International Review of Financial Analysis 93 (2024) 103221

Contents lists available at ScienceDirect

International Review of Financial Analysis

journal homepage: www.elsevier.com/locate/irfa

Review

Prediction of realized volatility and implied volatility indices using AI and machine learning: A review

Elias Søvik Gunnarsson, Håkon Ramon Isern, Aristidis Kaloudis, Morten Risstad *, Benjamin Vigdel, Sjur Westgaard

Norwegian University of Science and Technology (NTNU), Faculty of Economics and Management, Department of Industrial Economics and Technology Management, Norway

ARTICLE INFO

Keywords:
Volatility forecasting
Machine learning
Explainable artificial intelligence

ABSTRACT

In this systematic literature review, we examine the existing studies predicting realized volatility and implied volatility indices using artificial intelligence and machine learning. We survey the literature in order to discover whether the proposed methods provide superior forecasts compared to traditional econometric models, how widespread the application of explainable AI is, and to outline potential areas for further research. Generally, we find the efficacy of AI and ML methods for volatility prediction to be highly promising, often providing comparative or better results than their econometric counterparts. Neural networks employing memory, such as Long–Short Term Memory and Gated Recurrent Units, consistently rank among the top performing models. However, traditional econometric models are still highly relevant, commonly yielding similar results as more advanced ML and AI models. In light of the success with ensemble methods, a promising area of research is the use of hybrid models, combining machine learning and econometric models. In spite of the common critique of many machine learning models being of a black-box nature, we find that very few papers apply XAI to analyze and support their empirical results. Thus, we recommend that researchers strive harder to employ XAI in future work. Similarly, we see potential for applications of probabilistic machine learning, effectively quantifying uncertainty in volatility forecasts from machine learning models.

1. Introduction

Due to significant growth in the use of artificial intelligence and machine learning techniques in the recent years, there has been an equally growing interest in applications within prediction and forecasting of volatility using these methods. Volatility is a complex and dynamic phenomenon of interest to many stakeholders within finance and economics. Reliable volatility forecasts are important tools for these stakeholders in anticipating changes in market conditions to which they can adapt accordingly. For example, accurate volatility forecasts may help a fund manager to reduce their exposure to potential markets risks, by reducing their positions in assets that are likely to be affected by increased volatility or by hedging their positions using derivatives. Volatility is also an important factor in determining the value of such derivatives. Furthermore, risk managers may utilize accurate volatility forecasts to compute more precise Value at Risk (VaR)[1] measures. Traders can use accurate forecasts to make more informed decisions about when to enter and exit positions in financial markets. For instance, the low volatility during the summer of 2022 have caused traders large losses and stresses the importance of high conviction of volatility movements (Tsekova & Popina, 2022). Finally, governments and institutions can benefit from accurate volatility forecasts to assess the volatility and its impacts on the economy to make effective policy decisions regarding monetary and regulatory policy.

Estimating and forecasting volatility is complex, since volatility itself is a latent variable. Realized volatility, inferred from the sum of squared intradaily high-frequency returns, have since the seminal contribution of Andersen et al. (2001a), been considered the most appropriate representation of the true, unobservable integrated variance. While realized volatility is inherently backward-looking, implied volatility can be interpreted as the risk-neutral expectation of volatility. Implied volatility is an important determinant of option prices and is particularly appealing for forecasting purposes, due to its forward looking nature.

Machine learning has found applications across the banking and finance industry; often attributed to emergence of big data, access to

* Correspondence to: Alfred Getz vei 3, 7034 Trondheim, Norway.
E-mail addresses: eliassg@stud.ntnu.no (E.S. Gunnarsson), haakoris@stud.ntnu.no (H.R. Isern), aristidis.kaloudis@ntnu.no (A. Kaloudis), morten.risstad@ntnu.no (M. Risstad), benjamin.vigdel@ntnu.no (B. Vigdel), sjur.westgaard@ntnu.no (S. Westgaard).

[1] See Table D.5 for a list of abbreviations used in this study.

https://doi.org/10.1016/j.irfa.2024.103221
Received 21 March 2023; Received in revised form 26 December 2023; Accepted 6 March 2024
Available online 16 March 2024
1057-5219/© 2024 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

computing power and scalable statistical software. Machine learning is fundamentally different to traditional econometrics with regards to assumptions about the true data generating process. The former typically assumes data to be generated by some stochastic process, whereas the latter treats the data mechanism as unknown. Hence, the flexible nature of machine learning algorithms enables them to handle complex, non-linear dynamics in a model free framework, possibly encompassing highly correlated predictors and structured datasets. This has sparked interest in machine learning for volatility forecasting applications.

Over the last couple of years, there have been several literature reviews which study various deep learning-based neural network approaches for stock market forecasting, which mainly focuses on stock price prediction (Jiang, 2021; Thakkar & Chaudhari, 2021), while Sezer et al. (2020) studies financial time series forecasting in general, using deep learning. Henrique et al. (2019) and Kumbure et al. (2022) focus on machine learning approaches for various financial market predictions. Furthermore, Bustos and Pomares-Quimbaya (2020) focus on stock market movement prediction. Common for all the mentioned reviews is that volatility forecasting is either minor part of the scope, or not present. Poon and Granger (2003) conducted an early review concerning volatility forecasting in financial markets, but due to the more recent popularity and boom within machine learning and artificial intelligence, such approaches are not considered in that review.

The important foundation we know today about volatility forecasting originates from how known stylized facts about asset returns were incorporated in parameterized models. The autoregressive conditional heteroscedasticity (ARCH) model of Engle (1982), further developed by Bollerslev (1986) to the Generalized-ARCH (GARCH), which models the conditional variance as dependent on yesterday's disturbance term as well as yesterday's conditional variances. These were further extended in various forms to include asymmetric behavior and leverage effects, see for example Nelson (1991) and Glosten et al. (1993). Engle et al. (2013) introduced a family of univariate GARCH-models using Mixed Data Sampling Frequency (MIDAS) filters for inclusion of additional variables in order to both capture short- and long-term effects.

The availability of high frequency data has lead to more accurate volatility estimation such as realized volatility (RV), see Appendix A. The realized volatility proxy has no model-based relationship between the time periods, like the conditional variance, and thus can be modeled directly using for example autoregressive models. Inspired by the heterogeneous market hypothesis, Corsi (2009) utilized realized volatility to create a heterogeneous autoregressive model, see Appendix B. The availability of high-frequency data and realized volatility also motivated new extensions of the GARCH-model, such as the Realized-GARCH by Hansen, Huang et al. (2011). Alternative volatility measures such as implied volatility have together with realized volatility motivated alternative approaches for volatility forecasting, such as machine learning and AI. Malliaris and Salchenberger (1996) developed a neural network to forecast implied volatility where it was estimated in a similar manner as the well known CBOE Volatility Index (VIX). The VIX is a measure of the expected volatility of the S&P 500 index over the 30 days, and is calculated using real-time prices of options on the S&P 500 index (see Appendix C). Furthermore, is considered to be a leading indicator of investor sentiment and market volatility.

In this review, we give an overview and examine the current state of the art in the use of AI and machine learning for predicting realized volatility and implied volatility indices. To this end, we conduct searches in four major scientific literature databases and report on the result. We provide a brief summary of each study, highlighting what methods were used and what results were achieved. Generally, we find promising results for the efficacy of AI and machine learning in volatility forecasting, particularly through the use of hybrid and ensemble methods. To shed light on the general trends within the literature, we present descriptive statistics on the different facets inherent to the forecasting literature. We find that the US equity market comprises the largest area of study, and that the use of exogenous data often provides performance improvements, particularly over longer forecasting horizons. Furthermore, we find that memory-based neural networks like LSTM, random forest and boosting methods constitutes the majority of applied AI methods. Additionally, we present the different sampling frequencies and forecasting horizons, and find that intradaily and daily sampling frequencies are most widely used. Furthermore, we discuss important findings in light of our research questions, in addition to key challenges and opportunities for further research within the area.

We find that, going forward, the use of XAI and implied volatility forecasting proves an interesting direction for further research, as well as combination of the promising performing models and Bayesian approaches to quantify the uncertainty introduced by models to explain trustworthiness of the predictions. To the best of our knowledge, this is the first systematic review tackling the application of artificial intelligence and machine learning in volatility forecasting, where implied volatility and realized volatility are used as proxies.

The remainder of this paper is organized as follows. Section 2 provides a detailed description of our methodology. Section 3 describes typical empirical characteristics of the existing literature; such as markets, data and models investigated. Section 4 presents our most important findings and Section 5 concludes.

2. Methodology

The scope of this literature review is to summarize the existing research and methods concerning prediction of volatility using machine learning or artificial intelligence. Importantly, we also identify gaps in this research in order to advise areas for further research. We follow a systematic methodology, as literature reviews which are not thorough and transparent are of little scientific value (Kitchenham, 2004).

More specifically, the research questions focus on whether (i) the machine learning models are benchmarked on a fair basis with state-of-the-art econometric models, (ii) what the motivation for prediction of volatility is, (iii) which markets are studied, and (iv) whether volatility is forecastable with AI and machine learning. In relation to the last research question we investigate with particular scrutiny empirical evidence in favor of machine learning models relative to alternative approaches. As machine learning models tend to be less parsimonious than traditional econometric models, this brings up challenges regarding explainability of the models. Consequently, we provide an overview of the use of Explainable AI (XAI) in the volatility forecasting literature.

In order to make comprehensive and unbiased searches in literature databases, we identify keywords and search terms through the planning process of the systematic review which are based on titles, abstracts, and keywords from the studied material together with external and internal domain knowledge. The initial focus are studies which use any kind of implied volatility index as a proxy for volatility in forecasting or prediction. However, we find it appropriate to include studies using realized volatility as a proxy as well. This is because they both, unlike conditional volatility estimated by squared returns in the traditional GARCH-models, have no model-based relationships between the days. With indices like the VIX being directly observed, the realized volatility estimate of the latent volatility can be considered close to observable. Thus, these proxies can in theory be modeled using the same kind of modeling techniques. Additionally, there is limited amount of research conducted using implied volatility indices, such as the VIX, as a volatility proxy.

Based on these findings, through an iterative process which is consistent with guidelines from Kitchenham (2004), the search queries featured in Table 1 are constructed according to the required syntax of the databases to have as close meaning as possible. These search queries are combinations of keywords and subject terms to both broaden and focus our search according to the above-mentioned criteria. However, they are a compromise, as a purely focused search would result in missing

Table 1
The initial database searches. The queries are intended to be identical but are modified to be compatible with the databases syntaxes.

Database	Search query
Web of Science	((Realized volatility OR Implied volatility OR Volatility index) AND (Forecast OR Predict OR Forecasting OR Prediction) AND (AI OR Artificial intelligence OR Machine learning)) (Title)(Abstract)(Keywords)
Scopus	TITLE-ABS-KEY ((“Realized volatility” OR “Implied volatility” OR “Volatility index”) AND (“Forecast” OR “Predict” OR “Forecasting” OR “Prediction”) AND (“AI” OR “Artificial intelligence” OR “Machine learning”))
Science Direct	Title, abstract, keywords: ((Realized volatility OR Implied volatility OR Volatility index) AND (Forecast OR Predict OR Forecasting OR Prediction) AND (AI OR Artificial intelligence OR Machine learning))
ProQuest	SUMMARY, IF, TITLE (((Realized volatility OR Implied volatility OR Volatility index) AND (Forecast OR Predict OR Forecasting OR Prediction) AND (AI OR Artificial Intelligence OR Machine learning)))

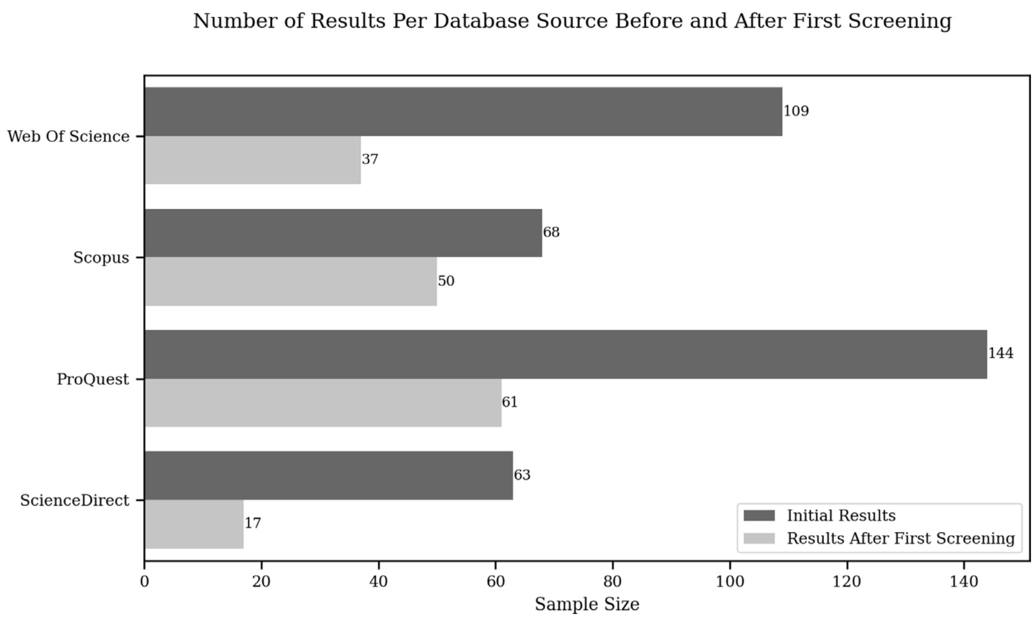

Fig. 1. The number of results from each database after initial search and screening.

out on some potentially relevant literature. Consequently, the search results are manually screened. The selected databases are Web Of Science, Scopus, Science Direct and ProQuest, which are among the largest and more commonly used for such database-driven searches. By conducting searches through multiple multi-publisher databases most publications from publisher-specific sources will also be included (Hiebl, 2021).

The first screening on the initial search results in the databases is done based on the information provided in the search summary of the databases with the following criteria: realized volatility or implied volatility must be included in either title or abstract. Secondly, the title or abstract must mention machine learning or artificial intelligence. Hybrid econometric models which include machine learning or artificial intelligence are accepted. The reason why traditional GARCH-models are not included is because they either implicitly or explicitly rely on squared returns as a volatility proxy for a model-based forecasting of conditional volatility between the days. Another important reason is that extensive research and reviews already exist within that area. Finally, prediction or forecasting must be a part of the paper's objective and be pointed out in the title or abstract. The initial combined search result consisted of 384 papers, not accounted for duplicates. As shown in Fig. 1, the number of papers decreased to 165.

Bibliometric files from all the databases are downloaded and stored in comprehensive Pandas DataFrames in a Jupyter Notebook. Due to different structures in the bibliometric files from the different sources, custom mappings are created to merge the files into a single data frame, with identical column names. This makes it possible to eliminate duplicates in the present sample. The sample is further decreased through a second screening where we investigate whether the papers meet the criteria that the objective of the paper is to predict volatility, where the volatility proxy is either realized volatility, or implied volatility indices. Additionally, the model used for the prediction must either be a machine learning model or a hybrid model. We also discuss included and excluded deviations through a second review. After removing duplicates, as shown in Fig. 2, the sample consists of 93 papers, and 54 after the last criterion based exclusions.

We gather quantitative metrics related to paper- and journal metrics, namely, Scopus citation score, Scimago journal rankings, Web of Science (Clarivate) journal impact factor and rankings, and total citations. Solely relying on such metrics can lead to favoring generalist journals instead of specialist journals and were therefore used with carefulness as exclusion criteria, see Hoepner and Unerman (2012). To ensure that such specialist journals were not excluded solely on the above metrics, we also considered a set of expert suggestions within the field financial forecasting. When papers have an origin from more than one of the selected database sources, the highest total citation number is chosen.

We also exclude unfinished works, unpublished papers, as well as conference papers from our sample. Hiebl (2021) presents different opinions whether such gray literature should be included or not in the systematic sample, but we chose to value peer-reviewed papers, and choose to only include those. Additionally, only work written in the English language was included.

The final sample consists originally of 28 papers, adjusted to 32 after inclusion of additional highly relevant papers which met the above mentioned requirements. These are discovered through snowball sampling while doing a comprehensive reading of all the papers following a predefined reading template and research questions.

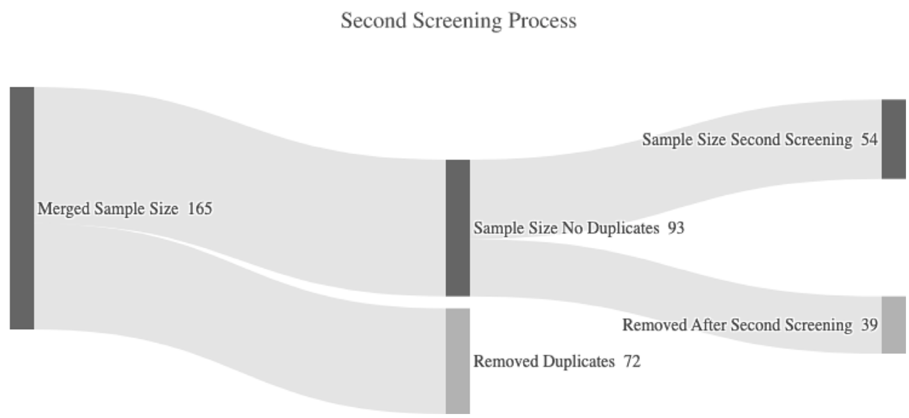

Fig. 2. Results after screening for whether the objective of the paper is to predict realized of implied volatility.

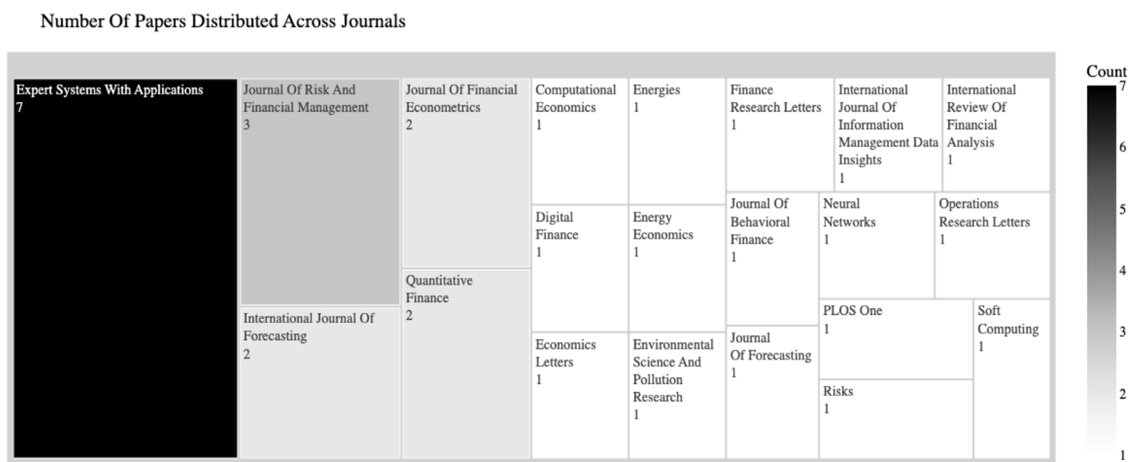

Fig. 3. Distribution of included papers across academic journals.

3. Existing literature — data and empirical design

This sections summarizes descriptive dimensions of the existing literature. More specifically, we categorize the literature by academic publishing channel, markets and data investigated — including sampling frequency, models and model evaluation criteria.

Journals

The sample originates from 21 different journals, and the composition is portrayed in Fig. 3. Expert Systems With Applications is the biggest contributor, aligning with the previous findings of Bustos and Pomares-Quimbaya (2020) and Henrique et al. (2019). Eight of the journals cover finance and finance-related matters, while four is primarily dedicated to research associated with computer science. PLOS One is a general journal covering findings within a wide array of fields and disciplines, leaving nine journals more specialized towards a specific area. These cover energy and environmental aspects, forecasting, general economics and operational research.

Scientific Productivity

There has been a rising prevalence of AI and machine learning in financial time series forecasting. In related reviews, see e.g. Kumbure et al. (2022) and Bustos and Pomares-Quimbaya (2020), the amount of papers applying different ML and AI methods increases drastically over the latter part of the previous decade. We find a similar trend in our sample, where Fig. 4 showcases the increasing popularity of AI methods for volatility estimation. The rapid surge of papers originating from 2021 and 2022 is of particular interest. This trend might indicate an increasing interest in the combination of AI and volatility estimation methods, perhaps in part due to the turbulent state of financial markets during the early parts of the 2020s.

Markets

The associated market of each target variable can be identified based on the underlying asset class: commodities, equities, or currencies. Then, a further split into the individual asset can be performed. This grouping is displayed in Fig. 5. Some studies used a selection of stocks from one stock exchange or index, and these are put into the "Stock" category. Note that some studies involves several target variables, causing an inflated count for some markets. However, none of the studies used assets across the initial market segmentation. As such, the overall grouping of the papers remains informative. The regional breakdown of each asset is displayed in Fig. 6. 19 of the studies dealt with equity market, where American stocks or indices comprised the biggest market. Some composite indices in this category consists of assets outside of the US, but for simplicity they will be classified based on the index' country of origin. Eight studies used commodities, six of which were oil. These were primarily based on WTI indices, but Tissaoui et al. (2022) focused their study on the OVX. As these are all based on American assets, they were placed into the American market. The remaining commodity-based studies used gold as the underlying asset for their volatility forecasting. Finally, five studies utilized currencies, where three handled Bitcoin and two used currency exchange rates. Due to the international nature of currency exchange rates and cryptocurrencies, these are not assigned a nationality.

From the above analysis, it is apparent that a majority of the studies in our sample were based on American oil and equity markets. The high volume of activity and market capitalization in these markets

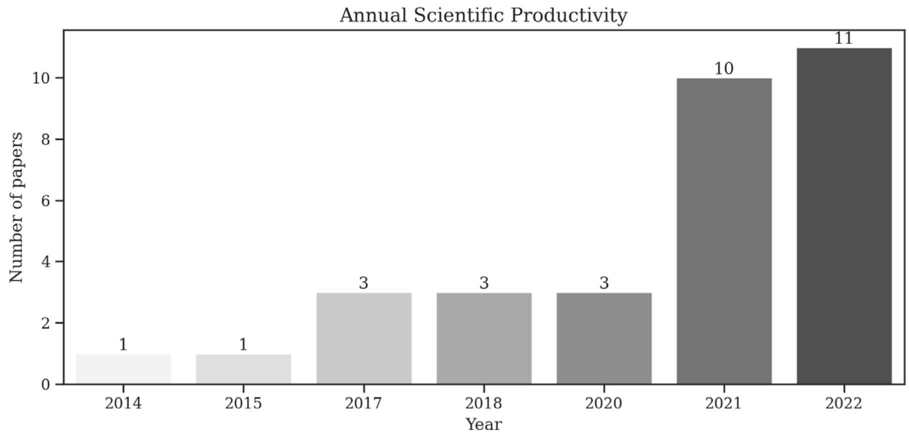

Fig. 4. The annual scientific productivity in the included sample.

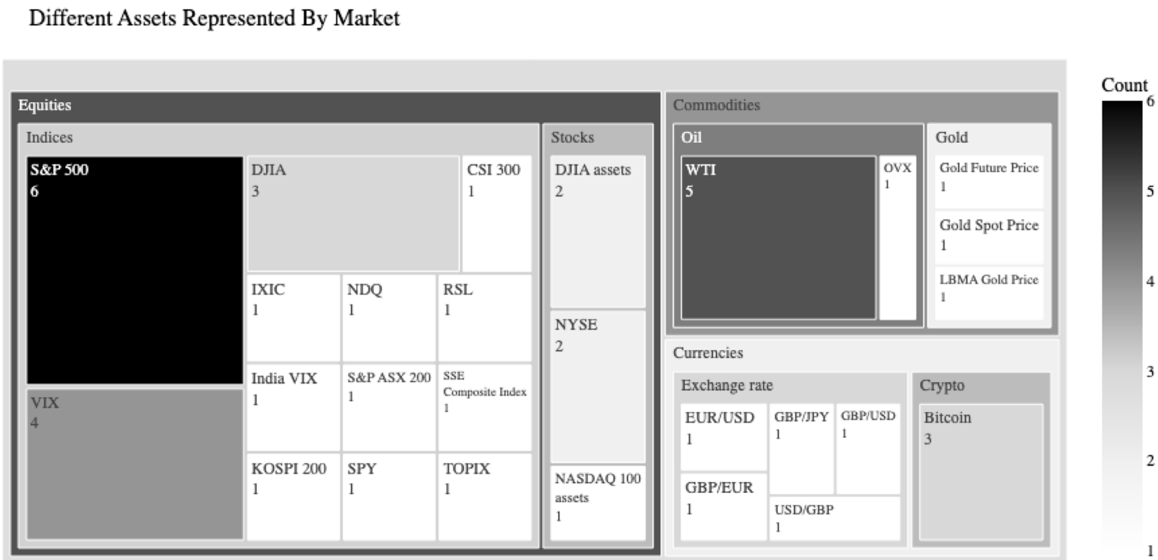

Fig. 5. The underlying target assets divided into market and asset type.

can make them more attractive for closer investigation, and they can be used as proxies for the global market. However, there are some benefits to approaching more localized markets, especially if measured against the larger indices. It might aid in discovering local differences, in turn allowing for a deeper understanding of the drivers of volatility. Additionally, some localized assets might exhibit different behavior compared to the larger counterparts, as shown by Chen and Hu (2022). The CSI 300 index portrayed more extreme statistical values than the S&P 500, allowing the authors to both compare the assets and the performance of each model in different environments. This can be utilized to assess the generality of the model, and how it copes with processes exhibiting different characteristics than the mainstream assets.

Data

Several of the studied papers utilized external data. Many of the authors specifically investigate effects of certain variables on forecasting performance. Others incorporated additional data to aid their given model. In total, 22 of the papers included some exogenous data in their set of predictors. This consisted of macroeconomic, financial, technical and uncertainty-related variables, specified in Table 3. The former category expresses information about the general economical environment through variables like inflation, credit spreads and industrial production growth. Financial variables represent details about the financial market, e.g. Fama–French factors, and firm-specific facts where applicable. The technical indicators recounts characteristics about a given financial time series, e.g., through momentum and moving average. Uncertainty-related data accounts for variables related to measurements of general uncertainty. This includes both political and news-based risk indices, in addition to quantification of publicly perceived risk through sentiment surveys, Google search trends and microblogging data.

The inclusions of exogenous data were mostly motivated by earlier works and recommendations, were there appeared to be wide support within the literature for including extra data. Several studies, see e.g. Bucci (2020), have found strong links between volatility and macroeconomic foundations, particularly for longer forecasting horizons, motivating the inclusion of this data for prediction purposes. Further, building on the approach of Zhang, He et al. (2022), using high-frequent sentiment data regressed at the residuals of an AR model might prove a promising avenue for including new data.

Of the remaining ten papers, two applied technical indicators based on their respective asset of choice: Petrozziello et al. (2022) provided their LSTM with open-close returns in addition to the realized volatility estimate, and Luong and Dokuchaev (2018) converted volatility into five technical indicators. Additionally, Vidal and Kristjanpoller

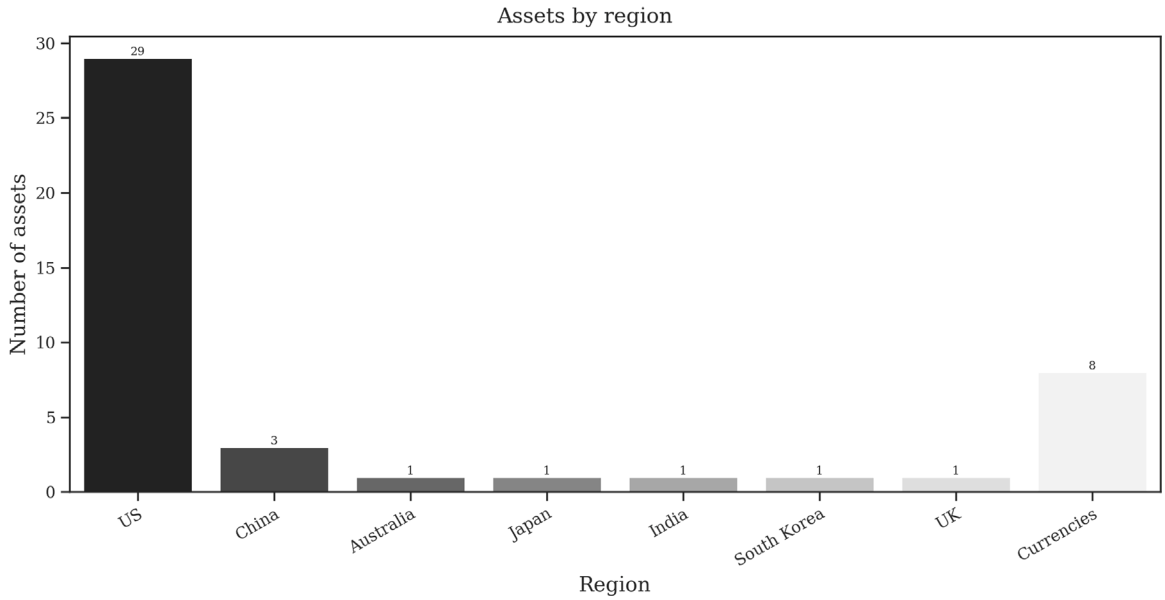

Fig. 6. The number of target assets by region. The region is determined by the source of the asset, or the location of its corresponding trading exchange.

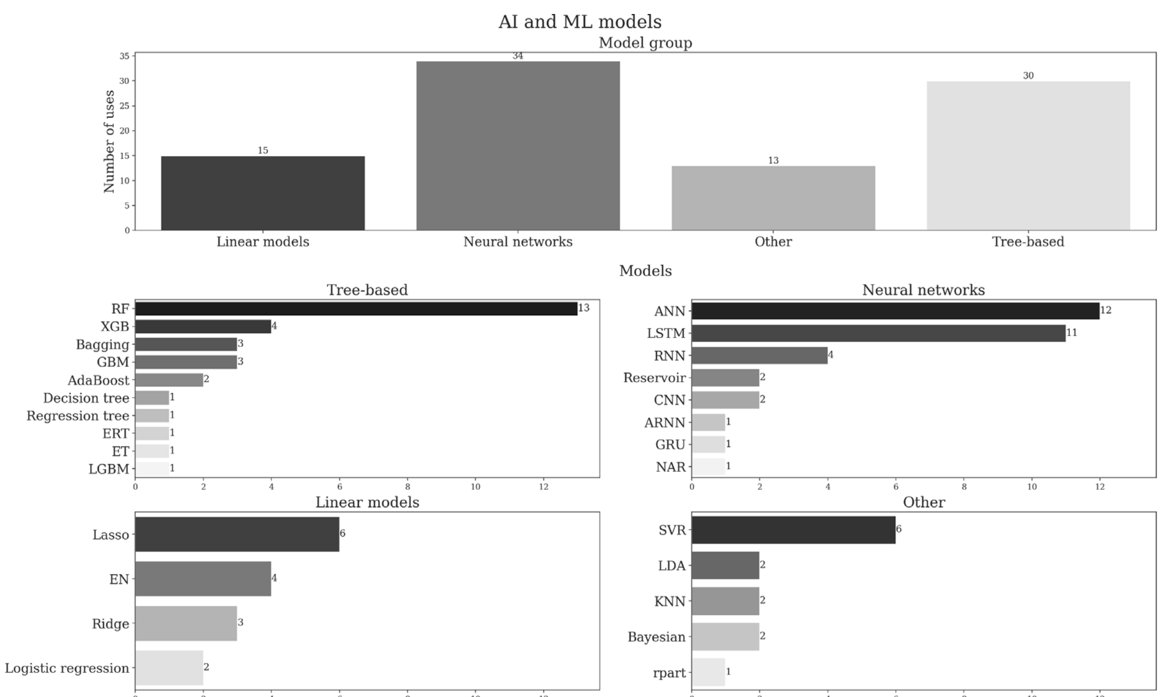

Fig. 7. The number of papers using each ML and AI model. The topmost figure details the usage of the overarching family of models, and the lower figure breaks down each group into its respective model frequency.

(2020) generated images based on returns as input to a CNN-LSTM hybrid network. The remaining seven papers used no extra data besides autoregressive components.

Models

As specified in Section 4.2, the motivations behind choice of models varied between papers. Some merely wanted to study the effect exogenous data had on predictive performance, thereby focusing on applying machine learning techniques for estimation purposes, while others were more concerned with using AI and ML models for the entire forecasting pipeline. This spurred the use of many different models in our sample, both econometric and ML and AI model. For simplicity's' sake, we have included AR and its family of models under the econometric umbrella of models.

In Fig. 7, the number of AI and ML models used in the sample are showcased. The topmost figure displays the general group of models, and gives a clear view of the most popular categories; neural networks and tree-based models constituted about 75% of these models. Moving to the lower section of the figure, a breakdown of each category is provided. Amongst the tree-based models, RF were the most widely used. This was partly caused by its favored use as a hybrid model in tandem with HAR and AR models, where these comprised six of its thirteen implementations. Generally, its ability to handle multicollinearity and nonlinearity in a data-driven way makes it an attractive option for including exogenous data. This was demonstrated by Christensen et al.

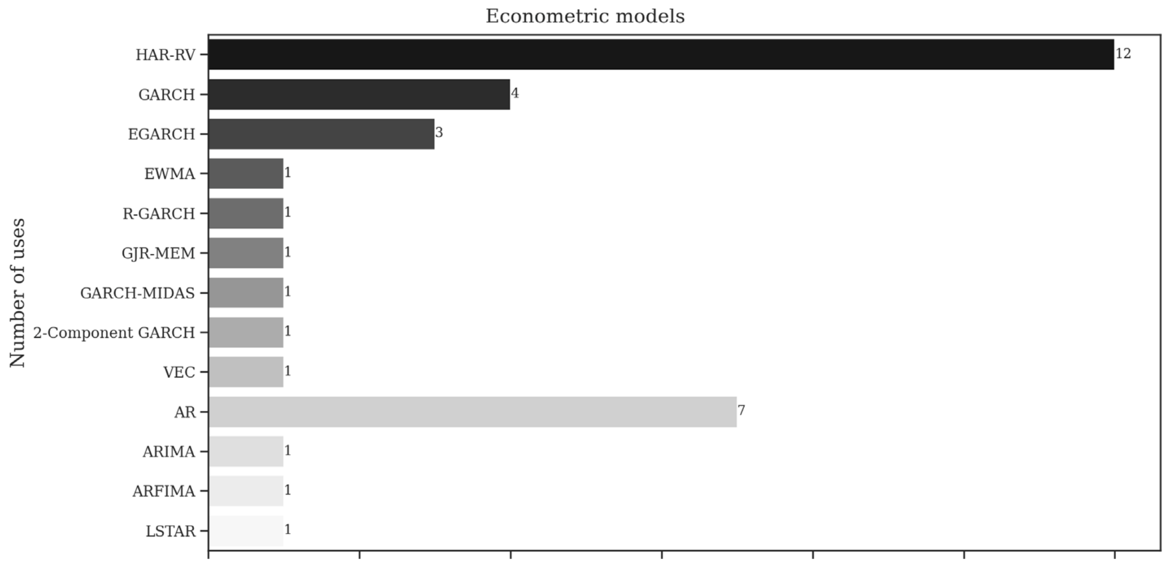

Fig. 8. The number of papers using each given econometric model. Except for the GARCH models, these counts includes extensions of the baseline model, e.g., HAR-RV accounting for jumps and leverage effects.

(2022), where increasing the forecasting horizon severely improved the performance of RF over econometric models when including external data. The other popular models in this category were mainly boosting methods. One advantage of this family of models comes with their variable importance feature; by inspecting splits made within the trees, it is possible to assess how much the model weighs different predictors.

The most popular category of models were neural networks. This includes simple feed-forward-networks, autoregressive networks and the recurrent network family of models. ANNs – both singular and multilayered perceptrons – and LSTMs were the most used models within this category. In particular, LSTM consistently ranked among the top performers when included, likely due to its ability to handle the inherent memory found within volatility time series. Additionally, it saw some use as the AI component of a hybrid model, especially when combined with GARCH models. The remaining NN models included some newer innovations, e.g. the reservoir and GRU, which tended to deliver good results.

Linear models and "other" makes up the final two categories. The former is composed of logistic regression – used for the directional classification problems – and penalized regression models: Ridge regression, LASSO and Elastic net. These three were used in different ways; either as independent models, as in Christensen et al. (2022), or to apply feature selection. In the latter case, LASSO or the likes were applied beforehand to select a subset of features, thereby reducing the model complexity, as shown in Zhang, Wahab et al. (2022), where the models applying feature selection greatly improved its performance. The "other" category consisted of models not fitting into a clear family of models, where SVR where the most used model (see Fig. 8).

For the econometric models, displayed in Fig. 8, HAR-RV remained the most popular choice, likely due to its role as one of the prevailing models within realized volatility forecasting. Additionally, its role as an AR-like model makes it suitable for applying several extensions, be it technical components – e.g., leverage effect and jump components – or exogenous data. In addition to the AR models, they were particularly favored as the baseline model for testing the effects of external variables; the family of papers investigating effect of EPU-like predictors mainly used AR or HAR-RV as their baseline, as mentioned in Section 4.2. After HAR-RV, the GARCH family of models composed a majority of the econometric models. These were either implemented as benchmarking models, as in the case of Petrozziello et al. (2022) with GJR-MEM and R-GARCH, or as inputs to hybrid AI models.

Sampling Frequency

The sampling frequency of the target asset is shown in Fig. 9. Three main groupings can be identified: monthly, daily and intradaily. Generally, the choice of sampling frequency can be attributed to the use of external data; it is often not available below a monthly level, particularly for macroeconomic and uncertainty variables. Thus, with some exceptions, intradaily frequencies were mainly used in studies only using endogenous data. Persio et al. (2021) – using only internal data – motivated their use of monthly realized volatility based on reducing day-to-day price fluctuations. Further, out of the 21 papers using exogenous data, two employed intradaily frequencies. Christensen et al. (2022) used five macroeconomic indicators available at a daily level, allowing them to utilize intradaily frequencies when estimating daily realized volatility. Similarly, Higashide et al. (2021) based their research on high frequent financial and technical data – such as trading volume and bid–ask spreads – accessible at an intradaily frequency. Some authors also applied two different sampling schemes. Prasad et al. (2022) used both daily and weekly macroeconomic variables to forecast the corresponding direction of the VIX, where the weekly data was included due to reporting delays for daily data.

Another aspect of the forecasting procedure of the papers regards the forecasting horizon. Table 2 shows the baseline volatility frequency and the forecasting horizon used by each paper. Contrary to the studies using implied volatility, all papers dealing with realized variance had to estimate their volatility proxy. Thus, the sampling frequency is usually higher than the corresponding volatility estimate. However, while the majority of studies computed realized volatility by summing square returns, Chen and Hu (2022), Kristjanpoller and Minutolo (2015) and Kim and Won (2018) used forward-looking approaches. Here, volatility was estimated by adding over terms consisting of the return at time i subtracted by the average return over the given period. Thus, they did not use intradaily data to estimate daily realized volatility, instead opting to base volatility on future values.

The forecasting horizons used varied greatly, from simple 1-step ahead up to several hundred steps. Ten studies reported only using 1-step ahead forecasts, while those exploiting several forecasting horizons differed in the used number of steps ahead. Yao et al. (2017) employed the longest horizon of 500 steps ahead for both hourly and daily forecasts. As for the number of different forecasting horizons, an average of 3.4 different horizons were investigated per study. Notably, Grigoryeva et al. (2014) tested 20 different forecasting horizons; from 1 up to 20 days ahead.

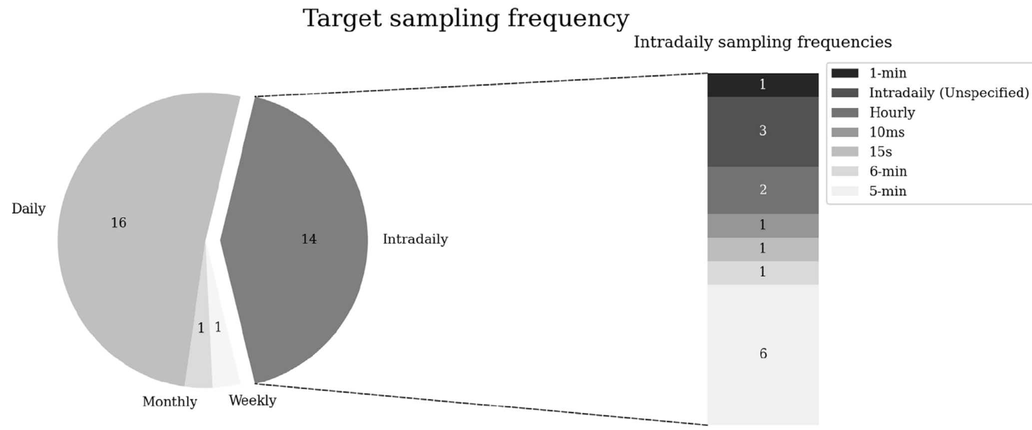

Fig. 9. The sampling frequency of the target asset. The leftmost figure displays the general sampling frequency, while intradaily frequencies are expanded upon in the right figure. One study used both daily and weekly frequencies, thereby giving a higher total count than the actual sample size.

Table 2
The table displays the sampling frequency of the target asset, the frequency of the volatility proxy and the different forecasting horizons utilized in each paper.

Reference	Sampling frequency	Proxy frequency	Forecasting horizon
Vrontos et al. (2021)	Monthly	Monthly	1-month-ahead
Petrozziello et al. (2022)	Intradaily (unspecified)	Daily	1-day-ahead
Luong and Dokuchaev (2018)	15 s interval	Daily	1-, 5- and 22-day-ahead
Tissaoui et al. (2022)	Daily	Daily	1-day- and week-ahead
Song et al. (2022)	Intradaily 5-min	Daily	1-and 5-day-ahead
Christensen et al. (2022)	Intradaily 5-min	Daily	1-day, 1-week, 1-month-ahead
Liu et al. (2018)	Intradaily 5-min	Daily	1-, 2-, 5-day-ahead
Çepni et al. (2022)	Daily	Monthly	1-, 3-, 6-, 12-, 18-, 24-, 32-, 48-month-ahead
Gkillas et al. (2021)	Hourly	Daily	1-day, 1-week, 1-month-ahead
Yao et al. (2017)	Intradaily 10 ms	Hourly and daily	5-, 20-, 100-, 200-, 360-, 500-h/days-ahead
Bouri et al. (2020)	Hourly	Daily	1-, 5-, 22-days-ahead
Gupta and Pierdzioch (2022)	Daily	Monthly	1-, 3-, 6-, 12-, 24-months-ahead
Plakandaras et al. (2017)	Monthly	Daily	1-month-ahead
Higashide et al. (2021)	Intradaily 5-min	Daily	1-day-ahead
Oliveira et al. (2017)	Intradaily (unspecified)	Annualized daily	1-day-ahead
Gupta et al. (2021)	Daily	Monthly	1-, 3-, 6-, 12-months-ahead
Grigoryeva et al. (2014)	Intradaily 6-min	Daily	1-up to 20-days-ahead
Prasad et al. (2022)	Daily and weekly	Daily and weekly	1-day- and 1-week -ahead
Persio et al. (2021)	Daily	Monthly	1-day-ahead
Lu et al. (2022)	Daily	Monthly	1-, 3-, 6-, 9-, 12-months-ahead
Qiu (2021)	Intradaily 5-min	Daily	1-, 7-, 14- and 30-days-ahead
Bucci (2020)	Daily	Monthly	1- and 5-months-ahead
Chen and Hu (2022)	Daily	Daily	5- and 10-days-ahead
Zhang, He et al. (2022)	Daily	Monthly	1-, 3-, 6-, 12-months-ahead
Zhang, Wahab et al. (2022)	Daily	Monthly	1-, 3-, 6-, 12-months-ahead
Ghosh and Sanyal (2021)	Daily	Daily	1-day-ahead
Osterrieder et al. (2020)	Intradaily 1-min	Seconds	1 min-ahead
Kristjanpoller and Minutolo (2015)	Daily	Daily	14-, 21-, 28-days-ahead
Ribeiro et al. (2021)	Intradaily 5-min	Daily	1-, 5- and 21- days-ahead
Vidal and Kristjanpoller (2020)	Daily	14-day	14-days-ahead
Kim and Won (2018)	Daily	Daily	1-, 14- and 21- days-ahead
Zolfaghari and Gholami (2021)	Intradaily (unspecified)	Daily	1-, 10-, 15-, 20-, 30-, 60-days-ahead

Model Evaluation

Several different techniques were used for testing and evaluation. First, a model should be tested with an out-of-sample forecast if its predictive performance is to be assessed. This can be achieved through different means, where the most popular methods were fixed, rolling and recursive window regressions. 17 of the studies used a rolling window evaluation scheme, eight applied a recursive window and eleven employed fixed size windows. Some used multiple methods for testing their model performance, allowing them to test the robustness of their results across different evaluation and estimation methods. Additionally, several studies used different window sizes when using recursive and rolling windows. In particular, Gupta et al. (2021) used five window sizes in increments of 20 between 60 and 140 daily observations.

Another facet of the evaluation procedure involves the choice of error statistics. The potential use of statistical tests comprises an additional factor in evaluation methods, where they can be used to assess the relative statistical significance of different forecasts and models, as well as the significance of the forecast compared with the target. In total, 37 different error metrics were found in the papers. The breakdown of the different metrics can be found in Fig. 10. For regression problems, the most prevalent metrics were MSE and MAE. However, some of the metrics include only minor differences, e.g., MSE and MSFE; the former is regular mean squared error, while the latter is mean squared forecasting error. The classification problems – those forecasting the direction of the target – were relatively few in numbers compared to the regression problems, but relied mainly on accuracy and F1 score for model evaluation. For the testing measures used, ten papers did not use any tests in their study. The remaining 22 papers used a total of eleven different tests, showcased in Fig. 11. The Diebold–Mariano, Clark–West and Model Confidence Set constituted the majority of applied tests by a large margin.

As evident from Figs. 10 and 11 there is variation as to how model performance is assessed. A plausible explanation could be the fact that there is currently no generally accepted standard for forecast evaluation in every possible scenario, as thoroughly discussed in Hewamalage et al. (2023). We note that MSE, and variations thereof, are frequently applied. This is reasonable, as parameters of machine learning algorithms typically are tuned using MSE in training and validation datasets. For volatility models trained on L1 loss, MAE might be more appropriate. As for statistical tests of forecasting accuracy, the Diebold and Mariano (1995) and Clark and West (2007) are the most often used. Both are pairwise tests of predictive accuracy; the DM test evaluates two competing forecasts, whereas CM specifically tests for equal predictability of nested models. The Model Confidence Set of Hansen, Lunde et al. (2011) serves a slightly different purpose, as it from a set of models selects a subset of best performing models whose statistical predictability cannot be distinguished from each other.

Alternative model evaluation schemes

To assess model robustness, it can be useful to inspect its performance under different specifications and conditions. Separating between "good" and "bad" volatility – volatility driven by positive and negative returns, respectively – can provide valuable information about the model's practical implications, as well as its ability to handle the oft present leverage effect in volatility series. Furthermore, the model's performance during different levels of financial stress might be particularly intriguing. A model that performs well during relatively calm periods, but fails to produce adequate results during turbulent times can be less appealing to practitioners than one better with better performance in the more volatile regimes. These considerations are not limited to periods defined by their volatility levels; they can also be applied to different business cycles. To summarize, including alternate forecasting evaluations and period specifications can help inform researchers about the robustness of their results.

In total, ten of the papers included some kind of alternative evaluation scheme in their methods. Bouri et al. (2020), Çepni et al. (2022) and Gupta et al. (2021) tested their selection of models' performance on good and bad volatility, and generally found results corroborating their main findings. Çepni et al. (2022) also joined (Christensen et al., 2022; Lu et al., 2022; Petrozziello et al., 2022) in applying different volatility regimes in their testing, e.g., low and high regimes. A general trend among those comparing AI and ML with econometric models were the tendency of the latter to struggle more with high-volatility regimes, especially when compared against the NN family of methods and ensemble methods. Along a similar vein, Petrozziello et al. (2022) and Lu et al. (2022) also compared model performance during pre- and in-crisis periods, e.g., before and during Covid-19 or the financial crisis of 2008. The results did not differ dramatically compared to their general tests. Tissaoui et al. (2022) compared model performance and variable importance before and during Covid-19 to assess different drivers of volatility, and found that different XGB performed better during the more turbulent period while SVM fared better for the more tranquil decade leading up til 2020. Bucci (2020) used a subsample of his data to assess the robustness of the overall results, finding generally equivalent results. Next, Chen and Hu (2022) looked at how models fared in different markets, testing them on both the CSI 300 and the S&P 500. Along with Zhang, He et al. (2022) and Lu et al. (2022) used an additional alternate evaluation scheme in the shape of business cycles. Lu et al. (2022) achieved good results with ensemble methods and NNs for both specifications, while the econometric models performed well in recessions. Zhang, He et al. (2022) tested their aligned GEPU model, and found that it performed better during expansions, while the more general AR models were better suited for the recessions.

4. Discussion of important findings

This section contains our most important findings. Section 4.1 presents typical sources of motivation in the current literature. Section 4.2 discusses the main research questions addressed, which we broadly group into (i) examining the effect of exogenous variables, (ii) comparing hybrid ML models to econometric models, and (iii) general model assessment. Section 4.3 discusses comparisons of machine learning and econometric models in further detail. Section 4.4 elaborates on the use of XAI to enhance explainability of volatility forecasts. Tables 3 and 4 contain an overview of the papers constituting or sample.

4.1. Overall motivation

Volatility prediction is of great interest due to its many implications on financial activities. These activities are the source of motivations for many of the researchers in the selected sample to research this area. Some of the most common ones are risk and portfolio management, derivatives pricing (Chen & Hu, 2022; Christensen et al., 2022; Luong & Dokuchaev, 2018; Yao et al., 2017), and more specifically various active portfolio immunization strategies (Vrontos et al., 2021). Some find volatility prediction a interesting and challenging task for newly introduced statistical and computational frameworks such as reservoir computing (Grigoryeva et al., 2014).

Oil is the most actively traded commodity in the world, consequently, many of these papers have their scope on that asset. Tissaoui et al. (2022) address the issues regarding this very volatile commodity because of the high risk it involves when small changes in the crude oil prices can damage national economies. Some of the more recent crisis such as Covid-19 have also sparked an interest in alternative underlying assets for diversification, where the impact of economic policy in the US, as an important player in the international market, on the US oil-price returns volatility is of interest (Çepni et al., 2022).

Another such alternative asset is Bitcoin, which has also motivated the need for better forecasting methods of volatility for managing Bitcoin positions, as well as investigate drivers of Bitcoin Gkillas et al. (2021). Lastly, Bouri et al. (2020) points out anecdotal evidence of Bitcoin acting as a flight to safety, or an asset independent of market risk. They are further motivated to investigate this phenomenon while considering the effect of US–China trade war and Bitcoin volatility.

4.2. Main research questions

To provide an outline of the sample and its characteristics, Table 3 presents a summarized view along the portrayed dimensions. As can be inferred from the table, the final sample includes a plethora of different models, underlying assets and forecasting targets. Hence, this section will provide a short overview over each paper by grouping them based on their main goal: examining effect of an exogenous variable, using a hybrid ML and econometric model, or more generally, assessing the forecasting performance of different models. Albeit several of the papers can fall into multiple of these categories, we have made this distinction based on how they relate to the two first criteria. Hence, those with a main approach not fitting either of those are put into the overall comparison category.

Examining Effect of Exogenous Variables

Several papers examine the overall effect from certain predictors on overall volatility forecasting. Generally, several model modifications are performed, wherein the data of interest is excluded in some and included in others. Thus, the authors can examine the impact the inclusion of the given data has on modeling performance. Çepni et al. (2022) investigated the effect of using uncertainty-related data on different levels of granularity for forecasting the realized variance of WTI oil price returns. They chose to use the US as a proxy for the global market, and constructed both nationwide and disaggregated state-level

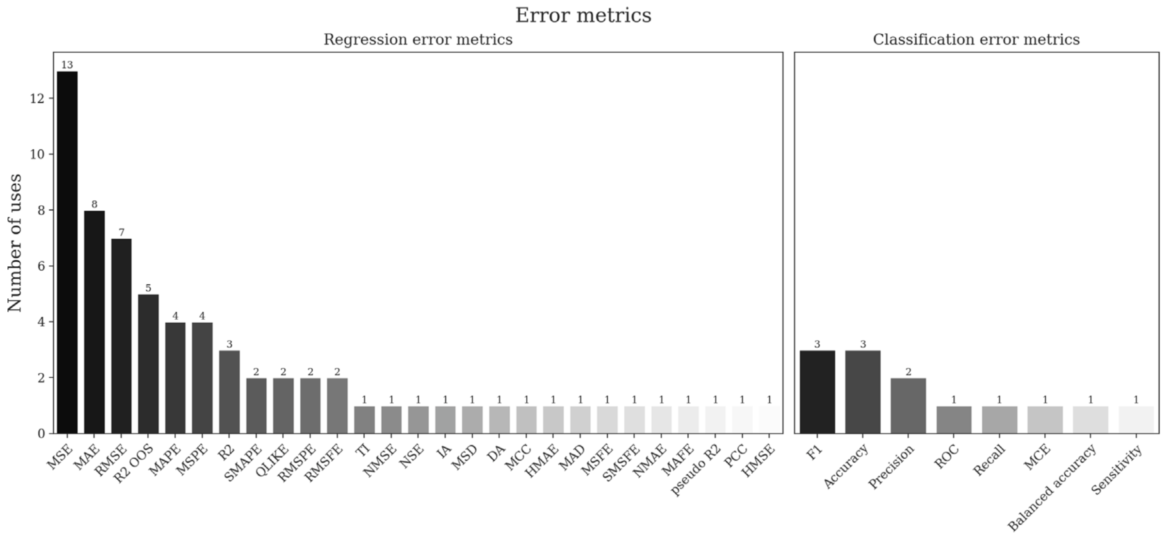

Fig. 10. The different error metrics used in the sample, grouped by the type of forecasting problem. See Appendix D for abbreviations.

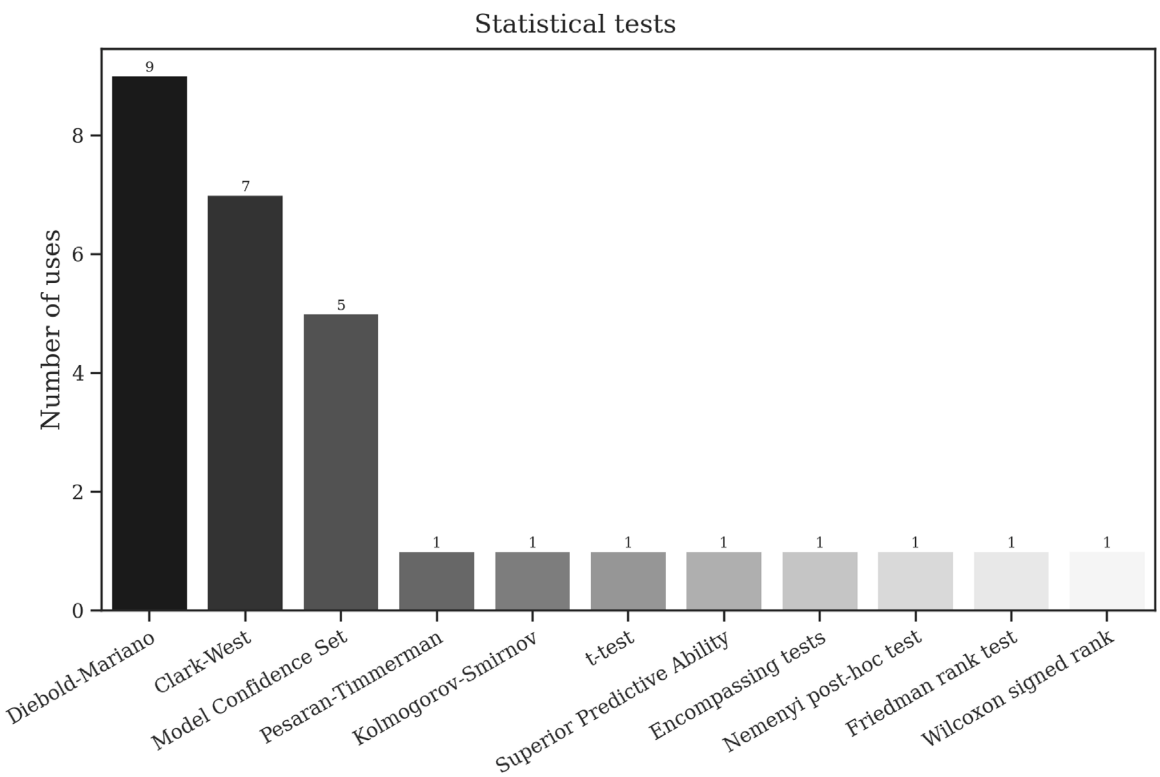

Fig. 11. The different testing measures utilized in the sample. Their major area of use was to compare models and forecasting results.

US uncertainty indices to test their impact on forecasting performance. As the uncertainty data was monthly, and the realized variance was based on daily log returns, they aggregated the daily data into monthly realized variance. A simple AR model estimated using OLS was used as the baseline, and predictors were sequentially added to assess the effect of including more information. To tackle the resulting increased model complexity, they resolved to use RF and LASSO in estimating the realized variance. Using the R^2 out-of-sample statistic for comparing models, they found that including uncertainty-related measures tended to increase the performance for longer forecasting horizons.

Gupta and Pierdzioch (2022) followed the approach of Çepni et al. (2022) to investigate the role of uncertainty and geopolitical risk in the realized volatility of oil price returns. Using the same methods as Çepni et al. (2022), they included the addition of a geopolitical risk index and an equity-market volatility tracker, both in an aggregated and disaggregated version. Similarly to Çepni et al. (2022), they found that the disaggregated metrics improves forecasting performance for longer horizons, particularly when using random forest as the estimation technique. These results might indicate that disaggregated uncertainty-related indices are useful for predicting long-run volatility.

The effect of an aligned global economic policy uncertainty (GEPU) index on forecasting monthly realized variances of WTI oil price returns are investigated by Zhang, He et al. (2022). They used a modified PLS (Partial Least Squares) approach to construct an aligned GEPU index targeted on the residuals of a simple AR model. Next, they compared the model with the benchmark AR model without uncertainty data. Additionally, they included other variants of the GEPU model to compare different approaches, and perform an in-sample comparison between exogenous economic variables and the GEPU term. They found that, for the in-sample results, the GEPU term provided complementary information not covered by existing GEPU and macroeconomical data. Further, the out-of-sample R^2 statistic of their resulting models shows that the GEPU-PLS model significantly increase the forecasting performance in both shorter and longer forecasting horizons. Finally, they implemented robustness checks and different trading strategies to assess the robustness and economic significance of their findings, and concluded that the GEPU-PLS model can provide substantial benefits for forecasting realized variance of oil prices.

Impact of uncertainty related to the US–China trade war in relation to Bitcoin volatility is of interest in the study from Bouri et al. (2020), where they used hourly log-returns to derive the daily realized volatility, and extract Google search volume intensity associated with the US–China trade war to capture the uncertainty aspect. As a baseline, they used the HAR-RV model, but included additional variables to account for different aspects related to jumps, leverage effects, as well as realized skewness and realized kurtosis. The model was estimated using random forest, and by comparing models with and without the uncertainty metrics they checked whether it improved the forecasting performance of a given model. Through a Diebold–Mariano test, results showed that the inclusion of trade uncertainty improved overall forecasting performance, especially for longer forecasting horizons.

With a special emphasis on determining whether the number of transactions can be of significance in forecasting, Gkillas et al. (2021) attempted to forecast daily Bitcoin realized volatility. Daily realized volatility was estimated using hourly returns, and total number of Bitcoin transactions were used as a proxy for transaction activity. To model the volatility, they utilized a random forest and HAR-RV-Jump hybrid to assess each model's performance. Through a DM-test and CW-test, they found that accounting for number of transactions improved the out-of-sample forecasting accuracy, especially when increasing the horizon of the forecast. The same results applied to the hybrid model against a regular HAR-RV model.

Plakandaras et al. (2017) used SVR variants to determine if EPU related to Brexit has an effect on forecasting the USD/GBP exchange rate realized volatility. They used both daily and monthly observations to derive realized volatility, and combined this with the logarithmic difference of the US and Brexit EPU indices. Starting with an AR model, they included different variations of the EPU data as extensions to the model, and compared OLS with four different SVR kernels to estimate the realized volatility. They observed that the models using the Brexit EPU estimated with SVR tended to produce the lowest RMSE on both a daily and monthly frequency. Additionally, they used a Quandt-test to prove that Brexit was a structural break, and demonstrated that a post-Brexit AR model estimated using OLS performs better than an otherwise equivalent model spanning the entire period.

Twitter data with sentiment values to deduce whether they have an effect on forecasting VIX and the annualized RV on the S&P 500, RSL, DJIA and NDQ was proposed by Oliveira et al. (2017). Furthermore, they compared several ML models against each other and an AR(5) benchmark model, and with a DM-test for comparing forecasting results, found that the usage of sentiment and attention indicators did not improve the forecasting performance on a statistical significant level.

In an attempt to forecast realized volatility of DIJA on a monthly basis, Gupta et al. (2021) suggested a random forest model as well as a HAR-hybrid model. They wanted to investigate the impact on forecasting accuracy when including investor sentiment indices and surveys. Using an out-of-sample R^2-statistic in tandem with a CW-test, they found that combining investor sentiment related data in addition to macroeconomical and financial data improved the forecasting performance of their model for shorter and intermediate forecasting horizons.

Prasad et al. (2022) examined the impact of including macroeconomical variables when predicting the direction of the VIX. Using logistic regression, LGBM and XGB, they found that the two latter models with the exogenous data exhibited statistically significant results through a t-test on the MCC, in both a weekly and daily forecasting horizon.

In a similar directional approach, Higashide et al. (2021) forecasted the aggregated Nikkei realized volatility direction using a random forest and HAR hybrid model. Their main focus was the inclusion of the TSE Co-location dataset in improving volatility forecasting. They compare both a HAR using a logistic function and their hybrid model, and find that including the data improves the F-measure the most for the hybrid variant. Additionally, the data appears to be more important when forecasting over a longer horizon.

Hybrid ML and Econometric Models

In recent years, there has been an increase in the number of papers combining ML and econometric approaches to forecast volatility. By combining the two approaches, one might be able to exploit the strengths of both to achieve a better forecasting performance. Yao et al. (2017) used a two-component short-term/long-term volatility approach to forecast the exchange rate realized volatility of different currencies through a hybrid ANN-AR model. They extracted the short- and long-term volatility using a Hodrick–Prescott filter, and used the latter as additional input to the autoregressive neural network. Then, the short-term volatility was estimated using an AR model. When comparing the performance of this model against an EGARCH and two-component GARCH, their proposed approach outperformed the traditional models on both an hourly and daily-based horizon for almost all scenarios.

Liu et al. (2018) used RNNs and combined them with a HAR-RV-Jump model to forecast the S&P 500, and compared the results of all models using a set of error measures. They found that the RNN appeared to produce better results than HAR with a smaller amount of data. When they increased the size of the training sample, their performance equalized. However, the hybrid model performed worse than the non-hybrid version for regular forecasting, hence, they conclude that there is no need for a hybrid model. Overall, they infer that RNN seems preferable when there is less available data.

In an attempt to capture both linear and nonlinear data generation processes (Qiu, 2021) proposed a method for combining forecasts based on complete subset least squares vector regressions (LSSVR^{CS}) as an estimator for the HAR-RV model. In a Monte Carlo simulation experiment, his LSSVR^{CS} outperforms many other competing estimators, while in an empirical forecasting exercise of Bitcoin realized volatility, the LSSVR^{CS} consistently achieved the best 1-, 7-, 14-, and 30-days out-of-sample forecasting performance.

A hybrid LSTM-GARCH(1,1) model is presented by Persio et al. (2021) to forecast the monthly S&P realized volatility. Their main task revolves around using the VolTarget strategy to assess its performance when compared to using regular historical volatility. Using their hybrid approach, they find that it results in a higher average return compared to the standard VolTarget strategy. Similarly, Kim and Won (2018) combined multiple GARCH-models with an LSTM. Specifically, they tried to combine LSTM and a deep feed-forward neural network with a GARCH, EGARCH and EWMA model while forecasting the realized volatility of the KOSPI 200 index. They concluded that combining two or more GARCH-models with an LSTM provided the best results, and through a DM- and WS-test proved their results to be statistically significant.

Ribeiro et al. (2021) used a combined HAR-PSO-ESN hybrid in order to forecast daily realized volatility of three Nasdaq stocks. They compared their model against multiple machine learning and econometric models, and found that the HAR-PSO-EN gave statistically significant better results for three different horizons. Furthermore, Zolfaghari and Gholami (2021) compared an AWT-LSTM and HAR-RV hybrid against an ANN and HAR-RV hybrid. They found that the AWT-LSTM-HAR hybrid gave better results with regards to their evaluation metrics.

Different machine learning models compared with HAR-RV and GARCH in predicting the SSE Composite Index realized volatility are studied by Song et al. (2022), where they further tested the effectiveness of including the short-term volatility found by a GARCH-MIDAS model. They found that LSTM, GRU and XGB produced the lowest errors according to their measures, and both deep learning methods benefited from including the GARCH-MIDAS volatility. Results from a DM-test revealed that GRU had the overall best forecasting performance among the models.

Kristjanpoller and Minutolo (2015) used a combined ANN-GARCH model to forecast gold spot and future price realized volatility. They found that including a GARCH forecast as input to an ANN, in addition to other macroeconomic variables, improved its forecasting performance measures over a regular GARCH approach. Also within gold price forecasting, Vidal and Kristjanpoller (2020) combined a CNN with a LSTM in order to include images of the realized volatility time series. They use this to forecast the LBMA gold price realized volatility, and compare their approach with several econometric and ML models, including an ANN-GARCH hybrid. Using both a statistical and performance related error measurement, they found that their proposed model outperforms all benchmark models.

General Model Assessment

In the final category we place papers where the main task was concerned with investigating a given model, and their method did not fit with the two aforementioned criteria. Tissaoui et al. (2022) used SVR and XGB to forecast the OVX, and compared their models against an ARIMAX model. Their out-of-sample results implied that XGB and SVR provided superior forecasting performance compared to ARIMAX. Additionally, they found that investor uncertainty related measures were important predictors for forecasting the OVX.

Lu et al. (2022) used machine learning models to predict the monthly realized volatility of WTI futures. They compared the machine learning models against several AR(3) models with different extensions using the out-of-sample R^2 and MSPE, and found that particularly random forest, boosting, ANN-related and ensemble models had relatively better performance in predicting the realized volatility of oil futures. Zhang, Wahab et al. (2022) also tried to forecast monthly realized volatility of WTI futures by using supervised PCA with variable selection. They compared the model against an AR, LASSO, ENet and kitchen sink model variant and found that their proposed approach outperformed all other models both in-sample and out-of-sample. Additionally, they used a portfolio exercise, see Bollerslev et al. (2018), to evaluate each models economic performance, and found that the PCA-VS model delivers the greatest average realized utility.

Christensen et al. (2022) studied a broad range of machine learning and HAR-RV models to forecast the realized volatility of several DIJA constituents. For shorter forecasting horizons, they found machine learning models capturing non-linearities to be superior. When increasing the forecasting horizon, random forest and ANN proved to be the best models for approximating a long-term structure in realized volatility. Finally, they used a VaR analysis to examine the economical utility of each model, and concluded that the basic HAR was hard to beat in this framework. However, the difference in performance between HAR-RV and the machine learning models was not substantial. Including more exogenous variables let the machine learning models deliver more precise measures for VaR.

To predict the realized volatility of several NYSE assets, Grigoryeva et al. (2014) examined reservoir computing ANNs and compared them against a VEC(1,1)-model. They found that the machine learning approach results in systematically better results compared to the VEC-GARCH model by the MSFE metric. Vrontos et al. (2021) compared several different machine learning models in predicting the direction of the VIX index. They compared the machine learning models against various logistic regression models, and based on numerous statistical evaluation measures found that the machine learning approach produced better results for out-of-sample forecasts.

Bucci (2020) examined the performance of neural networks in predicting the S&P 500 monthly realized volatility. They compared regular feed-forward NNs, ENN, JNN, LSTM and NARX-SP against ARFIMA, LSTAR and HAR-RV. Additionally, they included variants with and without exogenous macroeconomical and financial variables to further examine the resulting forecasting performance. They concluded that NNs, particularly LSTM and NAR, outperformed classical methods. However, the HAR model also provided good results, especially when increasing the forecasting horizon. When comparing against a benchmarking random walk, the HAR and NARX outperformed the other models by the most significant margin.

Utilizing ANN and LSTM using different preprocessing methods (Chen & Hu, 2022) compared different models for forecasting the realized volatility of both CSI 300 and S&P 500 futures. These were compared against a benchmark AR and EGARCH model with respect to different loss functions. The best overall models was the ANN and LSTM, both using PCA, and LSTM performed better for more statistically complex data.

Ghosh and Sanyal (2021) examined the predictability of market fear in India during the Covid pandemic through the India VIX. They combined the Boruta algorithm for variable selection with XGB, ERT, DNN and LSTM, and claimed they are successful in precise estimation of future India VIX. Further, through an MCS test procedure they conclude that XGB provides the best forecasting performance out of the selected models. Using option data, Osterrieder et al. (2020) utilized NNs to replicate the CBOE VIX, and consequently check for arbitrage opportunities. By training an LSTM model on S&P 500 options, they were able to outperform a naive approach using the previous lagged VIX value as its forecast and a random forest model. They specified how this method, by utilizing the intradaily deviations between VIX and its futures, might be used to find arbitrage opportunities.

Finally, Petrozziello et al. (2022) compared R-GARCH and GJR-MEM against a univariate and multivariate LSTM for forecasting the realized volatility of stocks from NASDAQ and DIJA. They found that LSTM had better forecasting performance during periods of high volatility, and comparable during more tranquil periods. They concluded that a multivariate LSTM better captures the spillover effects between assets, thereby its improved performance when forecasting volatility for multiple assets.

4.3. Comparisons of machine learning and econometric models

The majority (26) of the papers reviewed use realized volatility as an estimator or a proxy for the true volatility (see Appendix A). The minority of papers conduct a directional approach where the interest lies in predicting the direction of the volatility the next time-period (day, week, month). The more common approach is a regression problem, for which the HAR-RV model is a commonly used benchmark. Many findings suggest that this model is a quite good short-term prediction model (Christensen et al., 2022; Gkillas et al., 2021), some also report good results for long-term forecasting (Bucci, 2020). This is consistent with the conclusions of Corsi (2009) when proposing the model as a good performing long-memory model, despite its simple and parsimonious nature. When comparing the predictability of models, it is common to compare proposed models to well-known benchmarks.

Jia and Yang (2021) criticize the approach where machine learning methods are directly compared to GARCH-like econometric models, as it introduces two disadvantages. This approach is similar to many of the discussed papers, where historical data is used as input to forecast volatility, then the machine learning models are trained using distance based loss functions to obtain an optimized solution to forecast realized volatility. These forecasts are then compared point-to-point to the observed realized volatility. The first of the criticized problems is that volatility is always changing and a latent variable, hence, realized volatility is still just a statistical estimation of the true volatility. In other words, the machine learning methods are trained to predict volatility with errors. The second problem is the way econometric models estimate their parameters by maximizing the likelihood differs from the distance-based optimization in the machine learning approach. The out of sample evaluation using distance-based performance criteria is therefore only consistent with how the machine learning methods are trained.

To address the first issue, it is important to have in mind that since volatility is latent and not observable, one will never have perfect estimation techniques of the true volatility unless under ideal theoretical assumptions. These theoretical assumptions will arguably rarely happen in real world applications. Thus, we will never have perfect prediction or forecasting models which perfectly fit the true volatility. What might then be even more important is how the models and their predictions will help us reach a given end goal in an applied situation. This entails evaluating models from an economic point of view, measuring its performance through trading strategies and assessing the results. As outlined in Table 3, eight papers included such an analysis, albeit through different means. Petrozziello et al. (2022) and Christensen et al. (2022) utilized the VaR or CVaR framework for their analysis, rooted in their use of multiple assets, where the latter proved that even untuned nonlinear machine learning algorithms were able to give good VaR estimates. Meanwhile, Çepni et al. (2022), Liu et al. (2018), Vrontos et al. (2021) and Zhang, He et al. (2022) based their evaluation on a more direct trading strategy, either using the historical data, or through the use of Monte Carlo simulations. Finally, Persio et al. (2021) utilized the VolTarget strategy to assess the economic implications of their model. In particular, the trading strategy experiments of Liu et al. (2018) demonstrates the necessity of making an economical evaluation. RNN provided better statistical error measures than a benchmark HAR-RV jump model for most of their forecasting experiments, but in terms of financial performance no model clearly dominated the other. Thus, evaluating from an economical perspective might yield different results, and can help assess the practical implication of a given model.

Regarding the second issue, it will mainly be applicable when directly comparing to general GARCH-model which implicitly uses squared returns as a proxy for the volatility, especially in a high-frequency setting. Andersen et al. (2001b) describe this as GARCH having an inability to adapt to high frequency data. They emphasize that this does not reflect a failure of the GARCH-model, but rather the efficacy of exploiting the volatility. The slowly decaying weighted moving averages of past squared returns adapts only gradually to volatility movements, opposed to machine learning methods trained with a realized volatility proxy. Additionally, some of the machine learning models will also be able to benefit from the ability to account for non-linear features in the realized volatility. This confirms a rather unfair comparison for short horizon forecasting in high-frequency environments. Yao et al. (2017) do one-hour- and one-day ahead forecasts of realized volatility comparing a autoregressive neural network enhanced two-component GARCH-models to traditional GARCH and a two-component GARCH. This relatively short forecasting horizon in combination with the high frequency might impose these comparable issues, as the results shows even larger difference in RMSE for the shortest horizon in favor of the hybrid model built directly on realized volatility. However, in the reviewed papers, the more common approach is to compare with benchmark models which are built directly for the realized volatility proxy.

Frequency of Data and Forecasting Horizon

There are several issues related to microstructure problems that can arise when estimating realized volatility. One problem is the presence of transaction costs and bid–ask spreads, which can affect the prices of assets and introduce estimation errors. Another problem is the impact of market microstructure events, such as order imbalances, order cancellations and trading halts, which can also introduce bias. Consequently, when realized volatility is implemented in practice, the price process is often sampled sparsely to strike a balance between increased accuracy from using higher frequency data and the adverse effect of microstructure (Liu et al., 2015).

We can see from Fig. 9 that only 14 of the studies use intradaily observations. Within these, 5 min intervals are the most common used frequency. Liu et al. (2015) find little evidence that the 5 min realized volatility estimate is outperformed by other measures, which suggests its popularity. The high sampling frequency gives high accuracy in the volatility estimation and is appropriate for capturing the short-term dynamics of volatility. AI and machine learning models that are able to capture the additional information provided by the intradaily sampling will then be able to make more accurate forecasts in the short term horizons. This was demonstrated by Christensen et al. (2022), Higashide et al. (2021) and Liu et al. (2018) in their 1-day-ahead forecasts.

When considering longer forecasting horizons, the inclusion of exogenous data seems increasingly important. Song et al. (2022) demonstrated this by combining GARCH-MIDAS and a GRU model, and obtaining a significant reduction in the average MSE for 5-day ahead forecasting. As the GARCH-MIDAS exploited macroeconomic data, this result is twofold: it highlights the potential of hybrid models, and the benefits of exogenous data when increasing the forecasting horizon. Another illustration was done in Christensen et al. (2022), were longer forecasting horizons yielded significant benefits for the NN and RF models incorporating exogenous data. Finally, when comparing with an AR benchmark, Çepni et al. (2022), Gupta et al. (2021), Gupta and Pierdzioch (2022), Zhang, He et al. (2022) and Zhang, Wahab et al. (2022) all found that the models including additional data improved in tandem with an increasing forecasting horizon. These findings align with the general notion that exogenous data can assist in uncovering nonlinear long-term interaction effects between different predictors for volatility.

However, in spite of the results favoring realized volatility the aforementioned problems remains a nuisance. With the micro-structural problems in mind, the accuracy of the realized volatility as an estimator of the true volatility is dependent on the sampling frequency, forecasting horizon, and the liquidity of the underlying asset. Thus, when forecasting volatility, the researcher needs to make decisions regarding the sampling frequency, aggregation method, filtering and error handling. These effects becomes more pronounced as the sampling frequency increases, further complicating the forecasting process. Consequently, there might be a heightened risk of spurious estimation results, thereby weakening any proposed model built upon this estimator. Therefore, it is not consistent like another volatility proxy such as an implied volatility index.

Implied Volatility Indices

Out of the reviewed papers which did not use realized volatility as a proxy, an implied volatility index is used instead. Here, none of the papers use autoregressive conditional heteroskedastic models as benchmarks. This is due to the fact that it is directly observed and not an estimation based on returns like realized volatility. Thus, it would not make sense to use such conditional models. Another important aspect is the fact that implied volatility is often larger than the volatility obtained using a GARCH-model. This could either be because of the risk premium for volatility or the way daily returns are calculated (Tsay, 2010). The proposed benchmark models used in forecasting volatility indices are ARIMAX, an ARIMA model with additional explanatory variables, as well as AR-models. We observe less incentives to compare machine learning models with econometric models when using an implied volatility index as a proxy, see Ghosh and Sanyal (2021), Prasad et al. (2022) and Vrontos et al. (2021). Some use the naive prediction which consists of just using the current VIX value as the forecast for the next time-step, see for instance Osterrieder et al. (2020).

Tsay (2010) points out the unobservability of volatility which makes it difficult to evaluate the forecasting performance of conditional heteroskedastic models. Since implied volatility indices are directly observed, an accurate prediction of such index could then introduce arbitrage opportunities, as proposed by Osterrieder et al. (2020). Directional approaches for both implied volatility index- and realized volatility prediction also provide more intuitive performance measures, however these results also depend on how they are used. This also points back to our previous points in which it is important how the models and their predictions will help us reach a given end goal in an applied situation. The VIX can be especially attractive in this regard, since the VIX futures allows for direct trading evaluation of a forecast.

In total, six papers studied forecasting of an implied volatility index, where Oliveira et al. (2017) concurrently predicted the realized volatility of some stocks. Thus, a mere five papers concentrated their efforts on investigating the use of ML and AI for implied volatility forecasting. In light of the findings of Dai et al. (2020), implied volatility might have significant predictive power of stock return volatility. In the same vein, Christensen et al. (2022) found the VIX to be among the top ranked variables from their variable importance analysis. Additionally, Tissaoui et al. (2022) and Zhang, Wahab et al. (2022) found VIX to be one of the most important predictors for forecasting the OVX and WTI monthly realized variance, stressing its efficacy as an oil volatility predictor. However, the general literature remains inconclusive on its predictive power, see e.g. Becker et al. (2007). Thus, the paucity of this subject underlines a need for more research in this area, particularly compared against realized measures.

4.4. Use of XAI and explainability of models

With the advent of AI and ML methods as mainstream techniques in a plethora of disciplines, there has been some concerns regarding their rising popularity. Many AI methods can act as black boxes. Interpreting the models can be challenging, and it may be difficult to make inferences about the underlying interaction of features. Thus, the practicality of AI's ability to capture nonlinear relationships is somewhat hampered by their black-box nature. Neural networks and its family of models are particularly prone to this effect. Despite yielding superior results, Bucci (2020) highlights some of the disadvantages with their use of neural networks. The number of parameters to be estimated is substantially larger than its econometric counterparts, and the interpretation of causal relationships between variables cannot be easily performed. In tandem with the establishment of practical AI guidelines from several governmental institution, the need for explainability has become increasingly important. This has motivated the use of the Explainable Artificial Intelligence (XAI) framework to shed light on the underlying interactions of AI models.

In our sample, four papers employed XAI methods to analyze their findings. This was done using the Shapley framework and accumulated local effect, allowing the authors to assess the relative impact of each variable when for model predictions. Tissaoui et al. (2022) utilized Shapley values to determine the feature importance of their variables in predicting the OVX during two different regimes – before and during the Covid-19 pandemic – and found that VIX had the most impact on their predictions during both periods, but EPU played a larger part under Covid-19. Further, Christensen et al. (2022) employed the accumulated local effect to assess variable importance, and discovered that the ML models utilized implied volatility to a higher degree than its econometric counterparts. Lu et al. (2022) found that one lag of the monthly realized variance of WTI futures dominated the other variables with respect to the variable importance, but other variables provided significant contributions. Ghosh and Sanyal (2021) used both the Shapley and Lime framework in their forecasting experiments, and found that lagged values of the India VIX had the highest importance. However, many other variables played a large role in predictions, emphasizing a need for exogenous data when forecasting implied volatility.

As mentioned earlier, the use of exogenous data is especially useful for longer horizons. However adding more variables increases the complexity of the models. This makes the models difficult to interpret, which can make it challenging to understand the relationships between the variables and draw meaningful conclusions. Another consequence of too many additional variables is that it can lead to unreliable results, that is, when small changes in the data can have a significant impact on the model's performance. This further motivates future applications of XAI within AI and machine learning models as it can be necessary to identify key drivers of the volatility. This is of particular relevance when volatility forecasts are interpreted as early warnings of increased systemic risk, including contagion in financial markets. Additionally, as mentioned above, influence from institutions such as European Commission and Directorate-General for Communications Networks, Content and Technology (2019) will continue to promote further research and development in the field of XAI.

5. Conclusions

In this study we perform a systematic review of realized and implied volatility forecasting using artificial intelligence and machine learning methods. More specifically, we investigate how these models fare against the more traditional econometric approaches, the motivation for forecasting volatility, the markets under study, and whether AI and machine learning actually lends itself towards volatility forecasting.

We perform an extensive search through four major databases in scientific literature, and consequently exclude papers not fitting our topics, ending up with a final sample of 32 papers. We summarize their methods and findings, and provide descriptive statistics about the sample. Here, we find that the majority of the markets investigated are the US equity market, followed by oil and foreign exchange markets. Furthermore, we find the use of exogenous data valuable for the forecasting exercise, especially for longer forecasting horizons. In particular, the inclusion of predictors handling different aspects of uncertainty through newspaper-based indices yields favorable results. With the potential of extracting high-frequent public sentiment data from the web, this encourages more research into uncovering potential contemporaneous and long-term links to volatility.

Generally, we find the efficacy of AI and ML methods for volatility prediction to be highly promising, often providing comparative or better results than their econometric counterparts. Their capability of handling non-linear relationships, correlated predictors and large volumes of exogenous variables make them attractive options for future research within volatility forecasting. Neural networks employing memory consistently ranked among the top performers, where LSTM, GRU and reservoir computing proved apt in their ability to forecast volatility. Additionally, tree-based methods like boosting and RF were popular models, being able to handle large amount of external data to good results. However, traditional econometric models such as the HAR-RV are still highly relevant, commonly yielding similar results as more advanced ML and AI models. In light of the success with ensemble methods, a promising area of research is the use of hybrid models, as they have been shown to beat its constituting models on many occasions.

We also discuss some issues regarding comparison of models based on the volatility proxy of choice. Models are often designed around a specific process, and failing to account for this can provide unfair results. Furthermore, we discuss the issue of volatility and its latent nature, and how this complicates the modeling comparison process. This prompts us to recommend more focus be directed towards alternative evaluation methods, either by the way of robustness checks or through practical applications such as trading strategies. We also suggest that more focus should be aimed at investigating implied volatility forecasting using AI and ML, as the current literature to a large extent deals with forecasting realized volatility.

Another point of discussion revolves around explainability. ML and AI are particularly susceptible to issues surrounding this concept. This also become an issue when adding variables to econometric models in addition to unstable models. However, the issues could be less troublesome for ML and AI models with future developments within XAI. We find that very few papers included XAI to combat this, and

Table 3
A brief summary of all the featured studies in this paper.

Reference	Target /Proxy	Asset	Extra data	Sampling Horizon	Proxy Freq	Main Method/s	Benchmark	Performance measures	Financial evaluation
Vrontos et al. (2021)	VIX (Dir)	S&P 500	31 macroeconomic, financial market related indicators	01.01.1990 – 01.12.2019	Monthly	Naive Bayes, Ridge deviance, ADABoost, Discriminant analysis	Various LR models	ROC(AUC), MCE, Accuracy, Kappa statistic, Sensitivity, Precision, F1, Balanced accuracy	Annual return, Annual risk, Sharpe ratio, Sortino ratio, Downside risk, Avg. drawdown Alpha, Beta
Petrozziello et al. (2022)	RV (Kern. est.)	SPY, assets from DJIA, NASDAQ 100	None	01.01.2002 – 31.08.2008 and 01.12.2012 – 29.11.2017	Daily	Univariate and multivariate LSTM	R-GARCH, GJR-MEM	MSE, QLIKE, Pearson correlation index, DM	VaR, CVaR
Luong and Dokuchaev (2018)	RV, RV (Dir)	S&P, ASX 200	Converted RV into technical indicators	01.01.2008 – 31.12.2014	Daily	RF and HAR-RV hybrid	HAR-JL estimated using MLE	Accuracy, MAE, MAPE, RMSE, RMSPE	None
Tissaoui et al. (2022)	OVX	USO ETF	VIX, Lags, IDTI, EPC, EPU, GRI (VIX, lags and various uncertainty-related indices)	01.01.2010 – 01.08.2021	Daily	SVR and XGB	ARIMAX	JB, MSE, RMSE, R2	None
Song et al. (2022)	RV	Shanghai composite index	Technical, macroeconomical and financial indicators	26.11.2012 – 20.11.2020	Daily	SVM, RF, XBG, LSTM, GRU hybrids using GARCH-MIDAS as inputs	HAR-RV	KS test, MSE, RMSE, MAE, SMAPE, RMSPE, DM test	None
Christensen et al. (2022)	RV	DJIA constituents	IV, VIX, financial and macroeconomical variables	29.01.2001 – 31.12.2017	Daily	RF and NN	HAR-RV, LogHAR, LevHAR, HARQ	MSE, DM, MCS	VaR
Liu et al. (2018)	RV	S&P 500	SPY, VIX, VXX, ETN (Not used for training, only financial evaluation)	02.01.1996 – 02.06.2016	Daily	RNN and hybrid RNN/HAR-J	HAR-J, LogHAR-J	RMSE, MAE, MAPE	Annual return, Ann, volatility, Sharpe ratio
Çepni et al. (2022)	RV	WTI	State-level and national EPU of US	01.01.1984 – 12.01.2019	Monthly	AR (OLS), Lasso, RF (AR model, but RF/Lasso estimation technique)	AR using OLS without uncertainty variables	R2 for SE and AE, General loss function (Patton 2011)	Accumulated profits of option portfolio
Gkillas et al. (2021)	RV	Bitcoin	Total number of Bitcoin transactions	01.01.2014 – 07.03.2020	Daily	RF/HAR-DUJ hybrid	GARCH(1,1), HAR-RV, HAR-RV-DUJ	CW test, DM, MSPE	None
Yao et al. (2017)	RV	EUR/USD, BBP/EUR, GBP/JPY, GBP/USD FX	None	27.09.2009 – 12.08.2015	Hourly, Daily	NN and Enhanced two-component NN	Two-component GARCH, EGARCH	RMSE	None
Bouri et al. (2020)	RV	Bitcoin	Google trends	01.07.2017 – 30.06.2019	Daily	RF	Various HAR variants	DM, Abs and quad loss for unscaled and scaled forecast errors, Pseudo R2, CW	None
Gupta and Pierdzioch (2022)	RV	WTI	State-level and national EPU of US	01.01.1985 – 01.08.2021	Monthly	AR (OLS, Lasso and RF)	AR without uncertainty	RMSFE, MAFE, CW, DM	None
Plakandaras et al. (2017)	RV	USD/BPD FX	Brexit-related EPU and US-related EPU	01.01.2001 – 08.08.2016	Daily	SVR variants	OLS/SVM with simple AR-term	RMSE	None
Higashide et al. (2021)	RV (Dir)	Nikkei stocks	Market volume data, stock full-board dataset, TSE Co-Location data	01.03.2012 – 31.10.2019	Daily	RF/HAR hybrid	HAR-RV w/ logistic	F-measure	None
Oliveira et al. (2017)	VIX, RV	S&P 500, RSL, DJIA, NDQ	VIX, Twitter data, sentiment values	01.01.2012 – 01.10.2015	Annua-lized daily	MR, NN, SVM, RF, EA (Ensemble avg)	AR(5)	NMAE, DM	None
Gupta et al. (2021)	RV	DJIA	134 macroeconomic and 148 financial variables, AAII investor sentiment survey, 4 investor confidence indices	01.06.2001 – 01.06.2020	Monthly	RF/HAR-RV hybrid	HAR-RV	RMSFE, R2	None
Grigoryeva et al. (2014)	RV	NYSE assets	None	06.01.1999 – 31.12.2008	Daily	Reservoir computing (RNN)	VEC(1,1)	Standardized MSFE	None
Prasad et al. (2022)	VIX (Dir)	S&P 500	Macroeconomical and financial variables	01.05.2007 – 01.12.2021	Daily, Weekly	Logistic regression, LGBM, XGB	None	Accuracy, precision, recall, F1, MCC, t-test	VolTarget
Persio et al. (2021)	RV	S&P 500	None	01.01.2000 – 01.01.2020	Monthly	LSTM/ GARCH(1,1) hybrid	GARCH(1,1)	MSE, MAE, MAPE	None
Lu et al. (2022)	RV	WTI futures	Macroeconomical and financial variables	01.04.1986 – 30.09.2020	Monthly	SVR, KNN, Enet, LASSO, Ridge, RF, ET, Bagging, Adaboost, GBR, FCNN, CNN, C-LSTM	AR(3) using different number of variables	MSPE, R2	None

(*continued on next page*)

Table 3 (*continued*).

Qiu (2021)	RV	Bitcoin	None	01.09.2017 – 20.12.2018	Daily	Complete subset LSSVR, LSSVR	HAR-RV	MSFE, SPA test	None
Bucci (2020)	RV	S&P	Macroeconomical and financial variables	01.02.1950 – 31.12.2017	Monthly	ANN, ENN, JNN, LSTM, NARX-SP	AFIMA, LSTAR, HAR-RV	MSE, QLIKE, MCS, DM-test, Encompassing test	None
Chen and Hu (2022)	RV	CSI300, S&P 500 futures	7 major stock indices in China, and six major stock indices in US	01.01.2011 – 31.12.2018	Daily	ANN, LSTM	AR, EGARCH	MSE, MAE, MNMSE	None
Zhang, He et al. (2022)	RV	WTI	GEPU index, 15 individual EPU indices, 14 economic vars, 6 uncertainty vars	01.01.1985 – 31.12.2019	Monthly	PLS	AR, HAR	R2, OS, CW-test	Avg. return, St. deviation, Sharpe ratio
Zhang, Wahab et al. (2022)	RV	WTI futures	127 macroeconomic variables	01.01.1985 – 31.12.2018	Monthly	Supervised PCA	AR, PCA, LASSO, ENET	R2, MSPE, CW, Newey–West t-stat	None
Ghosh and Sanyal (2021)	India VIX	India VIX	9 macroeconomical and 11 technical indicators, 8 GSVI variables	01.03.2020 – 31.05.2021	Daily	XGB, ERT, DNN, LSTM	None	NSE, IA, TI, DA, MCS	None
Osterrieder et al. (2020)	VIX	VIX	SPX options	02.01.2018 – 02.28.2018	Seconds	LSTM, RF	VIX	MSE	None
Kristjanpoller and Minutolo (2015)	RV	Gold Spot, Gold future	DJI, EUR/USD, USD/Yen FX, FTSE, WTI	06.09.1999 – 20.03.2014	Daily	ANN/GARCH hybrid	GARCH	MSE, RMSE, MAE, MAPE	None
Ribeiro et al. (2021)	RV	CAT, EBAY, MSFT	None	549 weeks	Daily	HAR-PSO-ESN hybrid	ARIMA, HAR, MLP, PSO-ESN	Friedman-test, Post-hoc Nemenyi test, R2, MSE	None
Vidal and Kristjanpoller (2020)	RV	LBMA gold price	None	01.04.1968 – 01.10.2017	14-day	CNN-LSTM	GARCH, SVR, LSTM, CNN	MSE, MCS	None
Kim and Won (2018)	RV	KOSPI 200	None	01.01.2001 – 02.01.2017	Daily	LSTM, GARCH, EGARCH, EWMA hybrid	EGARCH, EWMA, DFN, LSTM	MSE, MAE, HMAE, HMSE, DM-test, WS-test	None
Zolfaghari and Gholami (2021)	RV	DJIA, IXIC	Dollar index, WTI, Trading indicators	04.01.2020 – 31.12.2020	Daily	AWT-LSTM, HAR-X-RV hybrid	ANN-HAR-X-RV	R2, RMSE, MAE, SMAPE, DM-test, ADM-test	None

recommend that future work should strive harder to employ it in their models.

Volatility prediction and forecasting using AI and machine learning are not new concepts. However, as documented in this review, the scientific literature in this combined field is currently not very mature. As is the case in many topics in finance, the academic literature centers around U.S. markets, predominantly equities and oil. Hence, many of the papers discussed in this review can be used as a starting point for further research in order geographical areas and other asset classes. This will shed light on the generality of the findings in the current literature. Various forms of artificial neural networks (ANNs) have proven to be effective in prediction and forecasting of volatility, as well as being able to compute precise VaR measures — the latter of significant importance for financial risk management. As discussed, it is very hard to achieve extremely precise predictions of the volatility estimators, and even more so the true volatility. Therefore, it might be beneficial to quantify the uncertainty related to the volatility forecasts provided by the models. We find very limited application of probabilistic AI in the current literature. Consequently, an interesting field for further research is to implement the most promising models within a Bayesian framework, for instance as Bayesian Neural Networks. This will enable quantification of the uncertainty introduced by the models in terms of outputs and weights to explain trustworthiness of the predictions.

CRediT authorship contribution statement

Elias Søvik Gunnarsson: Data curation, Methodology, Software, Formal analysis, Investigation, Writing – original draft. **Håkon Ramon Isern:** Data curation, Methodology, Software, Formal analysis, Investigation, Writing – original draft. **Aristidis Kaloudis:** Conceptualization, Validation. **Morten Risstad:** Writing – original draft, Writing – review & editing. **Benjamin Vigdel:** Conceptualization, Data curation, Methodology, Software, Formal analysis, Investigation. **Sjur Westgaard:** Conceptualization, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This research was partially funded by The Research Council of Norway throughout the project COMPAMA (https://www.ntnu.edu/compama/), with grant number 314609.

Appendix A. Realized volatility

Andersen and Bollerslev (1998) state that there is no contradiction between good volatility forecasts and poor predictive power for daily squared returns, which has widely been used as a measurement and proxy for volatility when evaluating traditional GARCH-based econometric models. This is demonstrated by how high-frequency intraday data may be used to construct a more accurate and meaningful interdaily volatility measurement using cumulative squared intraday returns. Corsi (2009) defines the standard definition of the realized volatility over a time interval of one day as

$$RV_t^{(d)} = \sqrt{\sum_{j=0}^{M-1} r_{t-j\cdot\Delta}^2}, \tag{A.1}$$

where $\Delta = 1d/M$, and the continuously compounded Δ-frequency returns, i.e. the intraday returns sampled at time interval Δ, is $r_{t-1\cdot\Delta} =$

Table 4
A brief summary of the main results of all featured studies in this paper.

Reference	Main result
Vrontos et al. (2021)	Compare several machine learning models against various logistic regression models in predicting the direction of the VIX index. Based on numerous statistical evaluation measures they find that the machine learning approaches produced better results for out-of-sample forecasts.
Petrozziello et al. (2022)	Compare R-GARCH and GJR-MEM against a univariate and multivariate LSTM for forecasting the realized volatility of stocks from NASDAQ and DIJA. They report that LSTM has better forecasting performance during periods of high volatility, and comparable during more tranquil periods. Conclude that a multivariate LSTM better captures spillover effects, hence the improved performance when forecasting volatility for multiple assets.
Luong and Dokuchaev (2018)	Improves forecasts of realized volatility on S&P500, including directional changes, by including implied volatility in a random forest framework.
Tissaoui et al. (2022)	Compared model performance and variable importance before and during Covid-19 to assess different drivers of volatility. Find that different XGB performs better turbulent periods while SVM fared better for the more tranquil decade leading up til 2020.
Song et al. (2022)	Predicting the SSE Composite Index realized volatility, they find that LSTM, GRU and XGB produce the lowest forecasting errors. All machine learning models benefit from including short-term GARCH-MIDAS volatility forecasts. GRU is the overall best performer.
Christensen et al. (2022)	Study a broad range of machine learning and HAR-RV models to forecast the realized volatility of DIJA constituents. They find machine learning models capturing non-linearities to be superior for short-term forecasts. When increasing the forecasting horizon, random forest and ANN prove to be best.
Liu et al. (2018)	Use RNNs in combination with HAR-RV-Jump models to forecast S&P500 volatility. RNNs is preferred when data is sparse.
Çepni et al. (2022)	Find that uncertainty-related measures in RF and LASSO models improve long-term WTI realized volatility forecasts, compared to an AR benchmark.
Gkillas et al. (2021)	Explore the relative importance of volume-related predictors when forecasting the realized volatility of Bitcoin. Report consistent results for RF and HAR-RV-Jump models.
Yao et al. (2017)	Develop a hybrid ANN-AR model which is superior various GARCH specifications for forecasting foreign exchange rate volatility on hourly and daily basis.
Bouri et al. (2020)	Study the impact of US–China trade war uncertainty in relation to Bitcoin volatility using RF. Report beneficial effects of including Google-search trends, also when RF is compared to various HAR-RV specifications.
Gupta and Pierdzioch (2022)	Building on Çepni et al. (2022) they report further improvements in long-term oil price volatility forecasts when including geopolitical risk indices.
Plakandaras et al. (2017)	Use SVR variants to determine that EPU related to Brexit had an effect on USD/GBP exchange rate realized volatility forecasts.
Higashide et al. (2021)	Forecasting the direction of Nikkei realized volatility they report that including certain exogenous variables is more beneficial in for random forest models compared to a hybrid HAR model.
Oliveira et al. (2017)	Use Twitter data with sentiment values to deduce whether they have an effect on forecasting VIX and the annualized RV on the S&P 500, RSL, DJIA and NDQ. Compare several ML models against each other and an AR(5) benchmark model, report that inclusion of sentiment variables does generally not lead to forecast improvements.
Gupta et al. (2021)	Find that combining investor sentiment related data in addition to macroeconomical and financial data improved the forecasting performance of their suggested random forest and hybrid HAR models for shorter and intermediate DIJA forecasts.
Grigoryeva et al. (2014)	Examine reservoir computing ANNs compared to a VEC(1,1)-model when predicting the realized volatility of several NYSE assets and report systematically better results.
Prasad et al. (2022)	Examined the impact of including macroeconomical variables when predicting the direction of the VIX. Using logistic regression, LGBM and XGB, they find that the two latter models with the exogenous data give better daily and weekly forecasts.
Persio et al. (2021)	Employ a hybrid LSTM-GARCH(1,1) model in a volatility target investment strategy, reporting improved risk-adjusted returns.
Lu et al. (2022)	Compare machine learning models to predict the monthly realized volatility of WTI futures. Report that particularly random forest, boosting, ANN-related and ensemble models perform better relative to several AR(3) models with different extensions.
Qiu (2021)	Proposed a method for combining forecasts based on complete subset least squares vector regressions as an estimator for the HAR-RV model. Report improved out-of-sample forecasts on horizons from one day to one month.
Bucci (2020)	Use a subsample of the Tissaoui et al. (2022) data to assess the robustness of the overall results, finding generally equivalent results.
Chen and Hu (2022)	Reports superior CSI 300 and S&P 500 futures volatility forecasting results for ANN and LSTM compared to AR and EGARCH across loss functions.

(*continued on next page*)

Table 4 (*continued*).

Zhang, He et al. (2022)	Suggest a modified Partial Least Squares approach to construct an aligned GEPU index targeted on the residuals of a simple AR model. Report substantial benefits for forecasting realized variance of oil prices.
Zhang, Wahab et al. (2022)	Forecast monthly realized volatility of WTI futures by using supervised PCA with variable selection. Report increased statistical and economic performance compared to AR, LASSO, ENet and kitchen sink models both in-sample and out-of-sample.
Ghosh and Sanyal (2021)	Use XGB, ERT, DNN and LSTM and obtain precise estimates of future India VIX. By means of the MCS test they conclude that XGB is superior.
Osterrieder et al. (2020)	Utilize NNs to replicate the CBOE VIX. By training an LSTM model on S&P 500 options they devise a profitable trading strategy driven by intradaily deviations between VIX and its futures.
Kristjanpoller and Minutolo (2015)	Use a combined ANN-GARCH model to forecast gold spot and future price realized volatility. They report that including a GARCH forecast as input to an ANN, in addition to other macroeconomic variables, improve performance compared to GARCH.
Ribeiro et al. (2021)	Use a combined HAR-PSO-ESN hybrid to forecast daily realized volatility of three Nasdaq stocks. Compared to multiple machine learning and econometric models, the proposed HAR-PSO-EN leads to statistically significant better results across different horizons.
Vidal and Kristjanpoller (2020)	Utilize image data of the time series of LBMA gold price evolution in a combined CNN-LSTM approach and find superior realized volatility forecasting performance.
Kim and Won (2018)	Combining various GARCH specifications and LSTM in a deep feed forward network they report that two or GARCH-models in combination with LSTM provides the best forecasting results on the KOSPI 200 index.
Zolfaghari and Gholami (2021)	Evaluating AWT-LSTM and HAR-RV hybrid against an ANN and HAR-RV hybrid, they find support for a AWT-LSTM-HAR hybrid model.

$p(t-j)\cdot\Delta - p(t-(j+1))\cdot\Delta$, with t indexing the day and j the time of day; also see Andersen et al. (2001a).

Appendix B. The HAR-RV model

Corsi (2009) suggests a three-factor stochastic volatility model, where the factors are past realized volatilities viewed at different frequencies, namely daily, weekly and monthly. The model's equation for the one-day-ahead realized volatility is

$$RV^{(d)}_{t+1d} = c + \beta^{(d)} RV^{(d)}_t + \beta^{(w)} RV^{(w)}_t + \beta^{(m)} RV^{(m)}_t + \omega_{t+1d}, \tag{B.1}$$

with $\omega_{t+1d} = \tilde{\omega}^{(d)}_{t+1d} - \omega^{(d)}_{t+1d}$, which are the volatility innovations, i.e. the difference between our best future realized volatility prediction and what is observed. Despite the model being seen as a simple AR-type model where a true long-term memory property is absent, it successfully reproduces the main empirical features of financial returns and forecasting performance (Corsi, 2009). Consequently, the model is often used as a benchmark for volatility prediction and forecasting using high-frequency data.

Appendix C. Volatility indices

Model-based

CBOE first introduced the VIX in 1993. Developed by Robert Whiley, the model derived implied volatility using a linear combination of eight near-the-money options on the S&P 100 (Whaley, 1993). The formulas used in the construction of the index are

$$\sigma_1 = \left(\frac{\sigma^{X_L}_{C,1} + \sigma^{X_L}_{P,1}}{2}\right)\left(\frac{X_U - S}{X_U - X_L}\right) + \left(\frac{\sigma^{X_U}_{C,1} + \sigma^{X_U}_{P,1}}{2}\right)\left(\frac{S - X_L}{X_U - X_L}\right), \tag{C.1}$$

$$\sigma_2 = \left(\frac{\sigma^{X_L}_{C,2} + \sigma^{X_L}_{P,2}}{2}\right)\left(\frac{X_U - S}{X_U - X_L}\right) + \left(\frac{\sigma^{X_U}_{C,2} + \sigma^{X_U}_{P,2}}{2}\right)\left(\frac{S - X_L}{X_U - X_L}\right), \tag{C.2}$$

$$VIX = \sigma_1\left(\frac{N_2 - 22}{N_2 - N_1}\right) + \sigma_2\left(\frac{22 - N_1}{N_2 - N_1}\right). \tag{C.3}$$

Here, σ_1 and σ_2 refers to the estimated implied volatility of the S&P 100 options with maturity just below and above 22 trading days. $\sigma^{X_j}_{C,i}$ and $\sigma^{X_j}_{P,i}$ is the implied volatility for puts and calls just above and below the current index S and time to maturity of 22 trading days. X_U and X_L are the strike prices of options with a strike price just above or below the current index price S, respectively. The final variables, N_1 and N_2, are the number of trading days to each nearby contracts maturity. As this model relies on the Black–Scholes option pricing model and its underlying assumptions, it is known as a model-based method.

Model-free

In 2003 the CBOE worked with Goldman Sachs to develop a new model-free implied volatility index, based on the work of Demeterfi et al. (1999). This methodology is centered around fair value pricing variance swaps, and is described as a model-free methodology. By assuming the underlying followed a non-jump diffusive process, and using the fair value of variance swaps as the basis for their calculations, Demeterfi et al. (1999) showed that the expected value of the risk-neutral variance Vt can be expressed as

$$V_t = \frac{2}{T}\left(rT - \left(\frac{S_0}{S_*}e^{rT} - 1\right) - \log\frac{S_*}{S_0} + e^{rT}\int_0^{S_*}\frac{1}{K^2}P(K)dK + e^{rT}\int_{S_*}^{\infty}\frac{1}{K^2}C(K)\,dK\right), \tag{C.4}$$

where r is the risk-free interest rate. T is the time of expiration of the listed contracts, S_0 is the current index price of the underlying and S_* defines the boundary between puts and calls. Finally, K is the strike price of the option, while $P(K)$ and $C(K)$ is the current fair value of a put and call. Based on these results, the CBOE updated the VIX in 2003 by performing a discretization of (C.4) to compute the market's expectation of future volatility. Additionally, the underlying asset was changed to the S&P 500. The procedure is described by the Chicago Board Options Exchange (2022) with the following formula for the VIX

$$VIX = \left(\frac{2}{T}\sum_i \frac{\Delta K_i}{K_i^2}Q(K_i) - \frac{1}{T}\left[\frac{F}{K_0} - 1\right]^2\right)\cdot 100. \tag{C.5}$$

T is the time to expiration in years and F is the forward price implied by the option. K_0 is the strike price closest to the forward index F. By sequentially moving away from the K_0, K_i is the strike price i steps away, and ΔK_i is the interval between strike prices. R is the risk-free interest rate, and $Q(K_i)$ denotes the midpoint between the bid–ask spread for every option with given strike K_i.

The general numerical procedure of fair pricing of variance swaps remains consistent among different indices (Fassas & Siriopoulos, 2021). Hence, the above method serves as a general foundation for how the model-free approach is undertaken in practice.

Table D.5
A summary of abbreviations used in this study.

Abbreviations	Explanation
RF	Random forest
XGB	eXtreme Gradient Boost
GBM	Gradient Boosting Machines
AdaBoost	Adaptive Boosting
ERT	Extremely Random Trees
ET	Extended trees
LGBM	Light Gradient Boosting Machine
ANN	Artificial Neural Network
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
ARNN	Autoregressive Neural Networks
GRU	Gated Recurrent Unit
NAR	Nonlinear Autoregressive Network
EN	Elastic Net
SVM	Support Vector Machine
LDA	Linear Discriminant Analysis
KNN	K Nearest Neighbors
rpart	Recursive Partitioning
XAI	Explainable Artificial Intelligence
PLS	Partial Least Squares
OLS	Ordinary Least Squares
MLE	Maximum Likelihood Estimation
AWT	Adaptive Wavelet Transform
PSO	Particle Swarm Optimization
ESN	Echo State Network
VaR	Value at Risk
CVaR	Conditional Value at Risk
EPU	Economic Policy Uncertainty
CBOE	Chicago Board Options Exchange
MSE	Mean Squared Error
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
R2 OOS	R2 Out-Of-Sample
MAPE	Mean Absolute Percentage Error
MSPE	Mean Squared Percentage Error
SMAPE	Symmetric Mean Absolute Percentage Error
RMSFE	Root Means Squared Forecasting Error
MSFE	Means Squared Forecasting Error
NMSE	Normalized Mean Squared Error
NSE	Normalized Squared Error
IA	Index of Agreement
MSD	Mean Squared Deviation
TI	Theil Inequality Index
DA	Directional Accuracy
HMAE	Heteroskedasticity adjusted Mean Absolute Error
MAD	Mean Absolute Deviation
MCC	Matthews Correlation Coefficient
SMSFE	Standardized Means Square Forecasting Error
NMAE	Normalized Mean Absolute Error
MAFE	Mean Absolute Forecasting Error
RMSPE	Root Mean Squared Percentage Error
PCC	Pearson Correlation Coefficient
DM	Diebold–Mariano test
CW	Clark–West test
MCC	Model Confidence Set

Appendix D. Abbreviations

See Table D.5.

References

Andersen, T. G., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, *39*(4), 885–905, URL: http://www.jstor.org/stable/2527343.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2001a). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, *96*(453), 42–55. http://dx.doi.org/10.1198/016214501750332965, arXiv:https://doi.org/10.1198/016214501750332965.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2001b). Modeling and forecasting realized volatility. *SSRN Journal Electronic Journal*, http://dx.doi.org/10.2139/ssrn.267792.

Becker, R., Clements, A. E., & White, S. I. (2007). Does implied volatility provide any information beyond that captured in model-based volatility forecasts? *Journal of Banking & Finance*, *31*(8), 2535–2549. http://dx.doi.org/10.1016/j.jbankfin.2006.11.013, URL: https://www.sciencedirect.com/science/article/pii/S0378426607000428.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327, URL: https://ideas.repec.org/a/eee/econom/v31y1986i3p307-327.html.

Bollerslev, T., Hood, B., Huss, J., & Pedersen, L. H. (2018). Risk everywhere: Modeling and managing volatility. *The Review of Financial Studies*, *31*(7), 2729–2773. http://dx.doi.org/10.1093/rfs/hhy041.

Bouri, E., Gkillas, K., Gupta, R., & Pierdzioch, C. (2020). Forecasting realized volatility of Bitcoin: The role of the trade war. *Computer and Economics*, *57*(1), 29–53. http://dx.doi.org/10.1007/s10614-020-10022-4.

Bucci, A. (2020). Realized volatility forecasting with neural networks. *Journal of Financial Econometrics*, *18*(3), 502–531. http://dx.doi.org/10.1093/jjfinec/nbaa008.

Bustos, O., & Pomares-Quimbaya, A. (2020). Stock market movement forecast: A systematic review. *Expert Systems with Applications*, *156*, Article 113464. http://dx.doi.org/10.1016/j.eswa.2020.113464.

Çepni, O., Gupta, R., Pienaar, D., & Pierdzioch, C. (2022). Forecasting the realized variance of oil-price returns using machine learning: Is there a role for U.S. state-level uncertainty? *Energy Economics*, *114*, Article 106229. http://dx.doi.org/10.1016/j.eneco.2022.106229.

Chen, X., & Hu, Y. (2022). Volatility forecasts of stock index futures in China and the US–A hybrid LSTM approach. In D. Kugiumtzis (Ed.), *PLoS One*, *17*(7), Article e0271595. http://dx.doi.org/10.1371/journal.pone.0271595.

Chicago Board Options Exchange (2022). Volatility index methodology: Cboe volatility index. *Chicago Board Options Exchange*, URL: https://cdn.cboe.com/api/global/us_indices/governance/Volatility_Index_Methodology_Cboe_Volatility_Index.pdf.

Christensen, K., Siggaard, M., & Veliyev, B. (2022). A machine learning approach to volatility forecasting. *Journal of Financial Econometrics*, http://dx.doi.org/10.1093/jjfinec/nbac020.

Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, *138*(1), 291–311.

Corsi, F. (2009). A simple long memory model of realized volatility. *Journal of Financial Econometrics*, *7*, 174–196. http://dx.doi.org/10.1093/jjfinec/nbp001.

Dai, Z., Zhou, H., Wen, F., & He, S. (2020). Efficient predictability of stock return volatility: The role of stock market implied volatility. *The North American Journal of Economics and Finance*, *52*, Article 101174. http://dx.doi.org/10.1016/j.najef.2020.101174, URL: https://www.sciencedirect.com/science/article/pii/S1062940820300711.

Demeterfi, K., Derman, E., Kamal, M., & Zou, J. (1999). A guide to volatility and variance swaps. *The Journal of Derivatives*, *6*(4), 6. http://dx.doi.org/10.3905/jod.1999.319129, 1999-may 31.

Diebold, F. X., & Mariano, R. S. (1995). Com paring predictive accu racy. *Journal of Business & Economic Statistics*, *13*(3), 253–263.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*(4), 987. http://dx.doi.org/10.2307/1912773.

Engle, R. F., Ghysels, E., & Sohn, B. (2013). Stock market volatility and macroeconomic fundamentals. *The Review of Economics and Statistics*, *95*(3), 776–797. http://dx.doi.org/10.1162/rest_a_00300.

European Commission and Directorate-General for Communications Networks, Content and Technology (2019). *Ethics guidelines for trustworthy AI*. Publications Office, http://dx.doi.org/10.2759/346720.

Fassas, A. P., & Siriopoulos, C. (2021). Implied volatility indices – A review. *The Quarterly Review of Economics and Finance*, *79*, 303–329. http://dx.doi.org/10.1016/j.qref.2020.07.004, URL: https://www.sciencedirect.com/science/article/pii/S1062976920300855.

Ghosh, I., & Sanyal, M. K. (2021). Introspecting predictability of market fear in Indian context during COVID-19 pandemic: An integrated approach of applied predictive modelling and explainable AI. *International Journal of Information Management Data Insights*, *1*(2), Article 100039. http://dx.doi.org/10.1016/j.jjimei.2021.100039.

Gkillas, K., Tantoula, M., & Tzagarakis, M. (2021). Transaction activity and bitcoin realized volatility. *Operations Research Letters*, *49*(5), 715–719. http://dx.doi.org/10.1016/j.orl.2021.06.016.

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, *48*(5), 1779–1801. http://dx.doi.org/10.1111/j.1540-6261.1993.tb05128.x.

Grigoryeva, L., Henriques, J., Larger, L., & Ortega, J.-P. (2014). Stochastic nonlinear time series forecasting using time-delay reservoir computers: Performance and universality. *Neural Networks*, *55*, 59–71. http://dx.doi.org/10.1016/j.neunet.2014.03.004.

Gupta, R., Nel, J., & Pierdzioch, C. (2021). Investor confidence and forecastability of US stock market realized volatility: Evidence from machine learning. *Journal of Behavioral Finance*, 1–12. http://dx.doi.org/10.1080/15427560.2021.1949719.

Gupta, R., & Pierdzioch, C. (2022). Forecasting the realized variance of oil-price returns: a disaggregated analysis of the role of uncertainty and geopolitical risk. *Environmental Science and Pollution Research*, *29*(34), 52070–52082. http://dx.doi.org/10.1007/s11356-022-19152-8.

Hansen, P. R., Huang, Z., & Shek, H. H. (2011). Realized GARCH: a joint model for returns and realized measures of volatility. *Journal of the Applications and Economics*, *27*(6), 877–906. http://dx.doi.org/10.1002/jae.1234.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, *124*, 226–251. http://dx.doi.org/10.1016/j.eswa.2019.01.012.

Hewamalage, H., Ackermann, K., & Bergmeir, C. (2023). Forecast evaluation for data scientists: common pitfalls and best practices. *Data Mining and Knowledge Discovery*, *37*(2), 788–832.

Hiebl, M. (2021). Sample selection in systematic literature reviews of management research. *Organizational Research Methods*, http://dx.doi.org/10.1177/1094428120986851.

Higashide, T., Tanaka, K., Kinkyo, T., & Hamori, S. (2021). New dataset for forecasting realized volatility: Is the Tokyo stock exchange co-location dataset helpful for expansion of the heterogeneous autoregressive model in the Japanese stock market? *JRFM*, *14*(5), 215. http://dx.doi.org/10.3390/jrfm14050215.

Hoepner, A. G. F., & Unerman, J. (2012). Explicit and implicit subject bias in the ABS journal quality guide. *Accounting Education*, *21*(1), 3–15. http://dx.doi.org/10.1080/09639284.2011.651291, arXiv:https://doi.org/10.1080/09639284.2011.651291.

Jia, F., & Yang, B. (2021). Forecasting volatility of stock index: Deep learning model with likelihood-based loss function. In B. M. Tabak (Ed.), *Complexity*, *2021*, 1–13. http://dx.doi.org/10.1155/2021/5511802.

Jiang, W. (2021). Applications of deep learning in stock market prediction: Recent progress. *Expert Systems with Applications*, *184*, Article 115537. http://dx.doi.org/10.1016/j.eswa.2021.115537.

Kim, H. Y., & Won, C. H. (2018). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. *Expert Systems with Applications*, *103*, 25–37. http://dx.doi.org/10.1016/j.eswa.2018.03.002.

Kitchenham, B. (2004). *Procedures for performing systematic reviews*: *Vol. 33*, Keele, UK: Keele Univ..

Kristjanpoller, W., & Minutolo, M. C. (2015). Gold price volatility: A forecasting approach using the Artificial Neural Network–GARCH model. *Expert Systems with Applications*, *42*(20), 7245–7251. http://dx.doi.org/10.1016/j.eswa.2015.04.058.

Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems with Applications*, *197*, Article 116659. http://dx.doi.org/10.1016/j.eswa.2022.116659.

Liu, F., Pantelous, A. A., & von Mettenheim, H.-J. (2018). Forecasting and trading high frequency volatility on large indices. *Quantitative Finance*, *18*(5), 737–748. http://dx.doi.org/10.1080/14697688.2017.1414489.

Liu, L. Y., Patton, A. J., & Sheppard, K. (2015). Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes. *Journal of Econometrics*, *187*(1), 293–311. http://dx.doi.org/10.1016/j.jeconom.2015.02.008.

Lu, X., Ma, F., Xu, J., & Zhang, Z. (2022). Oil futures volatility predictability: New evidence based on machine learning models. *International Review of Financial Analysis*, *83*, Article 102299. http://dx.doi.org/10.1016/j.irfa.2022.102299.

Luong, C., & Dokuchaev, N. (2018). Forecasting of realised volatility with the random forests algorithm. *JRFM*, *11*(4), 61. http://dx.doi.org/10.3390/jrfm11040061.

Malliaris, M., & Salchenberger, L. (1996). Using neural networks to forecast the S&P 500 implied volatility. *Neurocomputing*, *10*(2), 183–195. http://dx.doi.org/10.1016/0925-2312(95)00019-4, URL: https://www.sciencedirect.com/science/article/pii/0925231295000194, Financial Applications, Part I.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, *59*(2), 347. http://dx.doi.org/10.2307/2938260.

Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, *73*, 125–144. http://dx.doi.org/10.1016/j.eswa.2016.12.036.

Osterrieder, J., Kucharczyk, D., Rudolf, S., & Wittwer, D. (2020). Neural networks and arbitrage in the VIX. *Digital Finance*, *2*(1–2), 97–115. http://dx.doi.org/10.1007/s42521-020-00026-y.

Persio, L. D., Garbelli, M., & Wallbaum, K. (2021). Forward-looking volatility estimation for risk-managed investment strategies during the COVID-19 crisis. *Risks*, *9*(2), 33. http://dx.doi.org/10.3390/risks9020033.

Petrozziello, A., Troiano, L., Serra, A., Jordanov, I., Storti, G., Tagliaferri, R., & Rocca, M. L. (2022). Deep learning for volatility forecasting in asset management. *Software Computers*, *26*(17), 8553–8574. http://dx.doi.org/10.1007/s00500-022-07161-1.

Plakandaras, V., Gupta, R., & Wohar, M. E. (2017). The depreciation of the pound post-brexit: Could it have been predicted? *Finance Research Letters*, *21*, 206–213. http://dx.doi.org/10.1016/j.frl.2016.12.003.

Poon, S.-H., & Granger, C. W. (2003). Forecasting volatility in financial markets: A review. *Journal of Economic Literature*, *41*(2), 478–539. http://dx.doi.org/10.1257/002205103765762743, URL: https://www.aeaweb.org/articles?id=10.1257/002205103765762743.

Prasad, A., Bakhshi, P., & Seetharaman, A. (2022). The impact of the U.S. macroeconomic variables on the CBOE VIX index. *JRFM*, *15*(3), 126. http://dx.doi.org/10.3390/jrfm15030126.

Qiu, Y. (2021). Complete subset least squares support vector regression. *Economics Letters*, *200*, Article 109737. http://dx.doi.org/10.1016/j.econlet.2021.109737.

Ribeiro, G. T., Santos, A. A. P., Mariani, V. C., & dos Santos Coelho, L. (2021). Novel hybrid model based on echo state neural network applied to the prediction of stock price return volatility. *Expert Systems with Applications*, *184*, Article 115490. http://dx.doi.org/10.1016/j.eswa.2021.115490.

Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning : A systematic literature review: 2005–2019. *Applied Soft Computing*, *90*, Article 106181. http://dx.doi.org/10.1016/j.asoc.2020.106181.

Song, Y., Tang, X., Wang, H., & Ma, Z. (2022). Volatility forecasting for stock market incorporating macroeconomic variables based on GARCH-MIDAS and deep learning models. *Journal of Forecasting*, http://dx.doi.org/10.1002/for.2899.

Thakkar, A., & Chaudhari, K. (2021). A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions. *Expert Systems with Applications*, *177*, Article 114800. http://dx.doi.org/10.1016/j.eswa.2021.114800.

Tissaoui, K., Zaghdoudi, T., Hakimi, A., Ben-Salha, O., & Amor, L. B. (2022). Does uncertainty forecast crude oil volatility before and during the COVID-19 outbreak? Fresh evidence using machine learning models. *Energies*, *15*(15), 5744. http://dx.doi.org/10.3390/en15155744.

Tsay, R. S. (2010). *Analysis of financial time series*. John Wiley & Sons, Inc., http://dx.doi.org/10.1002/9780470644560.

Tsekova, D., & Popina, E. (2022). Amped-up bet on volatility goes awry with loss of $400 million. https://www.bloomberg.com/news/articles/2022-08-02/amped-up-bet-on-volatility-goes-awry-with-loss-of-400-million.

Vidal, A., & Kristjanpoller, W. (2020). Gold volatility prediction using a CNN-LSTM approach. *Expert Systems with Applications*, *157*, Article 113481. http://dx.doi.org/10.1016/j.eswa.2020.113481.

Vrontos, S. D., Galakis, J., & Vrontos, I. D. (2021). Implied volatility directional forecasting: a machine learning approach. *Quantitative Finance*, *21*(10), 1687–1706. http://dx.doi.org/10.1080/14697688.2021.1905869.

Whaley, R. E. (1993). Derivatives on market volatility: Hedging tools long overdue. *The Journal of Derivatives*, *1*(1), 71–84, URL: https://jod.pm-research.com/content/1/1/71.short.

Yao, Y., Zhai, J., Cao, Y., Ding, X., Liu, J., & Luo, Y. (2017). Data analytics enhanced component volatility model. *Expert Systems with Applications*, *84*, 232–241. http://dx.doi.org/10.1016/j.eswa.2017.05.025.

Zhang, Y., He, M., Wang, Y., & Liang, C. (2022). Global economic policy uncertainty aligned: An informative predictor for crude oil market volatility. *International Journal of Forecasting*, http://dx.doi.org/10.1016/j.ijforecast.2022.07.002.

Zhang, Y., Wahab, M., & Wang, Y. (2022). Forecasting crude oil market volatility using variable selection and common factor. *International Journal of Forecasting*, http://dx.doi.org/10.1016/j.ijforecast.2021.12.013.

Zolfaghari, M., & Gholami, S. (2021). A hybrid approach of adaptive wavelet transform, long short-term memory and ARIMA-GARCH family models for the stock index prediction. *Expert Systems with Applications*, *182*, Article 115149. http://dx.doi.org/10.1016/j.eswa.2021.115149.