*data*

MDPI

*Article*

# Leveraging Return Prediction Approaches for Improved Value-at-Risk Estimation

**Farid Bagheri [1], Diego Reforgiato Recupero [1,*] and Espen Sirnes [2]**

[1] Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, 09124 Cagliari, Italy; f.bagheri@studenti.unica.it
[2] School of Business and Economics, UiT The Arctic University of Norway, Breivangvegen 23, 9010 Tromsø, Norway; espen.sirnes@uit.no
* Correspondence: diego.reforgiato@unica.it

**Abstract:** Value at risk is a statistic used to anticipate the largest possible losses over a specific time frame and within some level of confidence, usually 95% or 99%. For risk management and regulators, it offers a solution for trustworthy quantitative risk management tools. VaR has become the most widely used and accepted indicator of downside risk. Today, commercial banks and financial institutions utilize it as a tool to estimate the size and probability of upcoming losses in portfolios and, as a result, to estimate and manage the degree of risk exposure. The goal is to obtain the average number of VaR "failures" or "breaches" (losses that are more than the VaR) as near to the target rate as possible. It is also desired that the losses be evenly distributed as possible. VaR can be modeled in a variety of ways. The simplest method is to estimate volatility based on prior returns according to the assumption that volatility is constant. Otherwise, the volatility process can be modeled using the GARCH model. Machine learning techniques have been used in recent years to carry out stock market forecasts based on historical time series. A machine learning system is often trained on an in-sample dataset, where it can adjust and improve specific hyperparameters in accordance with the underlying metric. The trained model is tested on an out-of-sample dataset. We compared the baselines for the VaR estimation of a day ($d$) according to different metrics (i) to their respective variants that included stock return forecast information of $d$ and stock return data of the days before $d$ and (ii) to a GARCH model that included return prediction information of $d$ and stock return data of the days before $d$. Various strategies such as ARIMA and a proposed ensemble of regressors have been employed to predict stock returns. We observed that the versions of the univariate techniques and GARCH integrated with return predictions outperformed the baselines in four different marketplaces.

**Keywords:** VaR estimation; machine learning; return prediction; walking forward optimization



**Citation:** Bagheri, F.; Reforgiato Recupero, D.; Sirnes, S. Leveraging Return Prediction Approaches for Improved Value-at-Risk Estimation. *Data* **2023**, *8*, 133. https://doi.org/10.3390/data8080133

Academic Editors: Edson Talamini, Letícia De Oliveira and Filipe Portela

Received: 3 July 2023
Revised: 7 August 2023
Accepted: 15 August 2023
Published: 17 August 2023



## 1. Introduction

Value at risk (VaR), dubbed the "new science of risk management", is a metric used to forecast the most significant potential losses over a given period. It represents a solution for reliable quantitative risk management used by risk managers and regulators. The need for risk metrics became apparent after events such as the market crash in October 1987, the following crises in emerging markets, and catastrophic losses brought on by the trading activities of institutions like Orange County, Long-Term Capital Management (LTCM), and Metallgesellschaft [1].

VaR is defined as the maximum loss within some level of confidence, usually 95 or 99%. Hence, if for example, VaR is −20 at 95% confidence, it means that in 95% of the days, if there is a loss, it will be less than 20. VaR can be calculated both in terms of amounts and relative returns.

VaR has become the industry standard and the most well-known measure of downside risk. It is currently a tool used by commercial banks and financial institutions to gauge the

size and likelihood of future losses in portfolios and, therefore, to gauge and manage the degree of risk exposure.

The optimal scenario for risk managers or regulators is a VaR that also works perfectly in retrospect; that is, if the desired confidence interval is 95%, then in retrospect, the actual loss should exceed VaR exactly 5% of the time. However, that is difficult to accomplish. Managers typically monitor the VaR, and if the agent takes more risk than agree or the market suddenly becomes riskier, the manager orders the agent to wind down his/her exposure.

If a normal distribution is assumed, the VaR is simply the volatility times the inverse of the standard normal distribution at the desired level of confidence. Given volatility, this is straightforward to calculate. However, since volatility is not constant, estimating it is far from trivial. Historical volatility is often chosen, but since volatility in financial markets greatly varies, historical volatility is usually very inaccurate.

It is well known that returns in financial markets are far from normally distributed because volatility varies. Most financial markets show "fat tails" of the return distribution, indicating that extreme results are far more likely than what the normal distribution implies. These extreme movements also tend to be clustered over time. The most likely and widely accepted explanation for this is that volatility is not constant but a process.

Hence, in retrospect, a regulator or risk manager rarely gets VaR exactly right. Therefore, the desired outcome is that the number of VaR "failures" or "breaches" (losses exceeding the VaR) is as close as possible to the desired rate, on average. In addition, it is preferred that the losses be as unclustered as possible. Evaluating this is called "back testing".

From the perspective of a regulator or risk manager, a breach event typically triggers some kind of risk management action, whereby risk exposure is reduced. The objective is to minimize the probability of disasters. For example, during the 2007–2008 financial crisis, we saw breaches multiple times, and in many cases, emergency actions were not triggered before it was too late. If the objective is to handle a potential crisis as early as possible, the optimal solution would, in principle, be to set a low absolute VaR so that breach events are triggered often.

However, for a regulated bank or entity, a low VaR has a cost. A higher VaR means that less capital can be allocated to activities with higher risks and returns, so a conservative VaR reduces revenue. This is often referred to as the capital charge of VaR. Both limiting the probability of disaster and, at the same time, inflicting minimum capital charge on the regulated entity is achieved by selecting the most precise failure rate, with as little clustering as possible.

There are several approaches to modeling VaR, the most rudimentary of which is to assume constant volatility and estimate VaR based on past returns. This method is called the normal method, as it assumes constant volatility and a normal distribution.

If we acknowledge that volatility is not constant, we can handle this by either assuming a distribution that is not normal, by modeling the volatility process itself, or both.

If we take the alternative distribution approach, the natural choice is to use the actual historical empirical distribution. This also happens to be the most frequently used approach in the industry [2], probably because it is both simple and intuitive. We call this method the historical simulation method, which is the term commonly used in the literature.

An alternative to empirical distribution is to use a mathematically formulated distribution, such as the extreme value or hyperbolic distribution. However, by definition, the empirical distribution always perfectly fits the data, so there is rarely the need to abstract to a less precise mathematically formulated distribution.

The third option is to model the volatility process itself. The state-of-the-art accepted solution for this is the generalized autoregressive conditional heteroskedasticity (GARCH) model [3]. An application of this model in the regulatory framework is called exponential weighted moving average (EWMA), a GARCH method with fixed parameters usually determined by regulatory authorities.

However, a GARCH model where the parameters are set to maximize the model's fit to the data is even more likely to predict future variance correctly. A more precise estimation of future variance yields a more precise VaR estimate.

In recent years, machine learning systems have been employed to execute stock market forecasts on the basis of historical time series. Usually, a machine learning system is trained on an in-sample dataset, where it might tune and optimize certain hyperparameters according to the underlying metric. Then, the trained model is tested on an out-of-sample dataset. It is important to quantify the influence of predictions on the economic level, in addition to simply evaluating the percentage of accurate predictions (i.e., accuracy). For instance, if we suffered significant financial losses for portions of a five-year period (for instance, two consecutive years), any further investment would have been stopped in a real-world scenario, and the measurement of good accuracy in the predictions for that period is not significant. For this reason, in addition to the accuracy metric, other metrics are taken into account, including maximum drawdown, coverage, and return over maximum drawdown.

The biggest problem with VaR is that it is not possible to know tomorrow's risk with certainty. Hence, there will always be a risk that the unwinding comes too late. The use of predictive methods such as machine learning and GARCH is a way of mitigating this problem. VaR essentially depends on the level of risk, so it is actually just a functional transformation of volatility. VaR is used instead of volatility because many people consider the maximum expected loss to be easier to understand.

Therefore, inspired by the recent success of machine and deep learning methods for return prediction, in this paper, we investigated the employment of predicted returns for a certain day ($d$) (performed by machine learning methods) to compute a VaR estimate of $d$. We used different approaches to stock return prediction, which we integrated within the univariate strategy for VaR estimation. To the best of our knowledge and in contrast to past works in the literature, this is the first attempt to combine market returns and predicted returns for VaR estimation. One more approach we tested is to add stock return predictions of a certain day ($d$) and stock returns of the days before $d$ to the GARCH model to compute a VaR estimate of $d$. We developed the Python 3.9.12 package Paneltime to do this. As far as we know, this is the only Python package that can take additional regressors into consideration in the GARCH model. The Paneltime package analytically calculates the Hessian matrix, in addition to the gradient. This makes it more likely to find parameters close to the real optimal parameters. In addition, Paneltime can be used to analyze the GARCH process in panels, which is novel. For this particular study, this function was not utilized, but it is a possible extension for future papers.

We compared the baselines (univariate strategies) against their respective versions integrated with stock return prediction information for the day for which the VaR was being estimated and the previous market returns and against a GARCH model integrated with the same information according to several metrics for VaR prediction. Different approaches for stock return prediction have been used (ARIMA, an ensemble of regressors we first introduced for statistical arbitrage and that we adapted for stock return prediction in this context [4]). We noticed how the integrated versions of the univariate strategies and GARCH provide benefits within several tested markets over the baselines, proving the validity of the provision of predicted return information for VaR estimation.

The remainder of this paper is organized as follows. Section 2 discusses related work on return prediction and VaR estimation using machine and deep learning strategies. Section 3 formulates the task we want to solve and provides information about the baseline machine learning approaches we used and the walking-forward mechanism. Section 4 describes the datasets we employed. The proposed approach for predicting returns is depicted in Section 5. The performance evaluation we carried out, the baselines for VaR estimations, the relative metrics we used, and the results for the adopted markets are discussed in Section 6. Finally, Section 7 ends the paper with conclusions and suggestions for future directions in research.

## 2. Related Works

In this section, we discuss a list of works involving VAR prediction using machine and deep learning. For this reason, we separate the literature review into two sections: one for machine learning and the other for deep learning. The final section details the differences between our approach and the state-of-the-art methods, highlighting the innovations of our method.

### *2.1. Machine Learning Approaches*

The work performed by the authors of [5] showed that using an exponentially weighted quantile regression via support vector machine (SVM) can forecast the multiperiod VaR with better accuracy than competing methods. In the same direction, employing the Tokyo Stock Exchange (Nikkei 225 index), the authors of [6] struggled to obtain results for different models to estimating VaR using realized volatility, non-linear support vector machines, and ARCH-type models. The goal was to find the best-performing model for computing one-day-ahead VaR. The authors found that the hybrid SVM–HAR–ARCH-type model performed better when 15 min intraday returns were used. In another study, the authors introduced a novel classification approach known as extended robust support vector machine (ER-SVM), which aims to minimize an intermediate risk measure positioned between conditional value at risk and VaR [7]. The objective of this method was to develop a model that exhibits reduced sensitivity to outliers present in the distribution's tail, thereby offering enhanced utility in the field of financial risk management. To evaluate the predictive performance of ER-SVM, they conducted numerical experiments and compared their outcomes with other classification methods, namely robust SVM, $\nu$-SVM, and E$\nu$-SVM. The findings of their analysis indicate that ER-SVM surpasses the performance of the alternative methods in scenarios involving outliers, thereby establishing its superiority in handling such instances. In a study conducted in 2015, the authors utilized SVR to forecast and estimate the volatility and VaR of the Belex 15 index [8]. The output of their SVR-based model consisted of a 5-day observed VaR. They compared their findings against the VaR estimations derived from the Markov regime switching model and a feed-forward neural network VaR. Their analysis demonstrated that the SVR tool provided superior VaR estimations when compared to the alternative methods. Other authors proposed an innovative non-linear and non-parametric framework for forecasting VaR that addresses the limitations of parametric models by adopting a fully data-driven approach [9]. In their approach, they employed SVR to model the mean and volatility, drawing inspiration from the standard GARCH formulation. To derive VaR, they used kernel density estimation (KDE). The effectiveness of the proposed framework was assessed through a comparative analysis with standard GARCH models, encompassing exponential and threshold GARCH models employing diverse error distributions. The results obtained from their study demonstrate that the SVR–GARCH–KDE hybrid model outperformed conventional linear and parametric models in terms of accuracy in forecasting the VaR. Others presented their research on portfolio optimization using a hybrid SVR–GARCH–KDE model [10]. Specifically, they focused on estimating the VaR for the LQ45 portfolio, which is a stock index in Indonesia. They found that their model was able to provide flexible return distribution characteristics, which is important for investors in managing risk. Last but not least, the authors of [2] analyzed revisions under Base1 III for market risks that allow for the conservative combination of short- and long-period VaRs. It was found that the combination of short and long historical observation periods improved the performance in regulatory backtests, resulting in lower penalties.

### *2.2. Deep Learning Approaches*

By using three different daily stock market datasets, like Brent Oil, Gold, and Copper, the authors of [11] attempted to forecast one-day-ahead VaR with three different neural network models: a multilayer perceptron (MLP) model, a recurrent neural network (RNN), and a higher-order neural network (HONN). In their work, RiskMetrics volatility and the

ARMA–GARCH (1,1) model were used as benchmark models to compare the achieved results. The final results demonstrated the fact that neural networks provided the best VaR predictions.

Furthermore, the authors of [12] utilized Psi Sigma neural networks to predict one-day-ahead VaR for Brent oil and gold bullion series. To benchmark their results, they used VaR forecasts from two different neural networks and genetic programming algorithms; some traditional techniques like the ARMA-Glosten, Jagannathan, and Runkle (1, 1) models; and the RiskMetrics volatility model. According to the results, their proposed model outperformed the baselines in estimating VaR with 5% and 1% levels of confidence.

The authors of [13] introduced a new model called a quantile autoregression neural network (QARNN),which is a combination of an artificial neural network (ANN) structure with the quantile autoregression (QAR) method. To evaluate its performance, they used Hong Kong Hang Seng Index (HSI), the US S&P500 Index (S&P500), and the Financial Times Stock Exchange 100 Index (FTSE100) time series. They used Monte Carlo simulation and empirical analyses of different real stock indices. The final results showed that QARNN generally outperformed other classical models in terms of the accuracy of VaR evaluation.

GELM is a non-linear random mapping model proposed by the authors of [14] that is a combination of the GARCH model and the extreme learning machine (ELM) used to compute the VaR. Its performance and precision have proven to be better those of other traditional models, like GARCH, SVM, and ELM.

In 2018, the authors of [15] introduced the EMD-DBN ensemble model for estimating VaR by associating the deep belief network ensemble model with the empirical mode decomposition (EMD) technique. The proposed model could identify more optimal ensemble weights and better integrate the partial information from extracted risk estimates. The authors used a forex market dataset to analyze and test their proposed model. The results illustrated that by employing this model, financial institutions and users of this model could obtain better insights, supporting the estimation of risk with accurate results.

Reinforcement learning has also been employed for VAR prediction. For example, the authors of [16] explored a deep reinforcement learning approach to minimize capital charge and improve risk models. Their work sought to establish a link between dynamic programming and reinforcement learning. The results showed that the deep reinforcement learning approach is capable of solving financial optimization problems characterized by a complex Markov decision process. In [17], a model-based deep reinforcement learning architecture was presented to solve the dynamic portfolio optimization problem. The goal was to develop an automatic trading system that could achieve a profitable and risk-sensitive portfolio using historical real financial market data. The proposed architecture consisted of an infused prediction module (IPM), a generative adversarial data augmentation module, and a behavior-cloning module. The authors observed that the use of IPM drastically improved the Sharpe and Sortino ratios. In [18], researchers introduced a mean-VaR-based deep reinforcement learning framework for practical algorithmic trading that outperformed other benchmark strategies on an ETF portfolio.

A hybrid and semiparametric model based on asymmetric Laplace (AL) quasi-likelihood and employing long short-term memory (LSTM) was proposed by the authors of [19]. Known as LSTM-AL, the proposed model was able to forecast and efficiently capture the underlying dynamics of VaR and expected shortfall (ES) in the financial sector.

In the same year, other researchers presented a novel model for measuring market risk based on variational autoencoders (VAEs) called encoded VaR [20]. The authors utilized VAE to produce a dataset similar to the real-world dataset, which allowed them to obtain a variance of return in a non-parametric way and, thus, approximate the VaR.

The authors of [21] constructed and simulated a market risk warning model based on LSTM-VaR. They used a dataset of stock market data that was preprocessed and standardized before being fed to the LSTM-VaR model. They selected 15 indicators to predict the standard deviation of stock returns. The probability distribution of the return rate under the conditional distribution was obtained according to the predicted results, and the VaR

was then determined. The results indicate a better VaR estimation of their proposed model compared to traditional prediction models.

Recently, the authors of [22] proposed a semiparametric, parsimonious VaR forecasting model based on quantile regression and machine learning methods combined with readily available market prices of option contracts from the over-the-counter foreign exchange rate interbank market. They employed ensemble methods and neural networks.

### *2.3. Differences with Respect to State-of-the-Art Approaches*

In this paper, we target the problem of estimating VaR. In contrast to previously proposed approaches, we provide the contributions and innovations:

- We leverage machine learning methods to predict the return for the day ($d$) for which VaR is being estimated, then integrate the obtained information with past returns to find the VaR estimate for $d$ using univariate strategies and GARCH;
- To predict returns for the day for which the VaR is being estimated, we use two approaches: ARIMA and an ensemble of regressors successfully employed for statistical arbitrage [4];
- We also developed a Python package called PanelTime, which implements a GARCH model that can integrate the predicted return for the underlying day ($d$) with the returns of past days. PanelTime can simultaneously estimate panels with fixed/random effects and time series with GARCH/ARIMA. As far as we know, it is the only package that does this simultaneously. Unlike alternative Python packages, Paneltime also allows for the specification of additional regressors in the GARCH model and calculates the Hessian matrix analytically, which makes it more likely to obtain estimates close to the true parameters.

## 3. Background

In this section, we detail the tools, approaches, and strategies that we employed in this paper.

### *3.1. VAR Prediction*

VaR was employed for the first time by companies in the late 1980s [23], and since then, it has attracted considerable attention among researchers. VaR indicates the maximum amount of investment that can be lost by a financial institution or company over a specified time horizon under normal market conditions within a certain confidence level [24]. It is a widely used measure in risk management to approximate potential losses. In this method, historical data and statistical models are used to predict potential losses in the future [25].

More formally, for a certain time horizon and probability ($p$), the $p$ VaR is defined as the largest possible loss during that time after excluding all worse outcomes whose combined probability is at most $p$. For instance, if a portfolio of equities has a one-day 95% VaR of \$1 million, it has a 0.05 likelihood of losing more than \$1 million in value over the course of a single day if there is no trading. On average, this portfolio is predicted to lose \$1 million or more on 1 out of every 20 days (based on a 5% probability).

Investors and risk managers utilize VaR estimations to make informed decisions, manage their exposure to risk, and allocate resources effectively. Daily returns are computed according to the following formula:

$$x_t = log(\frac{P_t}{P_{t-1}}) \tag{1}$$

where $P_t$ is the value of stock at $t$, $P_{t-1}$ is the value of stock on the previous day, and $X$ represents the distribution of these returns. Then, VaR is calculated as:

$$VaR_\alpha(X) = sup\{x \in R : F_X(x) < \alpha\} = F_X^{-1}(\alpha) \tag{2}$$

where $F_X$ is the cumulative distribution function of $X$, and $VaR_\alpha(X)$ is the VaR at confidence level $\alpha$ for the random variable ($X$).

### 3.2. ARIMA

The autoregressive integrated moving average (ARIMA) predicts future values based on past values. It consists of three distinct elements, namely autoregressive (AR), integrated (I), and moving average (MA) models [26].

- Autoregressive (AR) model: An Autoregressive model [27] with $p$, which represents the number of lagged observations, can be defined as:

$$Y = c + a_1 y_{t-1} + a_2 y_{t-2} + ... + a_p y_{t-p} \tag{3}$$

  where $Y$ represents the current value of the time series that we are attempting to predict, $c$ is the constant term or the intercept, $a_1, a_2, \ldots, a_p$ are the coefficients for the autoregressive terms, and $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$ are the lagged values of the time series. Here, the model relies on past values, and the objective is to derive estimates for the coefficients ($a_i$). In other words, the current observation depends on past observations, as it is assumed that the current value of a variable is related to its previous values.
- Integrated (I) model: To deal with non-stationary time series data, an integrated part of ARIMA called differencing is used to transform the data to remove trends or cycles that change over time, thereby making them stationary. In a stationary time series, the mean and variance are constant over time. It is easier to predict values when the time series is stationary. Differencing is denoted by $d$ in the ARIMA model and illustrates the number of differencing iterations needed to make the time series stationary. According to [28], if we define our original time series as $Y_t$, where $Y$ is the observation at time $t$, for general differencing of order $d$, the operation is defined as:

$$\nabla Y_t = Y_t - Y_{t-1} \tag{4}$$

- Moving average (MA) model: The moving average is expressed in [29] as:

$$y_t = \mu + \mu_t + \theta_1 \mu_{t-1} + \theta_2 \mu_{t-2} + \ldots + \theta_q \tag{5}$$

  where $y_t$ represents the current value of the time series to be predicted, $\mu$ is the mean value of the time series, $\mu_t$ refers to the error term (or residual) at time $t$, and $\theta_i$ represents the coefficients for the moving average terms. The forecasting process of the moving average model involves estimating the coefficients ($\theta_i$) through the utilization of past errors, as evident in Equation (5). It assumes that the current value of a variable is related to the errors made in previous forecasts, and it captures the influence of past forecast errors on the current observation. The order of the moving average component, as denoted by the parameter $q$, represents the number of considered lagged forecast errors.

### 3.3. Walk-Forward Mechanism

Walk-forward optimization is a bipartite method that separates a dataset into in-sample and out-of-sample portions. During each phase of the walk, different segments of the dataset are used for training and testing. This strategy employs a rolling method, where the out-of-sample set is progressively moved forward based on a predetermined window interval to become part of the in-sample dataset in the subsequent phase. It determines a trading strategy's ideal trading parameters over a predetermined time period (referred to as the in-sample or training data) and evaluates their performance over a subsequent time period (referred to as the out-of-sample or testing data). The steps to run this strategy are summarized as follows:

- Obtain all relevant data;
- Divide the data into several parts;
- Run an optimization on the first dataset (first in-sample) to determine the best settings;
- Apply those criteria to the second dataset (first out-of-sample);
- Run an optimization on the upcoming in-sample data to obtain the optimum settings;

- Apply those criteria to the following out-of-sample data;
- Continue until all the data parts have been covered;
- Merge the results of all out-of-sample data.

Figure 1 illustrates an example of walk-forward optimization with six walks. In-sample portions are double the out-of-sample portions (e.g., in-sample can be 1 year whereas out-of-sample can be 6 months). The rolling window is usually chosen as the size of the out-of-sample data.

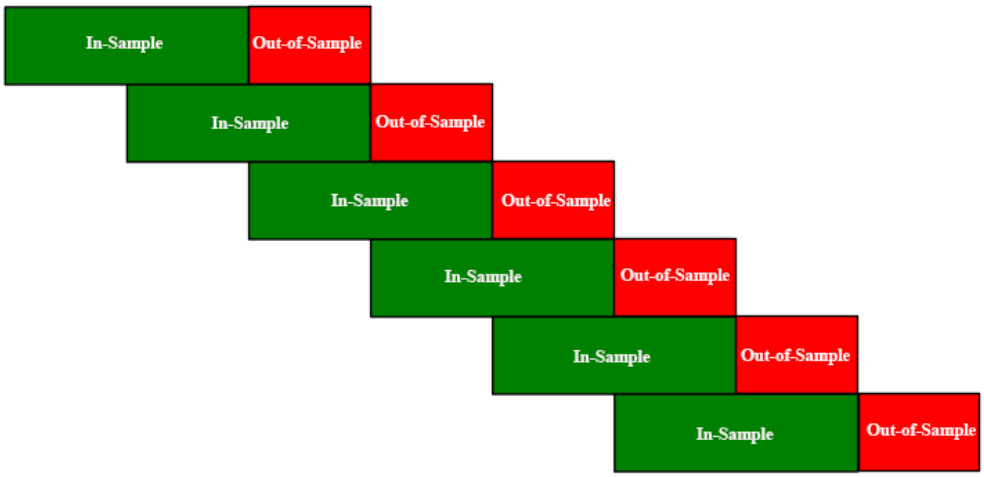


**Figure 1.** Walk-forward optimization chart with double the amount of in-sample data as out-of-sample data.

## 4. The Used Datasets

This section outlines the datasets employed to test our VaR prediction methodology. To present and interpret the outcomes of our proposed model, we adopted several distinct datasets from Yahoo Finance[1], a widely recognized and frequented financial website among traders and business professionals. Yahoo Finance offers an extensive array of financial data and services. It provides live stock quotations, financial market data, and charts for an array of financial instruments, like stocks, bonds, commodities, currencies, and indices. Four historical daily datasets from Yahoo Finance were used: 'S&P500', 'Crude Oil', 'Silver', and 'Gold'. They are described in the following sections.

### *4.1. Standard and Poor's 500*

The Standard and Poor's 500 or S&P 500[2] depicts the cumulative performance of 500 large U.S. companies. It serves as a benchmark for assessing the overall vitality and trend of the U.S. financial market and is a primary indicator for investment in the U.S. financial market. Consequently, this index was included in our model. The dataset we considered comprises 2837 daily observations of S&P stock prices (including close prices) from January 2012 to the middle of April 2023. Figure 2 shows the considered dataset.

### *4.2. Crude Oil*

This index represents the global market price of crude oil[3], a major contributor in the energy market. Spanning from January 2012 to the middle of April 2023, this dataset contains 2830 samples representing the closing prices. Its propensity to significantly change in a short time frame makes it a challenging candidate for VaR prediction. Figure 3 shows the considered dataset.

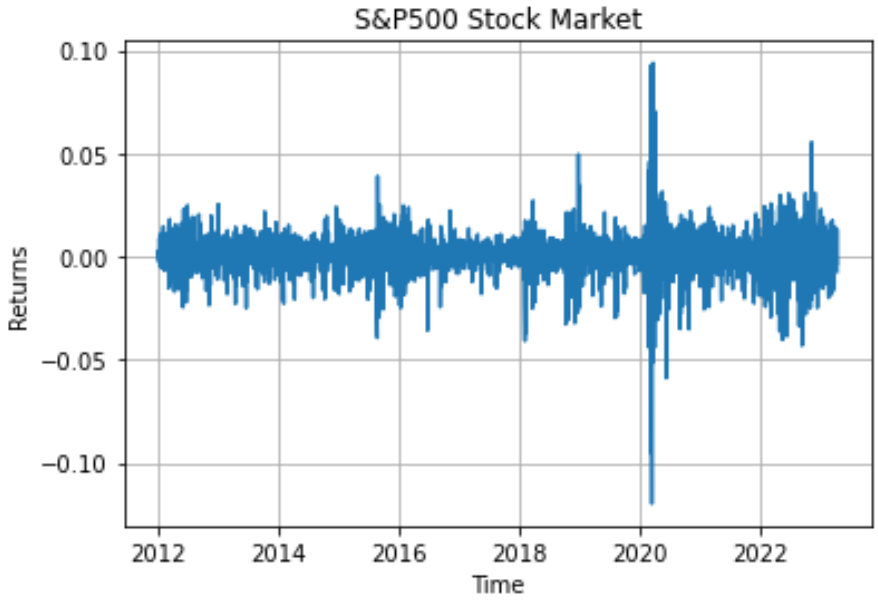


**Figure 2.** S&P 500 stock market returns in USD for the period under analysis.

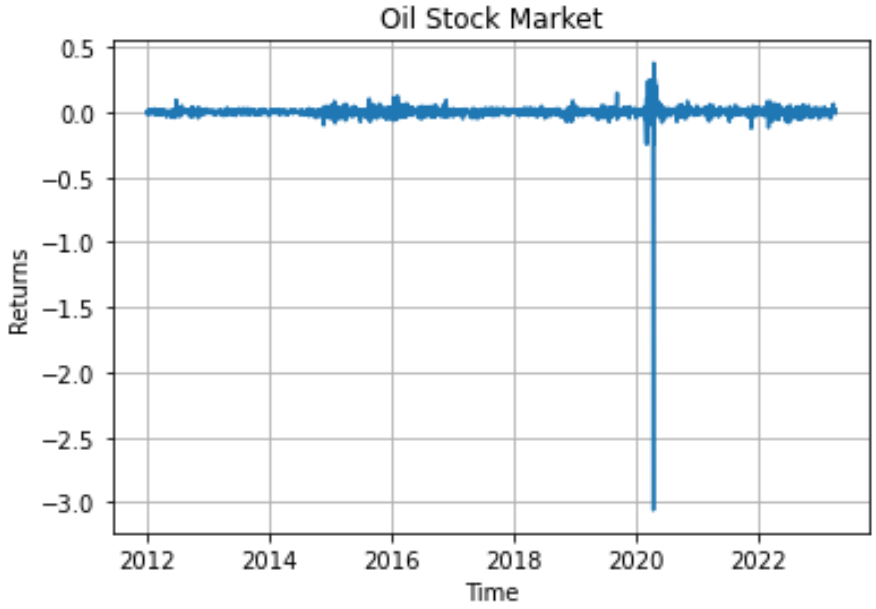


**Figure 3.** Oil stock market returns in USD for the period under analysis.

### *4.3. Silver*

This index represents the price of silver[4] stocks across financial platforms. The silver stock market typically refers to the trading of shares of companies engaged in the silver industry. Covering the period from January 2012 to the middle of April 2023, the dataset consists of 2827 daily closing prices. The peculiarity of these daily closing prices consists of prominent market fluctuations over time, which are crucial for accurate VaR prediction. Figure 4 shows the considered dataset.

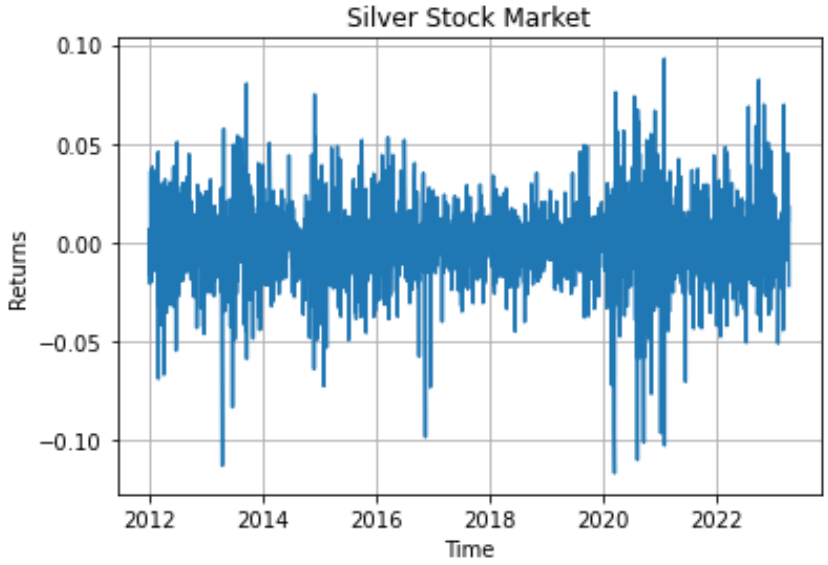


**Figure 4.** Silver stock market returns in USD for the period under analysis.

### *4.4. Gold*

This index supplies data related to the global gold[5] price, enabling market participants to trade a specified quantity of gold. It is frequently traded on various financial markets, such as the New York Mercantile Exchange (NYMEX)[6] or the Chicago Mercantile Exchange (CME)[7]. In this study, we utilized the daily closing prices of gold spanning from January 2012 to the middle of April 2023, consisting of 2824 daily data points. Figure 5 shows the considered dataset.

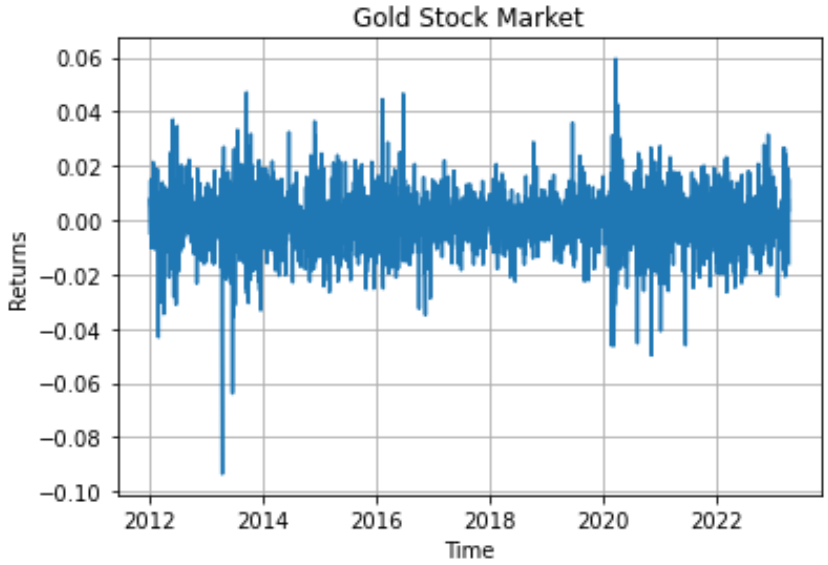


**Figure 5.** Gold Stock market returns in USD for the period under analysis.

## 5. The Proposed Ensemble for Stock Return Prediction

In this section, we provide details our proposed approach to predict market returns. As previously mentioned, we forecast the return for the day ($d$) for which the VaR is being estimated. Then, we use the past real returns and the predicted return for $d$ to compute the VaR estimation for $d$ using an extended version of the baseline strategies or a GARCH model. Then, by using a walk-forward strategy, we compute the VaR estimation for all the data in the OOS dataset. The machine learning approach we leverage takes inspiration

from [30], although it performs a regression task rather than classification. It consists of an ensemble of a set of regressors that are generated automatically after two sets of parameters (hyperparameters and intrinsic parameters) are optimized.

Intrinsic parameters consist of values related to the specific regressor, whereas hyperparameters consist of values related to the dataset, such as the size of the window of the walk-forward strategy.

Hyperparameters are transferred to the training portion of early past data, which we refer to as out-of-sample (OOS) data, once they have been optimized in in-sample (IS) late past data. In order to update the ensemble of regressors to more recent data, these hyperparameters assist in identifying a different set of parameters, referred to as intrinsic parameters, that are optimized. Then, the ensemble is ready to perform the predictions.

Through the proposed two-step ensemble, two sets of parameters are optimized such that the final ensemble can provide a predicted return for any market. First, using late past data from an IS dataset, we improve the hyperparameters while taking minimum square error (MSE) into account. These hyperparameters are then transferred to create ensembles for an early past (OOS) dataset. The validation portion of the current data updates the intrinsic parameters of each individual regressor before the final ensemble is constructed.

In the IS data and some of the OOS data, we use the non-anchored walk-forward approach to discover the optimum ensemble hyperparameters.

We used three regressors: gradient boosting, support vector machine, and random forest. The intrinsic parameters we had to find were chosen from the list shown in Table 1, whereas the hyperparameters to identify were chosen from the list shown in Table 2.

**Table 1.** Intrinsic parameter grid.

| Algorithm | Parameter | Values | Description |
|---|---|---|---|
| Gradient boosting | n_estimators | 10, 25, 50, 100 | Boosting stages to perform |
| | learning_rate | 0.0001, 0.001, 0.01, 0.1 | Contribution of each tree |
| | max_depth | 2, 4, 6, 8, 10 | Maximum depth of each estimator |
| Support vector machines | max_iter | 20, 50, 100 | Hard limit of iterations within solver |
| | tol | 0.0001, 0.001, 0.01, 0.1 | Tolerance for stopping criterion |
| | C | 1, 10, 20, 50 | Penalty of the error term |
| | gamma | 0.0001, 0.001, 0.01, 0.1 | Coefficient for the used kernel |
| Random forests | n_estimators | 20, 50, 100 | Trees in the forest |
| | max_depth | 1, 5, 10, 50 | Max depth of the tree |
| | min_samples_split | 0.2, 0.4, 0.8, 1.0 | Min samples to split a node |

**Table 2.** Hyperparameter grid.

| Parameter | Values | Description |
|---|---|---|
| window_size | 100, 150, 200, 250, 300 | Days used for the training set |
| train_size | 60, 65, 70, 75, 80 | Percentage of window_size used for the training set |
| lags | 1, 3, 5, 7, 9 | Previous days to use in order to predict the return |

The usefulness of ensemble techniques that use various algorithms is supported by numerous literature reports [31,32]. In many famous machine learning competitions (such as Kaggle, the Netflix Competition, KDD, and others), ensemble techniques typically produce the best results [33].

As a result, we used an ensemble learning approach, in which the final prediction is produced by merging the outputs of individual algorithms within the ensemble. An ensemble process like this can function independently or dependently. Since we take the independent framework approach, each classifier decision may be viewed as a separate vote from the others. We applied this scheme to the three algorithms we used (gradient boosting,

support vector machines, and random forests), with their ensemble hyperparameters initially found in the IS data and whose individual intrinsic parameters were found in the OOS data. Any other regressor can be added to our initial list. We chose them because they have already been successfully applied in the literature for return predictions [34–36].

The aggregation criterion we employed in our ensemble is the simple average.

To provide a quality assessment, we tested the proposed machine learning approach for return prediction in the four markets illustrated in Section 4 by using 1 year for the IS data and 1 month for OOS and compared the results against ARIMA (12,0,12). We computed the MSE between the predicted returns and the real returns. The MSE is computed as:

$$MSE = \frac{1}{n}\sum_{i=1} n(Y_i - \hat{Y}_i)$$

where $n$ is the number of data points, $Y$ is the vector of the real returns, and $\hat{Y}_i$ is the vector of the predicted returns.

Table 3 shows the obtained results (averaged along all elements in the OOS dataset), where we can notice how our proposed ensemble approach outperforms ARIMA for the four illustrated markets.

**Table 3.** Results in terms of MSE of stock market prediction.

| Market | ARIMA | Ensemble |
|---|---|---|
| S&P stock market | 0.00653 | 0.00589 |
| Oil stock market | 0.00712 | 0.00601 |
| Silver stock market | 0.00815 | 0.00612 |
| Gold stock market | 0.01247 | 0.00901 |

## 6. Performance Evaluation

This section presents a comprehensive evaluation of the approaches we propose for VaR estimation and a comparison against the baselines according to a set of metrics established in the literature.

### *6.1. Baselines*

As mentioned in Section 1, we considered the three following baselines for VaR estimation: normal distribution, historical simulation, and EWMA.

#### 6.1.1. Normal Distribution

Normal distribution is a parametric method that relies on specific rules or parameters to estimate VaR. It is particularly suited for portfolios with linear positions and assumes that asset returns adhere to a normal distribution. This assumption is largely derived from the central limit theorem, suggesting that the aggregate of numerous independent and identically distributed random variables approximates a normal distribution. In order to calculate the VaR, this method uses a historical dataset to obtain the mean and the standard deviation, which are computed within a designated time period. Once we have the mean and standard deviation, we can compute the VaR. A unique feature of this approach is that volatility is quantified in terms of standard deviation. One significant advantage of this technique is its capacity to deliver precise results [37] by leveraging just these two quantities. The following formula is used to calculate the VaR at a confidence level of $\alpha$ using the normal distribution approach:

$$VaR_\alpha = \mu + \sigma Z \quad (6)$$

where $Z$ represents the z score or standard normal deviation for the given confidence level ($\alpha$). This is directly tied to the reliance of the normal distribution method on the normal

distribution assumption. In other words, $Z$ is a value from the standard normal distribution corresponding to the desired confidence level. The $Z$ value can also be calculated according to the following equation:

$$Z_{=}\frac{VaR_{\alpha}-\mu}{\sigma} \tag{7}$$

where $\mu$ and $\sigma$ are the mean and standard deviation, respectively. The equation indicates that the $Z$ score is the number of standard deviations that the VaR is away from the mean. This interpretation further illustrates the emphasis of the normal distribution method on using standard deviation as a measure of risk and volatility.

Nonetheless, this method has its limitations. It assumes that returns are normally distributed, which may not always be true. Furthermore, due to its reliance on the left tail of the portfolio's normal distribution, the normal distribution tends to underestimate the exact VaR and the proportion of outliers. It may also underestimate the VaR when a high confidence level is applied [38]. Despite these limitations, it performs rather well when there is a linear association between portfolio positions and risk [39].

#### 6.1.2. Historical Simulation

Classified under non-parametric methods, the historical simulation approach employs empirical distribution and historical data for VaR estimation [40]. Essentially, it ranks the previous returns, and based on the target probability, it identifies the corresponding quantile of the distribution. The underlying assumption here is that current market conditions mimic future scenarios, resulting in similar outcomes [41]. A significant advantage of historical simulation is its simplicity [42]. It conveys that any change in the portfolio provides all the necessary information for computing VaR, eliminating the need to calculate variance and covariance. Capable of handling non-linear and non-normal portfolio distribution, it does not depend on any specific assumption. Historical simulation is comparable to an equally weighted moving average, assigning equal weights to each data point [39]. Unlike normal distribution, there is no need to assume that data are normally distributed. However, it has certain limitations, such as the need for a lot of observations. Historical simulation delivers accurate results, particularly under high confidence levels [38], but the premise that past market conditions accurately reflect future situations can sometimes be flawed. Although it is simpler and more reliable than the normal distribution method, it is more time-consuming [43]. To compute VaR using historical simulation, the following mathematical formula can be used [41]:

$$R_t^p = \sum_{i=1}^{n} W_i R_{i,t} \qquad t = 0, ..., T \tag{8}$$

where $R_t^p$ represents the return of the portfolio at time $t$. The term $(t)$ refers to a specific time period in the dataset that ranges from 0 to $T$, where $T$ represents the total number of time periods within the dataset. $R_{i,t}$ represents the return of asset $i$ at time $t$. $W_i$ is the weight attributed to asset $i$ in the portfolio, and $n$ denotes the number of assets in the portfolio. After computing the possible future return values for all time periods, an empirical distribution of potential portfolio returns is constructed. By ranking these returns from the lowest to the highest, it is possible to find the VaR at a specific confidence level. For instance, to calculate the 1-day VaR at a 95% confidence level, the fifth percentile of the ranked distribution of $R_t^p$ is identified. This means that based on the historical data, the portfolio is expect to return to be at or below this VaR value 5% of the time.

#### 6.1.3. EWMA

The EWMA model was introduced in [44] and assumes that the returns follow a normal distribution. Unlike equal weightage methods, EWMA assigns varying weights to observations depending on their relative recency and position within a given period. It prioritizes recent observations in the calculation of VaR, attributing more weight to

recent returns than older ones. This weighting scheme is based on the rationale that recent returns and market trends are likely to resemble future market conditions more closely. Furthermore, uncommon volatility in a given period could influence volatility in subsequent periods; hence, such periods are given higher weightage. This translates to recent observations having a greater impact on the VaR calculation in the EWMA method. The formula for the EWMA method introduced in [44], is defined as:

$$\sigma_t^2 = \lambda\sigma_{t-1}^2 + (1-\lambda)r_{t-1}^2 \tag{9}$$

In this formula, $\sigma_{t-1}$ and $r_{t-1}$ denote the volatility of returns and returns at time $t-1$, respectively. As per this formula, the volatility on day $t-1$ is employed to forecast the volatility for the day ahead, while $\lambda$ $(0 < \lambda < 1)$ represents the decay factor, which plays a crucial role in the EWMA method by assigning different weights to past returns depending on how recent they are. The more recent the return, the more weight it carries in the calculation of future volatility.

Despite its simplicity, the EWMA model has been found to perform remarkably well when compared to more complex conditional volatility models in VaR calculations [45–47]. By allocating higher weights to more recent data, it overcomes the limitation of assigning equal weight to all observations, a problem inherent in simple historical simulation methods. The EWMA model is adept at quickly adjusting to changes in market conditions, thanks to its sensitivity to recent returns. However, similar to the normal distribution method, it may also undertake extreme losses.

### *6.2. Used Metrics*

Backtesting is a procedure utilized to assess the proficiency and precision of VaR estimation methodologies prior to the deployment of actual capital. Backtesting offers crucial insights into the strengths and shortcomings of VaR estimation methods, enhancing risk management practices by quantifying the number of failures. Failure or exception in backtesting methods means that the actual losses have exceeded the predicted losses, indicating a prediction shortfall. The objective of backtesting lies in the verification of the alignment between the method and the model's assumptions. The Basel Accords[8] mandated banks to appraise the performance of VaR estimators through the application of diverse backtesting methodologies.

Time between failures (TBF) is an advanced backtesting methodology deployed in financial risk management specifically to gauge the accuracy of VaR estimators. TBF is a combination of two tests: the time between failures independence test (TBFI) and the probability of failure test (POF). TBFI tests for time independence: a high test statistic is associated with clustered violations. POF tests whether the overall frequency of failures is close to the reciprocal of the VaR confidence interval. The TBF is therefore a combined test of frequency and clustering. POF is a likelihood ratio test measuring whether the overall empirical probability matches the anticipated probability. Its test statistic is:

$$LRatioPOF = -2\log\left(\frac{pVar^x(1-pVaR)^{N-x}}{\left(\frac{x}{N}\right)^x\left(1-\frac{x}{N}\right)^{N-x}}\right) \tag{10}$$

where $x$ is the total number of failures, and $N$ is the total number of observations. This is essentially a test of whether the observed confidence level ($pVaR$) is close to the intended level. The $LRatioPOF$ is also chi-square distributed with 1 degree of freedom.

TBFI is an extension of the Kupiec TUFF test [48] and was proposed by Haas in 2001 [49]. The null hypothesis in this method postulates that exceptions are independent of each other. It measures the time until the first failure or exception and also calculates the time between subsequent failures. It builds upon Christoffersen's ideas [50] and

incorporates them into a more powerful metric. TBFI is a likelihood ratio test defined in [51] as:

$$LRatioTBFI = -2\sum_{i=1}^{x}\log\left(\frac{p(1-p)^{n_i-1}}{\left(\frac{1}{n_i}\right)\left(1-\frac{1}{n_i}\right)^{n_i-1}}\right) \quad (11)$$

where $x$ is the number of failures, $n_i$ represents the time between the $i$-th failure and the $(i-1)$-th failure, and $p$ is the reciprocal of the VaR confidence level ($pVaR$) so that $p = 1 - pVaR$. TBFI is a test of whether the statistic exceeds some critical level given by a chi-square distribution with degrees of freedom equal to the number of failures. Hence, the higher the $LRatioTBFI$, the more clustered the failures and the less useful the tested method is for predicting VaR.

The methods with lower $LRatioTBFI$ values generally exhibit less clustering and an acceptable frequency of failures and should therefore be preferred. Therefore, we score the methods using $LRatioTBFI$, and the best methods will have lower $LRatioTBFI$ values.

Besides $LRatioTBFI$, in our experiments, we report the number of failures, TBFMin, TBFQ1, TBFQ2, TBFQ3, and TBFMax. They are statistical measures related to the observed intervals between failures. They stand for the minimum value, first quartile, second quartile, third quartile, and maximum value of these observed intervals, respectively.

### *6.3. VaR Estimation*

This section shows the results we obtained in our study. To obtain the prediction of returns for the considered datasets, we used the ARIMA (12, 0, 12) technique and the ensemble proposed in Section 5 with the walk-forward mechanism mentioned in Section 3. For VaR estimation, we employed the walk-forward strategy with windows corresponding to 250 days for the IS and 1 day for the OOS.

Let us remark that, as mentioned in Section 4, the four used datasets included 2837, 2830, 2827, and 2824 samples for the S&P 500, oil, silver, and gold, respectively, with a final number of walks of 2586, 2579, 2576, and 2573, respectively.

The baselines (normal, historical, and EWMA) are referred to as $NameOfBaseline\_ConfidenceValue$.

The predicted returns were integrated into all the baselines to compute the VaR, obtaining three new methods (in the following, referred to as $NameOfBaseline\_ConfidenceValue\_PredictionMethod$, where $NameOfBaseline$ is normal, historical, or EWMA; $ConfidenceValue$ is either 95 or 99; and $PredictionMethod$ is either ARIMA or ENSEMBLE). To compute the VaR estimation for a day ($d$) we compute the predicted return for $d$, then feed one of the algorithms the predicted return and the real returns of the days before $d$. Then, we repeat the process for day $d+1$ and so on. More specifically, the proposed approaches leverages the predicted returns of day $i$ in their calculation of the VaR for day $i$. In such cases, for each walk, the 250 values of each IS consisted of the returns from day $i-249$ to day $i-1$ plus the predicted return for day $i$. Basically, the return of the first day of each walk is discarded.

The other approach we propose, which exploits PanelTime, uses the same walks and the same data (predicted return of $d$ and real returns of days before $d$) and combines them using the GARCH model.

All the experiments were run for two levels of confidence: 95% and 99%.

We also assessed the EWMA method (and our proposed version that takes the predicted returns) with distinct values for the decay factor ($\lambda$). We confirmed its best value as 0.94, as mentioned in [52]. In the results shown in the tables below, the decay value of the EWMA method is expressed at the end of the name in the Used Method column.

### *6.4. Results*

Table 4 illustrates the results pertaining to the S&P 500 stock market sorted by increasing values of the LRatioTBFI metric (the results for TBFMin, TBFQ1, TBFQ2, TBFQ3, and TBFMAX for this and the other markets were rounded to the closest integer).

**Table 4.** LRatioTBFI backtest results for the S&P 500 stock market.

| Used Method | LRatioTBFI | Failures | TBFMin | TBFQ1 | TBFQ2 | TBFQ3 | TBFMax |
|---|---|---|---|---|---|---|---|
| Paneltime_99_ENSEMBLE | 40.120 | 21 | 1 | 1 | 7 | 31 | 164 |
| Paneltime_99_ARIMA | 41.813 | 23 | 1 | 1 | 8 | 32 | 170 |
| Historical_99_ENSEMBLE | 149.932 | 43 | 1 | 2 | 17 | 68 | 430 |
| Historical_99_ARIMA | 149.932 | 43 | 1 | 2 | 17 | 68 | 430 |
| Historical_99 | 197.898 | 53 | 1 | 2 | 7 | 48 | 547 |
| Paneltime_95_ENSEMBLE | 222.193 | 63 | 1 | 2 | 6 | 39 | 316 |
| Paneltime_95_ARIMA | 244.809 | 65 | 1 | 2 | 7 | 40 | 320 |
| Normal_99_ENSEMBLE | 277.689 | 72 | 1 | 2 | 7 | 43 | 324 |
| Normal_99_ARIMA | 280.258 | 76 | 1 | 2 | 8 | 45 | 327 |
| Normal_99 | 343.423 | 90 | 1 | 2 | 7 | 35 | 351 |
| Historical_95_ENSEMBLE | 344.921 | 119 | 1 | 2 | 4 | 23 | 133 |
| Historical_95_ARIMA | 344.109 | 139 | 1 | 2 | 5 | 25 | 134 |
| Normal_95_ENSEMBLE | 350.790 | 143 | 1 | 2 | 5 | 17 | 166 |
| Normal_95_ARIMA | 355.022 | 150 | 1 | 2 | 5 | 18 | 172 |
| Historical_95 | 419.061 | 176 | 1 | 2 | 5 | 14 | 169 |
| Normal_95 | 420.112 | 175 | 1 | 2 | 5 | 14 | 169 |
| EWMA_99_0.94 | 530.177 | 157 | 1 | 4 | 10 | 25 | 66 |
| EWMA_95_0.94 | 565.432 | 282 | 1 | 2 | 6 | 13 | 61 |
| EWMA_99_0.3 | 968.365 | 268 | 1 | 4 | 8 | 13 | 39 |
| EWMA_99_0.2 | 1088.338 | 289 | 1 | 4 | 7 | 12 | 39 |
| EWMA_99_0.1 | 1198.688 | 308 | 1 | 4 | 7 | 11 | 31 |
| EWMA_95_ENSEMBLE_0.94 | 1306.871 | 440 | 1 | 1 | 3 | 6 | 110 |
| EWMA_95_ARIMA_0.94 | 1316.012 | 452 | 1 | 1 | 3 | 6 | 114 |
| EWMA_99_ENSEMBLE_0.94 | 1615.671 | 340 | 1 | 2 | 3 | 7 | 116 |
| EWMA_99_ARIMA_0.94 | 1900.344 | 342 | 1 | 2 | 3 | 7 | 118 |

For both the confidence values (95% and 99%), the best approaches are the PanelTime and historical approaches with predicted returns. In fact, for the confidence value of 99% PanelTime is the best method with either the ensemble or ARIMA, followed by the historical method with either the ensemble or ARIMA, the first baseline, i.e., the historical method. For a confidence value of 95%, we noticed a similar trend: PanelTime (either with the ensemble or ARIMA) achieves the best performance, followed by the historical method (either with the ensemble or ARIMA), the historical method, and the normal method. The normal method combined with the predicted returns (either ensemble or ARIMA) performs well and better than its baseline counterparts for both confidence values. Neither return prediction method (ARIMA and ensemble) seems to produce good results when combined with the EWMA method. In general and for all the methods, the predictions returned by the ensemble seem to result in better VaR estimates than those returned by ARIMA.

Table 5 show the results pertaining to the crude oil sorted by increasing values of LRatioTBFI metric. Even in this case, our proposed approaches (predicted returns integrated with PanelTime, historical, or normal methods) beat the baselines. In particular, for the confidence value equal to 99%, PanelTime (either with the ensemble or ARIMA) achieves

the best performance, followed by the historical method integrated with the predicted returns (either with the ensemble or ARIMA) and the normal method integrated with the predicted returns (with either the ensemble or ARIMA). The first baseline is the historical method. For the confidence value of 95%, PanelTime confirms its superiority, followed by the normal method integrated with the predicted returns and the historical method integrated with the predicted returns. The first baseline for such a confidence value is the normal method. Furthermore, for the oil stock market, in all cases except when using the normal method for the confidence value equal to 99%, the ensemble approach provides better VaR estimation than ARIMA.

**Table 5.** LRatioTBFI backtest results for the oil stock market.

| Used Method | LRatioTBFI | Failures | TBFMin | TBFQ1 | TBFQ2 | TBFQ3 | TBFMax |
|---|---|---|---|---|---|---|---|
| Paneltime_99_ENSEMBLE | 45.310 | 2 | 1 | 1 | 3 | 50 | 430 |
| Paneltime_99_ARIMA | 47.663 | 3 | 1 | 2 | 4 | 52 | 429 |
| Historical_99_ENSEMBLE | 145.110 | 43 | 1 | 3 | 8 | 56 | 367 |
| Historical_99_ARIMA | 151.813 | 46 | 1 | 4 | 11 | 63 | 379 |
| Normal_99_ARIMA | 208.652 | 57 | 1 | 3 | 9 | 34 | 442 |
| Normal_99_ENSEMBLE | 209.167 | 55 | 1 | 3 | 8 | 35 | 448 |
| Paneltime_95_ENSEMBLE | 209.259 | 3 | 1 | 2 | 3 | 61 | 390 |
| Paneltime_95_ARIMA | 209.578 | 3 | 1 | 2 | 4 | 43 | 378 |
| Historical_99 | 295.018 | 68 | 1 | 2 | 5 | 16 | 693 |
| Normal_95_ENSEMBLE | 303.610 | 113 | 1 | 3 | 5 | 11 | 309 |
| Normal_95_ARIMA | 310.868 | 140 | 1 | 3 | 6 | 15 | 312 |
| Historical_95_ENSEMBLE | 319.671 | 142 | 1 | 3 | 5 | 17 | 275 |
| Historical_95_ARIMA | 324.145 | 148 | 1 | 3 | 5 | 16 | 269 |
| Normal_99 | 363.302 | 81 | 1 | 2 | 5 | 12 | 419 |
| Normal_95 | 386.712 | 151 | 1 | 2 | 5 | 10 | 367 |
| Historical_95 | 465.254 | 205 | 1 | 2 | 5 | 11 | 223 |
| EWMA_99_0.94 | 516.944 | 159 | 1 | 4 | 11 | 20 | 99 |
| EWMA_95_0.94 | 550.936 | 308 | 1 | 3 | 6 | 11 | 75 |
| EWMA_99_0.3 | 1209.963 | 300 | 1 | 4 | 6 | 11 | 65 |
| EWMA_95_ENSEMBLE_0.94 | 1268.127 | 309 | 1 | 5 | 6 | 11 | 78 |
| EWMA_99_0.2 | 1274.892 | 313 | 1 | 4 | 6 | 10 | 65 |
| EWMA_95_ARIMA_0.94 | 1395.281 | 446 | 1 | 1 | 3 | 5 | 147 |
| EWMA_99_ENSEMBLE_0.94 | 1399.112 | 301 | 1 | 1 | 3 | 5 | 157 |
| EWMA_99_0.1 | 1405.215 | 333 | 1 | 3 | 6 | 9 | 65 |
| EWMA_99_ARIMA_0.94 | 1822.801 | 316 | 1 | 1 | 3 | 5 | 163 |

The results for the silver stock market sorted by increasing values of the LRatioTBFI metric are displayed in Table 6. The proposed PanelTime method (with either the ensemble or ARIMA) is confirmed to outperform the baselines for the confidence value of 99%. For this market, the historical method performs better than its integrated version with the predicted returns. The presence of several fluctuations in the silver stock market is the likely reason for this behavior. On the other hand, the normal approach integrated with predicted returns beats its baseline version. For the confidence value of 95%, the normal method integrated with the predicted returns (either the ensemble or ARIMA) is the best approach for VaR estimation, followed by its baseline version. Similarly, the historical method integrated with the predicted returns beats its baseline counterpart.

**Table 6.** LRatioTBFI backtest results for the silver stock market.

| Used Method | LRatioTBFI | Failures | TBFMin | TBFQ1 | TBFQ2 | TBFQ3 | TBFMax |
|---|---|---|---|---|---|---|---|
| Paneltime_99_ENSEMBLE | 38.750 | 26 | 1 | 2 | 14 | 60 | 395 |
| Paneltime_99_ARIMA | 41.813 | 30 | 1 | 2 | 15 | 65 | 410 |
| Historical_99 | 149.932 | 43 | 1 | 2 | 17 | 68 | 430 |
| Historical_99_ARIMA | 156.359 | 44 | 1 | 2 | 14 | 67 | 428 |
| Historical_99_ENSEMBLE | 167.119 | 49 | 1 | 2 | 17 | 68 | 431 |
| Normal_99_ENSEMBLE | 237.014 | 52 | 1 | 2 | 5 | 32 | 297 |
| Normal_99_ARIMA | 244.809 | 59 | 1 | 2 | 6 | 39 | 301 |
| Normal_99 | 304.284 | 81 | 1 | 2 | 7 | 43 | 327 |
| Normal_95_ENSEMBLE | 335.097 | 86 | 1 | 2 | 5 | 30 | 248 |
| Normal_95_ARIMA | 343.423 | 90 | 1 | 2 | 7 | 35 | 351 |
| Normal_95 | 355.022 | 150 | 1 | 2 | 5 | 18 | 172 |
| Historical_95_ENSEMBLE | 359.230 | 135 | 1 | 2 | 4 | 20 | 130 |
| Historical_95_ARIMA | 362.303 | 144 | 1 | 2 | 5 | 21 | 134 |
| Historical_95 | 419.061 | 176 | 1 | 2 | 5 | 14 | 169 |
| Paneltime_95_ENSEMBLE | 420.010 | 174 | 1 | 2 | 5 | 13 | 169 |
| Paneltime_95_ARIMA | 420.112 | 175 | 1 | 2 | 5 | 14 | 169 |
| EWMA_95 | 530.177 | 157 | 1 | 4 | 10 | 26 | 66 |
| EWMA_99 | 565.432 | 282 | 1 | 2 | 6 | 13 | 61 |
| EWMA_95_ENSEMBLE_0.94 | 798.012 | 243 | 1 | 3 | 6 | 11 | 41 |
| EWMA_99_0.3 | 968.365 | 268 | 1 | 4 | 8 | 13 | 39 |
| EWMA_99_ENSEMBLE_0.94 | 1043.797 | 276 | 1 | 3 | 4 | 10 | 43 |
| EWMA_99_0.2 | 1088.338 | 289 | 1 | 4 | 7 | 12 | 39 |
| EWMA_95_ARIMA_0.94 | 1198.688 | 308 | 1 | 4 | 7 | 11 | 31 |
| EWMA_99_0.1 | 1346.119 | 457 | 1 | 1 | 3 | 6 | 114 |
| EWMA_99_ARIMA_0.94 | 1900.344 | 342 | 1 | 2 | 3 | 7 | 118 |

Finally, the results for the gold stock market sorted by increasing values of the LRatioTBFI metric are displayed in Table 7. As previously seen in the other markets, PanelTime (either with the ensemble or ARIMA) is the best method for the confidence value of 99%, followed by the historical approach with the predicted returns (with either the ensemble or ARIMA) and its baseline version. For a confidence value equal to 95%, the normal approach integrated with predicted returns (with either the ensemble or ARIMA) and the historical approach integrated with predicted returns (with either the ensemble or ARIMA) are the best methods, followed by their baseline counterparts, as in the previous market using PanelTime (with the Ensemble and ARIMA). In this market, the ensemble seems to provide better values than ARIMA when used to predict returns.

Different Python libraries such as Scikit-learn[9], `Numpy`[10], `Pandas`[11], `Statsmodels`[12], `Scipy`[13], and `Matplotlib`[14] were used to develop the ARIMA and ensemble approaches and calculate the metrics used for the backtest methodology for VaR estimation. We also leveraged the `Risk Management Toolbox`[15] of MatLab for all the metrics related to VaR estimation.

**Table 7.** LRatioTBFI backtest results for the Gold stock market.

| Used Method | LRatioTBFI | Failures | TBFMin | TBFQ1 | TBFQ2 | TBFQ3 | TBFMax |
|---|---|---|---|---|---|---|---|
| Paneltime_99_ENSEMBLE | 30.091 | 24 | 1 | 7 | 30 | 87 | 312 |
| Paneltime_99_ARIMA | 41.813 | 28 | 1 | 13 | 40 | 94 | 389 |
| Historical_99_ENSEMBLE | 54.185 | 30 | 1 | 13 | 40 | 97 | 397 |
| Historical_99_ARIMA | 61.681 | 31 | 1 | 15 | 45 | 101 | 401 |
| Historical_99 | 64.585 | 21 | 1 | 8 | 35 | 70 | 817 |
| Normal_99 | 81.221 | 33 | 1 | 9 | 34 | 59 | 700 |
| Normal_99_ENSEMBLE | 82.917 | 38 | 1 | 12 | 37 | 68 | 209 |
| Normal_99_ARIMA | 83.000 | 46 | 1 | 14 | 40 | 73 | 217 |
| Normal_95_ENSEMBLE | 139.290 | 117 | 1 | 6 | 14 | 29 | 101 |
| Normal_95_ARIMA | 158.421 | 129 | 1 | 6 | 14 | 30 | 107 |
| Historical_95_ENSEMBLE | 160.109 | 131 | 1 | 5 | 13 | 28 | 91 |
| Historical_95_ARIMA | 162.279 | 138 | 1 | 5 | 14 | 29 | 95 |
| Historical_95 | 181.480 | 130 | 1 | 5 | 13 | 29 | 156 |
| Normal_95 | 185.255 | 103 | 1 | 5 | 12 | 32 | 261 |
| Paneltime_95_ENSEMBLE | 212.475 | 98 | 1 | 4 | 8 | 13 | 82 |
| Paneltime_95_ARIMA | 244.809 | 103 | 1 | 4 | 8 | 13 | 87 |
| EWMA_95 | 394.123 | 274 | 1 | 4 | 7 | 14 | 45 |
| EWMA_99 | 428.975 | 147 | 1 | 6 | 14 | 26 | 80 |
| EWMA_99_0.3 | 1352.436 | 327 | 1 | 3 | 6 | 11 | 36 |
| EWMA_99_0.2 | 1484.849 | 351 | 1 | 3 | 6 | 10 | 34 |
| EWMA_95_ENSEMBLE_0.94 | 1509.104 | 513 | 1 | 2 | 3 | 5 | 71 |
| EWMA_95_ARIMA_0.94 | 1587.327 | 555 | 1 | 2 | 3 | 5 | 76 |
| EWMA_99_0.1 | 1657.845 | 375 | 1 | 3 | 5 | 10 | 31 |
| EWMA_99_ENSEMBLE_0.94 | 2168.110 | 409 | 1 | 2 | 3 | 7 | 95 |
| EWMA_99_ARIMA_0.94 | 2200.140 | 415 | 1 | 2 | 3 | 7 | 97 |

## 7. Conclusions and Future Directions

In this paper, we considered the problem of VaR estimation. VaR modeling determines the potential for loss in the entity being analyzed, as well as the likelihood that the specified loss will occur. Using VaR in risk measurement has a number of benefits. It is a single number that can be easily understood, is frequently used by experts in the financial sector, and can be stated as a percentage or in price units. VaR calculations can be compared across a variety of asset classes or portfolios, including shares, bonds, derivatives, currencies, and more. VaR is frequently featured and calculated in different financial software tools due to its popularity. We discussed the baseline approaches usually used to calculate VaR for different confidence values. Then, we presented different machine learning regressors to predict stock market returns and indicated how to benefit from the combination of predicted returns and real returns by extending the baselines or using a GARCH model. One regressor that we employed is the well-known ARIMA, and another regressor that we proposed in this paper is an ensemble of different machine learning approaches that operates in two steps: the first step is used to tune hyperparameters on an IS set of data. Then, the identified hyperparameters are transferred to create ensembles for an early past OOS dataset. The validation portion of this set updates the intrinsic parameters of each individual regressor before the final ensemble is built. For each baseline used to estimate the VaR, we created an extended version that integrates the predicted return (using either ARIMA or the ensemble of regressors) for day ($d$) for which the VaR was being estimated with the real returns of the days before $d$. We also proposed PanelTime, which, to the best of

our knowledge, is the only Python package that can take additional regressors into account in the GARCH model and combine them with stock returns to compute VaR estimations. The experiments that we carried out indicate, according to the proposed metrics, that our proposed methods are always superior to the baselines; therefore, the predicted return information is useful for a more precise VaR computation. In future work, we would like to extend the ensemble with more regressors consisting of deep learning approaches. Moreover, an analysis of the best deep learning approaches (e.g., transformer-based) will be carried out in order to identify the most promising approaches that can work well in financial markets.

**Author Contributions:** Conceptualization, F.B., D.R.R. and E.S.; methodology, D.R.R. and E.S.; software, F.B. and E.S.; validation, F.B. and D.R.R.; formal analysis, D.R.R.; investigation, F.B., D.R.R. and E.S.; writing—original draft preparation, F.B., D.R.R. and E.S.; writing—review and editing, D.R.R. and E.S.; supervision, D.R.R. and E.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5-Call for tender No.3277 published on 30 December 2021 from the Italian Ministry of University and Research (MUR) funded by the European Union—NextGenerationEU. Project Code ECS0000038—Project Title eINS Ecosystem of Innovation for Next Generation Sardinia—CUP F53C22000430001-Grant Assignment Decree No. 1056 adopted on 23 June 2022 by the Italian Ministry of University and Research (MUR).

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Notes

1 https://finance.yahoo.com/, accessed on 2 July 2023.
2 https://finance.yahoo.com/quote/%5EGSPC, accessed on 2 July 2023.
3 https://finance.yahoo.com/quote/CL=F?p=CL=F&.tsrc=fin-srch, accessed on 2 July 2023.
4 https://finance.yahoo.com/quote/SI=F?p=SI=F&.tsrc=fin-srch, accessed on 2 July 2023.
5 https://finance.yahoo.com/quote/GC=F?p=GC=F&.tsrc=fin-srch, accessed on 2 July 2023.
6 https://www.cmegroup.com/company/nymex.html, accessed on 2 July 2023.
7 https://www.cmegroup.com/, accessed on 2 July 2023.
8 https://en.wikipedia.org/wiki/Basel_Accords, accessed on 2 July 2023.
9 https://scikit-learn.org, accessed on 2 July 2023.
10 https://numpy.org, accessed on 2 July 2023.
11 https://pandas.pydata.org/, accessed on 2 July 2023.
12 https://www.statsmodels.org/stable/index.html, accessed on 2 July 2023.
13 https://scipy.org/, accessed on 2 July 2023.
14 https://matplotlib.org, accessed on 2 July 2023.
15 https://it.mathworks.com/help/risk/, accessed on 2 July 2023.

## References

1. Gallati, R.R. (Ed.) *Risk Management and Capital Adequacy*; McGraw-Hill: New York, NY, USA, 2003.
2. Sharma, M. Evaluation of Basel III revision of quantitative standards for implementation of internal models for market risk. *IIMB Manag. Rev.* **2012**, *24*, 234–244. [CrossRef]
3. Engle, R. GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics. *J. Econ. Perspect.* **2001**, *15*, 157–168. [CrossRef]
4. Carta, S.M.; Consoli, S.; Podda, A.S.; Recupero, D.R.; Stanciu, M.M. Ensembling and Dynamic Asset Selection for Risk-Controlled Statistical Arbitrage. *IEEE Access* **2021**, *9*, 29942–29959. [CrossRef]
5. Xu, Q.; Jiang, C.; He, Y. An exponentially weighted quantile regression via SVM with application to estimating multiperiod VaR. *Stat. Methods Appl.* **2016**, *25*, 285–320. [CrossRef]
6. Khan, M.A.I. Modelling daily value-at-risk using realized volatility, non-linear support vector machine and ARCH type models. *J. Econ. Int. Financ.* **2011**, *3*, 305.

7. Takeda, A.; Fujiwara, S.; Kanamori, T. Extended robust support vector machine based on financial risk minimization. *Neural Comput.* **2014**, *26*, 2541–2569. [CrossRef]
8. Radović, O.; Stanković, J. Tail risk assessment using support vector machine. *J. Eng. Sci. Technol. Rev.* **2015**, *8*, 61–64. [CrossRef]
9. Lux, M.; Härdle, W.K.; Lessmann, S. Data driven value-at-risk forecasting using a SVR-GARCH-KDE hybrid. *Comput. Stat.* **2020**, *35*, 947–981. [CrossRef]
10. Wara, S.S.M.; Prastyo, D.D.; Kuswanto, H. Value at risk estimation with hybrid-SVR-GARCH-KDE model for LQ45 portfolio optimization. *AIP Conf. Proc.* **2023**, *2540*, 080013.
11. Dunis, C.L.; Laws, J.; Sermpinis, G. Modelling commodity value at risk with higher order neural networks. *Appl. Financ. Econ.* **2010**, *20*, 585–600. [CrossRef]
12. Sermpinis, G.; Laws, J.; Dunis, C.L. Modelling commodity value at risk with Psi Sigma neural networks using open–high–low–close data. *Eur. J. Financ.* **2015**, *21*, 316–336. [CrossRef]
13. Xu, Q.; Liu, X.; Jiang, C.; Yu, K. Quantile autoregression neural network model with applications to evaluating value at risk. *Appl. Soft Comput.* **2016**, *49*, 1–12. [CrossRef]
14. Zhang, H.G.; Su, C.W.; Song, Y.; Qiu, S.; Xiao, R.; Su, F. Calculating Value-at-Risk for high-dimensional time series using a nonlinear random mapping model. *Econ. Model.* **2017**, *67*, 355–367. [CrossRef]
15. He, K.; Ji, L.; Tso, G.K.; Zhu, B.; Zou, Y. Forecasting exchange rate value at risk using deep belief network ensemble based approach. *Procedia Comput. Sci.* **2018**, *139*, 25–32. [CrossRef]
16. Banhudo, G.S.F.D. Adaptive Value-at-Risk Policy Optimization: A Deep Reinforcement Learning Approach for Minimizing the Capital Charge. PhD Thesis, ISCTE Business School, Lisbon, Portugal, 2019.
17. Yu, P.; Lee, J.S.; Kulyatin, I.; Shi, Z.; Dasgupta, S. Model-based deep reinforcement learning for dynamic portfolio optimization. *arXiv* **2019**, arXiv:1901.08740.
18. Jin, B. A Mean-VaR Based Deep Reinforcement Learning Framework for Practical Algorithmic Trading. *IEEE Access* **2023**, *11*, 28920–28933. [CrossRef]
19. Li, Z.; Tran, M.N.; Wang, C.; Gerlach, R.; Gao, J. A bayesian long short-term memory model for value at risk and expected shortfall joint forecasting. *arXiv* **2020**, arXiv:2001.08374.
20. Arian, H.; Moghimi, M.; Tabatabaei, E.; Zamani, S. Encoded Value-at-Risk: A machine learning approach for portfolio risk measurement. *Math. Comput. Simul.* **2022**, *202*, 500–525. [CrossRef]
21. Zhao, L.; Gao, Y.; Kang, D. Construction and simulation of market risk warning model based on deep learning. *Sci. Program.* **2022**, *2022*, 3863107. [CrossRef]
22. Blom, H.M.; de Lange, P.E.; Risstad, M. Estimating Value-at-Risk in the EURUSD Currency Cross from Implied Volatilities Using Machine Learning Methods and Quantile Regression. *J. Risk Financ. Manag.* **2023**, *16*, 312. [CrossRef]
23. Linsmeier, T.J.; Pearson, N.D. Value at risk. *Financ. Anal. J.* **2000**, *56*, 47–67. [CrossRef]
24. Alexander, C. *Market Risk Analysis, Value at Risk Models*; John Wiley & Sons: Hoboken, NJ, USA, 2009.
25. Jorion, P. Risk2: Measuring the risk in value at risk. *Financ. Anal. J.* **1996**, *52*, 47–56. [CrossRef]
26. Box, G.E.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
27. Yoo, J.; Maddala, G. Risk premia and price volatility in futures markets. *J. Futur. Mark.* **1991**, *11*, 165–177. [CrossRef]
28. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*; OTexts: Melbourne, Australia, 2018.
29. Slutzky, E. The summation of random causes as the source of cyclic processes. *Econom. J. Econom. Soc.* **1937**, *5*, 105–146. [CrossRef]
30. Carta, S.; Corriga, A.; Ferreira, A.; Recupero, D.R.; Saia, R. A Holistic Auto-Configurable Ensemble Machine Learning Strategy for Financial Trading. *Computation* **2019**, *7*, 67. [CrossRef]
31. Sharkey, A. On Combining Artificial Neural Nets. *Connect. Sci.* **1996**, *8*, 299–314. [CrossRef]
32. Tsymbal, A.; Pechenizkiy, M.; Cunningham, P. Diversity in search strategies for ensemble feature selection. *Inf. Fusion* **2005**, *6*, 83–98. [CrossRef]
33. van Wezel, M.; Potharst, R. Improved customer choice predictions using ensemble methods. *Eur. J. Oper. Res.* **2007**, *181*, 436–452. [CrossRef]
34. Davis, J.; Devos, L.; Reyners, S.; Schoutens, W. Gradient Boosting for Quantitative Finance. *J. Comput. Financ.* **2020**, *24*, 1–40. [CrossRef]
35. jae Kim, K. Financial time series forecasting using support vector machines. *Neurocomputing* **2003**, *55*, 307–319.
36. Sadorsky, P. A Random Forests Approach to Predicting Clean Energy Stock Prices. *J. Risk Financ. Manag.* **2021**, *14*, 48. [CrossRef]
37. Ammann, M.; Reich, C. VaR for nonlinear financial instruments-linear approximation or full Monte Carlo? *Financ. Mark. Portf. Manag.* **2001**, *15*, 363–378. [CrossRef]
38. Hendricks, D. Evaluation of value-at-risk models using historical data. *Econ. Policy Rev.* **1996**, *2*, 39–69. [CrossRef]
39. Damodaran, A. *Strategic Risk Taking: A Framework for Risk Management*, 1st ed.; Wharton School Publishing: Philadelphia, PA, USA, 2007.
40. Wiener, Z. Introduction to VaR (value-at-risk). In *Risk Management and Regulation in Banking: Proceedings of the International Conference on Risk Management and Regulation in Banking (1997)* ; Springer: Berlin/Heidelberg, Germany, 1999; pp. 47–63.
41. Dowd, K. *Beyond Value at Risk: The New Science of Risk Management*; John Wiley & Son Limited: Hoboken, NJ, USA, 1999; Volume 96.

42. Cabedo, J.D.; Moya, I. Estimating oil price 'Value at Risk' using the historical simulation approach. *Energy Econ.* **2003**, *25*, 239–253. [CrossRef]
43. Seyfi, S.M.S.; Sharifi, A.; Arian, H. Portfolio Value-at-Risk and expected-shortfall using an efficient simulation approach based on Gaussian Mixture Model. *Math. Comput. Simul.* **2021**, *190*, 1056–1079. [CrossRef]
44. Morgan, J.P. *RiskMetrics*; Technical Document, J.P., Morgan/Reuters: New York, NY, USA, 1996.
45. Alexander, C.O.; Leigh, C.T. On the covariance matrices used in value at risk models. *J. Deriv.* **1997**, *4*, 50–62. [CrossRef]
46. Boudoukh, J.; Richardson, M.; Whitelaw, R.F. Investigation of a class of volatility estimators. *J. Deriv.* **1997**, *4*, 63–71. [CrossRef]
47. Ding, J.; Meade, N. Forecasting accuracy of stochastic volatility, GARCH and EWMA models under different volatility scenarios. *Appl. Financ. Econ.* **2010**, *20*, 771–783. [CrossRef]
48. Kupiec, P.H. Techniques for verifying the accuracy of risk measurement models. *J. Deriv.* **1995**, *3*, 73–84. [CrossRef]
49. Haas, M. *New Methods in Backtesting*; Financial Engineering Research Center: Bonn, Germany, 2001.
50. Christoffersen, P.F. Evaluating interval forecasts. *Int. Econ. Rev.* **1998**, *39*, 841–862. [CrossRef]
51. Shaik, M.; Padmakumari, L. Value-at-risk (VAR) estimation and backtesting during COVID-19: Empirical analysis based on BRICS and US stock markets. *Invest. Manag. Financ. Innov.* **2022**, *19*, 51–63. [CrossRef]
52. Nieppola, O. Backtesting Value-at-Risk Models. Master Thesis, Helsinki School of Economics, Espoo, Finland, 2009.