	0.000721	-9.089061
FTSE	ANN_noVIX	0.000775	-9.085845
DAX	GRU_noVIX	0.000678*	-8.517756**
DAX	LSTM_noVIX	0.000702	-8.516402
DAX	LSTM_VIX	0.000751	-8.514982
DAX	GRU_VIX	0.000754	-8.514977
DAX	ANN_VIX	0.000849	-8.511958
DAX	ANN_noVIX	0.000868	-8.510126
SP500	GRU_VIX	0.000844*	-8.911962**
SP500	GRU_noVIX	0.000978	-8.902167
SP500	LSTM_VIX	0.001019	-8.898648
SP500	LSTM_noVIX	0.001000	-8.898220
SP500	ANN_VIX	0.001045	-8.895793
SP500	ANN_noVIX	0.001050	-8.890080

*Best RMSE performance for each market; ***Best QLIKE performance for each market.*

The RMSE and the QLIKE for LSTM are systematically lower than for ANN as shown in Table 1. This reflects the capacity of the LSTM to capture the persistence of the volatility in the latent state of the cell and of the gates. Adding VIX to the LSTM, as shown in Table 2, does not bring any consistent and economically significant improvement across markets. The Diebold–Mariano test between the two types of models shows that the differences between the forecasts are generally not statistically different. This suggests that the LSTMs are able to extract most of the information contained in the

past realized volatilities, implying that the VIX and similar variables measuring future volatility do not bring much predictive content.

For all the other markets we see that the GRU model generally performs best and gives the smallest forecast errors (Table 1). In the high liquid and deep markets of the FTSE and the S&P 500 indices we find that the RMSE and the QLIKE value are lower for the GRU with VIX model than for the GRU w/o VIX model. GRU has statically significant values for S&P500 and DAX (Table 2). Hence, even though the GRU model has a less complex gate structure, it seems that it is better able to adapt to shifts in the volatility regime. It seems that the regime dependent and forward looking information in the VIX index can be more easily incorporated into the model.

Table 2 Diebold–Mariano Tests for VIX Contribution (noVIX vs. VIX Models)

Market	Architecture	Comparison	DM Statistic	p-value
HSI	ANN	noVIX_vs_VIX	-0.496028	0.620882
HSI	LSTM	noVIX_vs_VIX	-1.815068	0.072263*
HSI	GRU	noVIX_vs_VIX	-0.535078	0.593686
NIKKEI	ANN	noVIX_vs_VIX	-1.795064	0.075440*
NIKKEI	LSTM	noVIX_vs_VIX	-0.187219	0.851837
NIKKEI	GRU	noVIX_vs_VIX	-0.787376	0.432771
FTSE	ANN	noVIX_vs_VIX	0.890040	0.375423
FTSE	LSTM	noVIX_vs_VIX	-0.363306	0.717080
FTSE	GRU	noVIX_vs_VIX	0.409602	0.682902
DAX	ANN	noVIX_vs_VIX	0.585339	0.559541
DAX	LSTM	noVIX_vs_VIX	-1.012412	0.313583
DAX	GRU	noVIX_vs_VIX	-2.572292	0.011448*
SP500	ANN	noVIX_vs_VIX	0.770252	0.442832
SP500	LSTM	noVIX_vs_VIX	0.145506	0.884580
SP500	GRU	noVIX_vs_VIX	2.100099	0.038027*

** p < 0.10, ** p < 0.05, *** p < 0.01.*

4.2 Incremental Contribution of VIX Information

The impact of the implied volatility on the volatility forecast is not instantaneous or unconditional (Table 3). The VIX is only marginally significant for the ANN models and its inclusion in the model does not generally improve the performance of the model. The role of VIX in the LSTM model is limited and not stable across markets. The LSTM model can absorb a large amount of historical information on past realizations of volatility and the VIX futures seem not to bring any significant additional information for the forecast. In contrast to LSTM, the VIX plays a more stable and economically meaningful role in the GRU model, especially in the more financially interconnected markets. Our findings confirm the conditional role of VIX information for volatility forecasting. This conditionality strongly depends on the underlying model and market.

While all markets are subject to the same global risk drivers, there is significant market level heterogeneity. The market level where the VIX has the largest and most significant impact is in the

most directly global market indices such as the S&P 500 and the FTSE. On the other hand, for markets that are further away from the global risk drivers such as the HSI and the NIKKEI for the Asian region, the VIX has a weaker and less stable impact on the forecast of the volatility of the market at the particular level. This suggests that for the case of some markets in the Asian region, the impact of their risk drivers is significant to the determination of the volatility of the markets, thus noisier and less useful VIX.

Table 3 Incremental Effect of VIX on Forecast Accuracy (difference of VIX and noVIX)

Market	Architecture	Δ RMSE (VIX – noVIX)	Δ QLIKE (VIX – noVIX)
HSI	ANN	0.000039	0.000915
HSI	LSTM	0.000072	0.001867
HSI	GRU	0.000047	0.001036
NIKKEI	*ANN*	*0.000159*	*0.010047*
NIKKEI	*LSTM*	*-0.000015**	*0.000370*
NIKKEI	*GRU*	*-0.000022**	*0.001584*
FTSE	ANN	-0.000054*	-0.003216*
FTSE	LSTM	0.000013	0.000431
FTSE	GRU	-0.000003*	-0.000589*
DAX	*ANN*	*-0.000019**	*-0.001832**
DAX	*LSTM*	*0.000049*	*0.001421*
DAX	*GRU*	*0.000077*	*0.002779*
SP500	ANN	-0.000005*	-0.005713*
SP500	LSTM	0.000019	-0.000428*
SP500	GRU	-0.000134*	-0.009795*

**Models including VIX outperform models without VIX (negative values show VIX outperforming).*

4.3 Summary of Empirical Findings

We will refer to the performance order of the volatility forecasts as the volatility forecasting rank. The performance order is defined based on the respective performance of the four models. The worst performance is recorded for the ANN models that is not surprising. As we have already discussed, the feedforward structure of ANNs prevents them from exploiting the clustering and persistence properties of realised volatility, which are properties that require some form of memory in the model. The better performance of the LSTM and GRU models is consistent with the path dependence of volatility. What is not expected is that the GRU models perform better than the LSTM models. Even though the LSTM models have more complex gates, which could provide more flexibility, the additional parameters introduced in the GRU architecture seem to cause overfitting on the relatively short sample of daily data of moderate length.

The VIX index was present in some of the models, but not in others, and it could be past VIX, current VIX or future VIX depending on the model. Our interpretation of this result is that VIX is not an independent variable but a variable that is strongly correlated with the information already included in the variables of the model. Models with a priori accounting for the volatility clustering (i.e. through

an autoregressive or long memory component) can easily account for the inclusion of VIX as an extra piece of information. In such a case VIX and the autoregressive components of volatility can be seen as complementary variables. In feedforward models that have not accounted a priori for the volatility clustering through memory, the extra information that VIX provides (i.e. the forward-looking component of VIX) is not easily employed, as it is not a generic variable, but a variable that should be employed in a specific context.

The following tables 1, 2 and 3 present the empirical evidence for our three hypotheses, listed in order. H1 is supported by the results. Table 1 shows that the differences in performance between the different models are statistically significant. The expected result that the ANNs have a higher value of the error statistics than the recurrent models is confirmed for all loss functions and all markets. H2 is supported only to a certain extent. VIX is a statistically significant variable for the case of the GRU models, but its impact is not the same for all markets, and it varies to a greater or lesser extent over time, depending on the market and the type of model used, as shown by the results in Tables 1–3. H3 is confirmed by the findings as well. As already mentioned, we hypothesised that the memory-based models would perform better than the feedforward ANNs. Given the high degree of persistence and autocorrelation of financial volatility, the confirmation of H3 is indeed unsurprising.

4.4 Discussion

The results of this study demonstrate that model architecture is crucial in short-term volatility forecasting. No single model can be said to be superior in all cases; forecasting performance depends on how volatility dynamics are represented within the model. The findings show that memory-based networks produce stronger results. LSTM and GRU models perform better than feedforward networks (ANN) over a one-day forecast horizon. This result is expected given the persistent nature and regime sensitivity of financial volatility (Engle 1982; Cont 2001). Temporal dependence is strong in volatility processes, and architectures that can internally model this dependence can make more effective forecasts. However, this flexibility is more limited in feedforward networks. The contribution of the VIX variable also differs among architectures. While some models show improved forecasting performance, others show no significant change. This indicates that forward-looking risk indicators alone do not guarantee forecasting power. The contribution depends on its relationship with the existing information structure.

The results indicate that VIX's contribution is heterogeneous and limited depending on the architecture. VIX is only useful when it complements the information provided by historical volatility. There are several structural explanations for this result. First, VIX reflects the implicit volatility derived from option prices and is calculated under a risk-independent measure. The realized volatility used in this study, however, is based on a physical measure. Therefore, there is a natural mismatch between the two volatility measures (Carr and Wu 2009; Bollerslev, Tauchen, and Zhou 2009). The strong persistence of realized volatility is also important. Historical volatility already provides strong predictive power in many cases. Memory-based architectures, such as LSTM and GRU, can utilize this information quite effectively. In such cases, the marginal contribution of additional risk indicators may remain limited. Another factor is market differences. VIX is a global risk indicator derived from

US option markets. It indirectly represents risk conditions in European and Asian markets. This may partially explain the market-dependent results.

5. Conclusion

This study focuses on model architecture and forward-looking risk information in short-term volatility forecasting. The analysis was conducted within a rolling out-of-sample framework covering global equity markets. The results show that model structure is crucial. Network architectures including GRU and LSTM memory provide higher forecasting accuracy. It is known from the literature that volatility processes are persistent and regime dependent. Therefore, structures that can model temporal dependence produce better performance. According to the findings, the contribution of the VIX variable was conditional. While forecasting accuracy increased in GRU architectures, the effect remained weak in the ANN model. However, VIX only provides benefits when it complements the information provided by past volatility; no systematic contribution was observed. An important conclusion emerges from the application perspective. In practice, model architecture has been shown to be a critical choice in volatility forecasting. Flexible and memory-inclusive structures offer advantages in short-term forecasting, and additional risk indicators should not always be assumed to contribute to the forecast. In each case, they should be empirically tested. This study has some limitations. The analysis is limited to specific architectures and a single risk indicator. Different network structures or larger datasets may produce different results. Future research may follow this direction. Ultimately, volatility forecasting cannot be explained by a single model or a single data source. Forecasting success depends on the fit between the model structure and the data content, but market conditions also shape this relationship.

Author Contributions

Conceptualization, O.O. & A.A.K.; methodology, O.O. & A.A.K.; software, O.O. & A.A.K.; validation, O.O. & A.A.K.; formal analysis, O.O. & A.A.K.; investigation, O.O. & A.A.K.; resources, O.O. & A.A.K.; data curation, O.O. & A.A.K.; writing—original draft preparation, O.O. & A.A.K.; writing—review and editing, O.O. & A.A.K.; visualization, O.O. & A.A.K.; supervision, O.O. & A.A.K.; project administration, O.O. & A.A.K. The author has read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

The data used in this study are publicly available. Daily equity index data were obtained from Stooq, and VIX historical data were obtained from the Chicago Board Options Exchange (CBOE).

Conflicts of Interest

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was not funded by any grant.

References

[1] Andersen, T. G., Bollerslev, T., & Diebold, F. X. (2007). Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. Review of Economics and Statistics, 89(4), 701–720. https://doi.org/10.1162/rest.89.4.701.

[2] Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. Econometrica, 71(2), 579–625. https://doi.org/10.1111/1468-0262.00418.

[3] Barndorff-Nielsen, O. E., & Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. Journal of the Royal Statistical Society: Series B, 64(2), 253–280. https://doi.org/10.1111/1467-9868.00336.

[4] Basel Committee on Banking Supervision. (2016). Minimum capital requirements for market risk. Bank for International Settlements. https://www.bis.org/bcbs/publ/d352.pdf.

[5] Bekaert, G., & Hoerova, M. (2014). The VIX, the variance premium and stock market volatility. Journal of Econometrics, 183(2), 181–192. https://doi.org/10.1016/j.jeconom.2014.05.008.

[6] Bishop, C. M. (1995). Neural networks for pattern recognition. Oxford University Press.

[7] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1.

[8] Bollerslev, T., Gibson, M., & Zhou, H. (2011). Dynamic estimation of volatility risk premia and investor risk aversion from option-implied and realized volatilities. Journal of Econometrics, 160(1), 235–245. https://doi.org/10.1016/j.jeconom.2010.03.033.

[9] Bollerslev, T., Tauchen, G., & Zhou, H. (2009). Expected stock returns and variance risk premia. Review of Financial Studies, 22(11), 4463–4492. https://doi.org/10.1093/rfs/hhp008.

[10] Carr, P., & Wu, L. (2009). Variance risk premiums. Review of Financial Studies, 22(3), 1311–1341. https://doi.org/10.1093/rfs/hhn038.

[11] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1724–1734). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1179.

[12] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv. https://doi.org/10.48550/arXiv.1412.3555.

[13] Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. Quantitative Finance, 1(2), 223–236. https://doi.org/10.1080/713665670.

[14] 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), 1–24. https://doi.org/10.1080/07350015.2014.983236.

[15] Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics, 13(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599.

[16] Ding, Z., Granger, C. W. J., & Engle, R. F. (1993). A long memory property of stock market returns and a new model. Journal of Empirical Finance, 1(1), 83–106. https://doi.org/10.1016/0927-5398(93)90006-D.

[17] Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 50(4), 987–1007. https://doi.org/10.2307/1912773.

[18] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

[19] Hornik, K. (1989). Multilayer feedforward networks are universal approximators. Neural Networks, 2(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8.

[20] McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, 5, 115–133. https://doi.org/10.1007/BF02478259.

[21] Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. Journal of Econometrics, 160(1), 246–256. https://doi.org/10.1016/j.jeconom.2010.03.034.

[22] Whaley, R. E. (2000). The investor fear gauge. Journal of Portfolio Management, 26(3), 12–17. https://doi.org/10.3905/jpm.2000.319728.