Forecasting Commodity Market Returns Volatility: A Hybrid Ensemble Learning GARCH-LSTM based Approach

Kshitij Kakade - Graduate Researcher, BITS Pilani KK Birla Goa Campus. *Email:* kshitijkakade2705@gmail.com

Aswini Kumar Mishra- Associate professor and Head, Department of Economics, BITS Pilani KK Birla Goa Campus. *Email:* *aswini@goa.bits-pilani.ac.in; aswinimishra1@gmail.com*

Kshitish Ghate - Graduate Researcher, BITS Pilani KK Birla Goa Campus. *Email:* ghatekshitish@gmail.com

Shivang Gupta - Graduate Researcher, BITS Pilani KK Birla Goa Campus. *Email:* shivangg99@gmail.com

Corresponding Author: Aswini Kumar Mishra. ORCID ID: http://orcid.org/0000-0002-7754-0633

Forecasting Commodity Market Returns Volatility: A Hybrid Ensemble Learning GARCH-LSTM based Approach.

Abstract:

This study investigates the advantage of combining the forecasting abilities of multiple generalized autoregressive conditional heteroscedasticity (GARCH)-type models such as the standard GARCH (GARCH), exponential GARCH (eGARCH), and threshold GARCH (tGARCH) models with advanced deep learning methods to predict the volatility of five important metals (Nickel, Copper, Tin, Lead & Gold) in the Indian commodity market. This paper proposes integrating the forecasts of one to three GARCH-type models into an ensemble learning-based hybrid LSTM (long short-term memory) models to forecast commodity price volatility. We further evaluate the forecasting performance of these models with respect to standalone LSTM and GARCH-type models using the root mean squared error (RMSE), mean absolute error (MAE), and mean fundamental percentage error (MAPE). The results highlight that combining the information from the forecasts of multiple GARCH-types into a hybrid LSTM model leads to superior volatility forecasting capability. The SET-LSTM, which represents the model that combines forecasts of the GARCH, eGARCH, and tGARCH into the LSTM hybrid, has shown the best overall results for all metals, barring a few exceptions. Moreover, the equivalence of forecasting accuracy is tested using the Diebold-Mariano and Wilcoxon signed-rank tests.

Keywords: Hybrid ensemble learning model; Volatility; GARCH models; LSTM, Forecasting

JEL classification: G17, G12, Q02

1. Introduction

The commodity markets serve as a vehicle of investment for big corporations and play an indispensable role in allowing giant corporations to hedge their risks and protect their long-term value. This market is highly relevant for manufacturing countries as metals play a crucial role in industrial production and economic activity. The Indian manufacturing sector contributes about 16-17 percent to its GDP (Confederation of Indian Industry.[1]). With the growth of India's manufacturing industries, the demand for metals has increased. In addition to manufacturing utility, households, especially in developing countries, invest heavily in precious metals like Gold for hedging inflation (Ghosh et al. (2004)). For these reasons, it is one of the most liquid and actively traded commodities in the Indian commodity market. Alterations in the underlying supply and demand factor alterations are significant for frequent metal price fluctuations throughout the year. In India's context, metals are of crucial strategic and economic significance and serve as the economy's vital backbone.

Financial asset class volatility estimation has always been a popular research area due to its synonymy with uncertainty in asset prices. The risk associated with asset prices is generally measured in terms of their volatility (Markowitz, 1952). Volatility analysis is essential for managing risk in hedging and asset allocations (Jondeau and Rockinger, 2003). Predicting asset price volatility has been a significant and challenging task for

[1] "CII." https://www.cii.in/. Accessed 20 Mar. 2021.

financial analysts. Given the importance of India's commodity market, it is of substantial relevance to understanding the time-varying volatility dynamics to gain a strong perception of the market's risk.

Numerous studies have employed financial and statistical time series models to model the volatility of financial markets. Ederington and Lee (1993) have used standard deviation on return to measure the stock market's price risk. The traditional regression models assume homoscedasticity in error terms, usually not valid for financial time series data (Hamilton,1994). Thus, Engle (1982) proposed the autoregressive conditional heteroscedasticity (ARCH) model. Bollerslev (1986) then presented a generalized ARCH model (GARCH) to solve the ARCH model's issues related to parameter estimation. Watkins and McAleer (2006) use GARCH (1,1)-ARMA (1) on five metals using daily data to model the time-varying volatility. The study also examines the existing literature on empirical regularities arising from nonlinear volatility-based estimation. It concludes that speculation in commodities is one of the critical reasons for the highly volatile nature of such asset classes. McMillan and Speight (2001) and Dooley and Lenihan (2005) analyze the non-ferrous metals' price and volatility using various financial time series models.

Nonlinear GARCH models such as threshold GARCH (tGARCH) and exponential GARCH (eGARCH) are often used to explain the asymmetry in volatility changes due to positive and negative shocks. A study on corn returns volatility shows that eGARCH models gave superior forecasting out sample results than the GARCH model (Musunuru, 2014). Lim and Sek (2013) use the symmetric and asymmetric GARCH models to study the Malaysian stock market's volatility. The mean squared error (MSE), root means squared error (RMSE), and mean absolute percentage error (MAPE) were used to estimate the models' GARCH predictability across three different time frames. Zhu et al. (2017) study the impact of leverage and after-hour information on volatility forecasting for Chinese non-ferrous metals. This study uses an extension of HAR-GARCH models, which have better prediction ability attributed to their superior performance than the benchmark GARCH model.

Artificial neural networks (ANN) are nonlinear functions. A neural network's significant advantage is identifying patterns in data without any underlying assumptions. Hence, they have been considered an essential algorithm to model time series with nonlinear inputs. Several studies have used either standalone neural network models or hybrid GARCH-type neural network models to obtain improvements while modeling volatility. Donaldson and Kamstra (1996) used a GARCH-ANN to forecast international stock return volatility. The comparison between the in-sample and out-sample data suggests that the hybrid ANN model better captures the volatility features overlooked by standard GARCH & eGARCH models. Ormoneit and Neuneier (1996) used a multi-layer density estimating network to predict the DAX German index. The study shows that the multi-layer density network performs better than the multi-layer perceptron. Donaldson and Kamstra (1997) use Neural Networks-GARCH to investigate the volatility trend in stock returns.

Meissner and Kawano (2001) use a similar approach to model the volatility of Spanish tech stocks. There have also been comparative studies that show that ANN models have performed better than the ARCH and GARCH models (Hamid and Iqbal, 2004; Dhamija and Bhalla, 2010). Kristjanpoller et al. (2014) conducted a volatility

forecasting study using hybrid neural networks and GARCH models in three Latin American markets. They show that the hybrid model performs better than the simple GARCH model by reducing the mean absolute percentage error (MAPE). Kristjanpoller and Minutolo (2015) conduct a similar study to show the superior performance of neural network hybrid models for gold prices. Another study (Lu et al.,2016) on volatility forecasting uses an artificial neural network and GARCH hybrid model in the Chinese energy markets. The results indicate the eGARCH-ANN neural network's better performance in forecasting log-returns of the Chinese energy market. The performance is compared based on RMSE error on the out-sample data. The results also show significant leverage effects in the Chinese energy market.

There exist numerous applications of extensions of the ANN in improving the forecasting performance of models. One such class of ANN, namely, recurrent neural networks (RNN), has feedback connections to identify the data patterns. RNN models can also remember things learned from prior inputs while generating outputs. The long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), a class of RNN model, is a well-known neural network model that performs exceptionally well while modeling and forecasting time-series data (Nelson et al. 2017; Wu et al., 2019). Kim and Won (2018) utilize hybrid GARCH-type LSTM models to predict the volatility of the KOSPI 200 index. They show that combining forecasts from multiple GARCH-type models into the hybrid neural network can significantly improve forecast performance. Ensemble methods are one of the widely used techniques to improve prediction performance. It has thus found several applications in the financial world. Lahmiri and Boukadoum (2015) use it to forecast the intraday volatility of the S&P 500. Hu et al. (2020) synthesize the GARCH and the ANN models to generate better volatility forecasts.

Recent advancements in deep learning techniques can significantly enhance traditional financial time series models' forecasting accuracy, as shown by recent studies. The literature is scant in terms of studies that explore the advantage of combining the forecasting abilities of multiple well-known GARCH models with advanced deep learning models to predict the one-step-ahead volatility of metal prices. This study also conducts the Diebold-Mariano (D.M.) and Wilcoxon signed-rank (W.S.) test for equal forecast accuracy to verify whether the model forecasts are significantly different from each other and that the performance comparisons are therefore meaningful. This is one of the early studies to propose forecasting models where information from multiple GARCH-type models is combined into an ensemble learning-based hybrid-LSTM neural network model and apply it to model the volatility of the metal commodity markets. We extend the literature in using nonlinear deep learning techniques to enhance the performance of traditional financial time series models in modeling asset price volatility. We find a significant improvement in the forecasting accuracy of the proposed hybrid neural network models compared to the standalone GARCH and LSTM models.

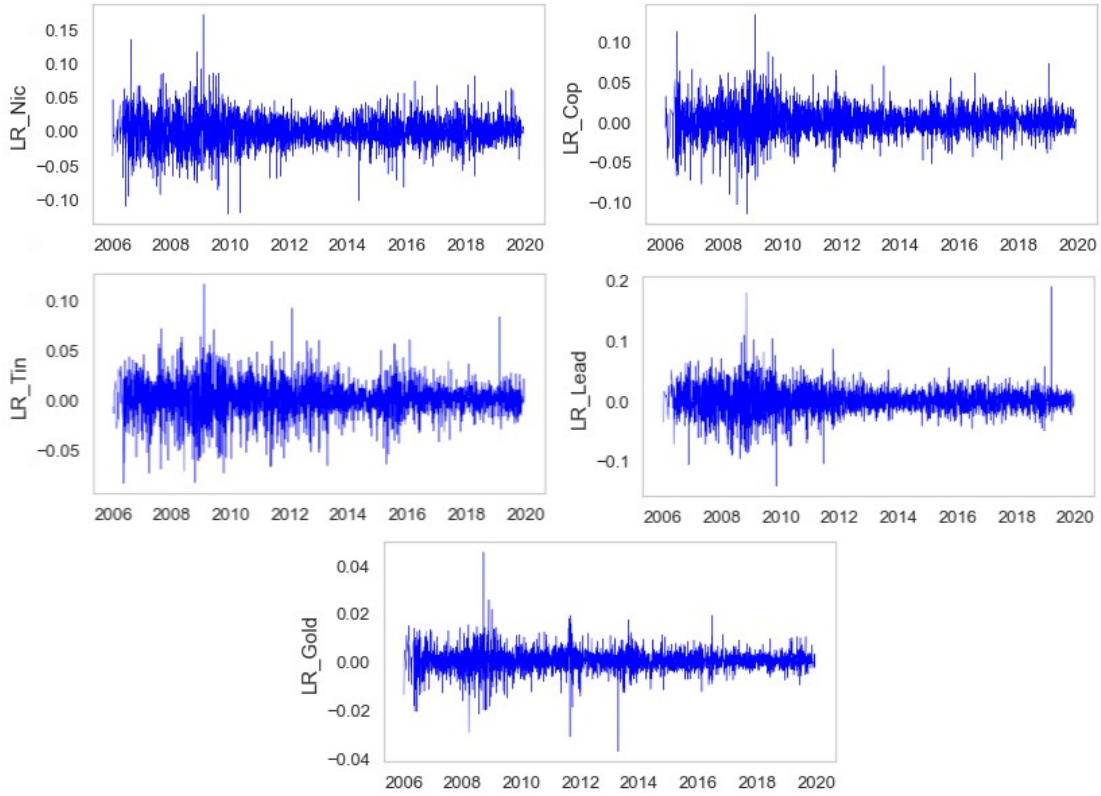
The rest of this paper proceeds as follows. Section 2 outlines the data and methodology used in this study. Section 3 describes the experimental procedures, and Section 4 discusses the empirical results. Section 5 concludes the paper.

2.Methodology:

2.1. Data:

This study considers the volatility of five metals, namely Nickel, Copper, Tin, Lead, and Gold. We utilized daily time-series data from 3/30/2006 to 3/20/2020, consisting of 3575 data points for each metal. The spot market closing prices for all the metals used in this study are collected from the MCX website.[2]. A day-wise data matching exercise was carried out for the four metals to bring the data to a model-building framework and ensure consistency across the metals.

Figure 1: Daily log returns distribution



The logarithmic return time series plots in Figure 1 show that the log-returns oscillate based on the average value of zero and are stationary.

Table 1 below describes the summary statistics of the log return series of the five metals. The Tin return series has the highest standard deviation followed by Lead, whereas Tin has the least among the four metals. As Figure 2 demonstrates, all five-metal series tend to reveal a high kurtosis, with Gold having the highest value.

Table 1: Summary statistics for log-returns of the five metals.

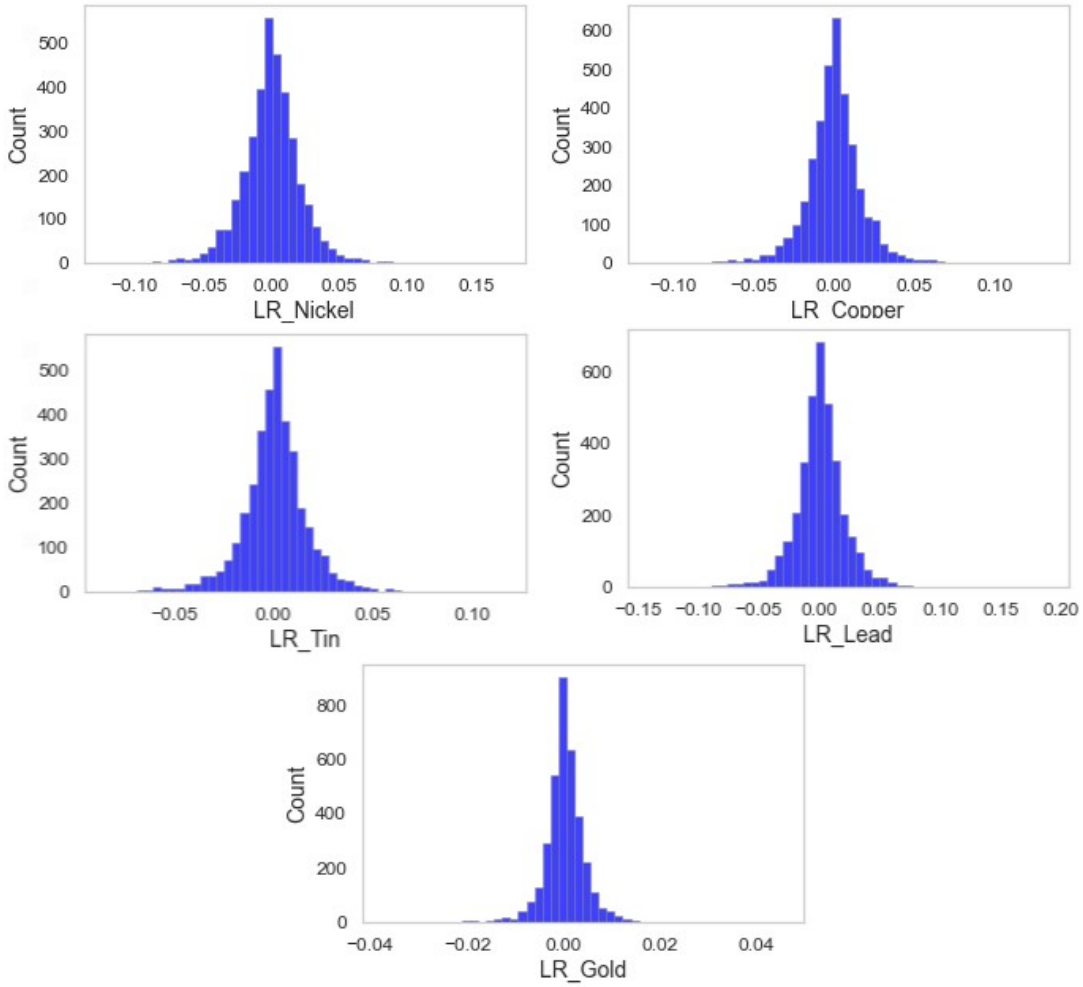
Stats	Copper	Tin	Nickel	Lead	Gold
Mean	0.00027	0.0000067	0.000047	0.00027	0.00019

[2] "MCX." https://www.mcxindia.com/. Accessed 20 Mar. 2021.

Count	3575	3575	3575	3575	3575
Variance	0.00028	0.00045	0.00031	0.00044	0.00002
Std. Dev.	0.026	0.021	0.017	0.021	0.004
Skewness	-0.178	0.036	-0.024	0.133	-0.176
Kurtosis	6.553	6.954	7.483	9.264	12.565
Jarque-Bera	0.0	0.0	0.0	0.0	0.0
ADF	-59.482***	-55.047***	-53.441***	-53.356***	-58.736***

Note: ***, **, and * denote rejection of the null hypothesis at 1%, 5%, and 10% significance level.

Figure 2: Daily log returns a histogram



The Jarque-Bera test is used to test the normality of the underlying time series. All the five metal return series have excess kurtosis. Thus, the p values mentioned in Table 1 for the JB-test are consistent, indicating a non-normal distribution for log returns. The critical value of the t-statistic for the ADF test at 10% significance value is -2.56. Thus, we reject the null hypothesis of a unit root's presence for the log return series.

Further tests are done to check the presence of heteroscedasticity and muti-collinearity in the return series. We have used the ARMA model to get the best fit mean function

of the return series. The ARMA model's lag values are then selected using the Partial Autocorrelation function at a 95% confidence interval and having the minimum Bayesian information criterion (BIC) values and Hannah–Quinn information criterion (HQIC) values. Breusch-Godfrey Serial Correlation L.M. test indicated no auto-correlation in the underlying data. The presence of heteroscedasticity in the return series was tested using the ARCH-LM test. The test showed that the error values for the underlying data were correlated with the explanatory variables. Hence, GARCH models were run after the initial tests. For a model to be optimal on underlying data, the residuals obtained from running the model should be free from ARCH effects. We have used the Box-Pierce test to verify the multicollinearity present in the residual data. The null hypothesis for a Box-Pierce Q-statistic is the presence of serial correlation in the series. The null hypothesis is rejected for the models indicating the absence of multicollinearity in the residual series. Further, The L.M. test on the residual series shows no ARCH effects implying that the residual data is homoscedastic.

2.2 Model setup:

2.2.1 Return estimation:

The closing prices play a crucial role in determining the volatility in the metals market. This paper uses the log-returns of metals to account for the heterogeneity and non-stationarity in the data. The log returns for date t are computed as follows -

$$return_t = lo(p_t) - lo(p_{t-1}) \qquad (1)$$

The return values with equal prices on the previous days are filled with infinitesimally small values to prevent null values from being used in the prediction models, ensuring that the results are not spurious.

2.2.2 Volatility estimation:

Volatility serves as a critical metric for various risk-based assessments. It acts as a statistical measure used to measure the dispersion of a given asset class around its mean or average value. The higher the asset's volatility, the higher the perceived risk in the asset and, the higher are the returns expected by investors from that asset class. This paper compares the predicted and realized volatility of the log-returns of metal prices while estimating the prediction models' correctness and fit. The Realized volatility (RV_t), which is a measure of the change in asset price during a given period, is calculated for a given time t as follows,

$$RV_t = \frac{1}{T}\sum_{j=t}^{j=t+T-1}\left(R_j - \bar{R}\right)^2 \quad (2)$$

Where T is the number of days after time t, R_j is the log return of the given metal at time t, and $\bar{R}$ is the mean log return value over the period T. We have chosen a 22-day rolling window size for calculating the volatility following Hu et al., (2020). The R.V. for all the four metals computed using a 22-day rolling window size is shown in Figure 8 (see in Appendix)

2.2.3 Data standardization

Data Standardization is essential for enhancing the performance of neural networks. Running the neural networks on non-standardized data can lead to the neural network learning incorrect parameters resulting in inaccurate predictions. The standardization is done by subtracting each time series from the minimum and dividing it with the difference between the maximum and the minimum value. The standardization is done by

$$x_{i,t} = \frac{x_{i,t} - x_{i,min}}{x_{i,max} - x_{i,min}} \qquad (3)$$

Where x_{min} and x_{max} are the minimum and maximum values for the particular time series while x_i is the volatility values for the particular day. Finally, after the standardization of values, we get the values of x between 0 and 1, which is then fed to the neural network.

2.2.4 Measures of prediction errors

The root means squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) are the three-error metrics that have been used to evaluate the out-sample forecast of the various prediction's models.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(PV_t - RV_t)^2} \qquad (4)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|PV_t - RV_t| \qquad (5)$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}|1 - PV_t/RV_t| \qquad (6)$$

R.V. is the realized volatility for time t, while P.V. is the predicted values for the particular time t, and N is the number of predictions.

The RMSE is a key metric defined as the average of the squared error. It measures the deviation of the predicted values around the actual values. MSE is the most used algorithm in ML applications; however, a significant drawback in the MSE approach is that it is subjected to biases related to the sample size of the data (Walthter et al., 2005). Thus, the RMSE is a better metric for estimating the error terms in such cases. The MAPE is the most commonly used measure for evaluating forecast accuracy. However, the MAPE is a poor accuracy indicator in some instances of Machine Learning applications.

Several past works have utilized these metrics to measure the out-sample efficiency of the training models. Fuertes et al. (2009) have applied MSE, MAE, and MAPE while analyzing the volatility in the stock markets. Hu et al. (2020) also use the measures mentioned above to evaluate the LSTM and BLSTM models.

2.2.5 Test for comparing equivalence of forecast accuracy

The Diebold-Mariano (D.M.) test and Wilcoxon signed-rank (W.S.) test are employed to check the equivalence of the forecasting accuracy of the time series models used in the study. The tests are used to show statistically significant differences in the out-of-sample forecast accuracy.

2.2.5.1 Diebold-Mariano (DM) test

The D.M. test evaluates each forecast's quality by a predefined loss function g of the forecast error (Diebold and Mariano, 1995). The null hypothesis of equal predictive accuracy is defined as $E(d_t) = 0$, where $d_t \equiv g(u_{1,t}) - g(u_{2,t})$

The D.M. statistics are obtained as follows:

$$DM = \frac{\overline{d}}{\sqrt{2\pi \hat{f}_d(0)/T}} \qquad (7)$$

where, $\overline{d} = \frac{1}{T}\sum_{t=1}^{T}(g(u_{1,t}) - g(u_{2,t}))$ and $\tilde{f}_d(0)$ is a consistent estimate of $f_d(0)$

2.2.5.2 Wilcoxon signed-rank (W.S.) test

The Wilcoxon-Signed (W.S.) Rank Test tests whether the difference of forecasting accuracy based on zero-median loss differential is statistically significant. The null hypothesis is given by median (dt) = 0 (Pratt, 2012). If the out-of-sample loss distribution is symmetric, both the tests should give consistent results. W.S. statistic is given as follows:

$$WS = \sum_{t=1}^{T} I_+(d_t)\text{rank}(|d_t|), \qquad (8)$$

$$\text{where } I_+(d_t) = \begin{cases} 1, d_t > 0 \\ 0, \text{otherwise} \end{cases}$$

2.3 Financial Time-series models:

2.3.1 ARCH (p) model

This model establishes the variance as a regression dependent on the squared residual values of previous periods (Engle, 1982). The following ARCH equations describe the conditional mean, conditional error distribution, and the series' conditional variance. The mean equation of the model is:

$$Y_t = \alpha + \beta' X_t + u_t, \frac{u_t}{\Omega_t} \sim N(0, h_t) \qquad (9)$$

X_t is a *k x 1* vector of explanatory variables, and β is a *k x 1* vector of coefficients. The error term u_t follows a normal distribution, given that all information at the time (t − 1) is available. The variance of the ARCH (p) equation is established as

$$h_t = \gamma_0 + \sum_{j=1}^{p} \gamma_j {u_t}^2_{-j} \qquad (10)$$

2.3.2 GARCH (p, q) model

A significant disadvantage of the ARCH(p) model is that there is no standard method for determining the order p of the model. Furthermore, for a more significant value of p, a greater number of parameters are required to identify the model. Bollerslev (1986) proposed a generalized model of ARCH (GARCH) to solve the above problems (now referred to as standard GARCH or GARCH).

The mean equation of the GARCH model is quite similar to that of the ARCH model. The GARCH equation for conditional variance is modeled as

$$GARCH(p,q): h_t = \gamma_0 + \sum_{i=1}^{p} \delta_i h_{t-i} + \sum_{j=1}^{q} \gamma_j {u_t}^2_{-j} \quad (11)$$

Where h_t is the conditional variance predicted using its own lagged values and the squared residuals of the errors.

2.3.3 eGARCH (p, q) model

In addition to modeling both positive and negative shocks, the eGARCH models determine the dominating shock and its impact on the time series. The variance equation for the eGARCH is modeled as

$$eGARCH\ (p,q): \log\left(h_t\right) = \alpha + \sum_{j=1}^{q} \gamma_j \left| \left(\frac{u_{t-j}}{(h_{t-j})^{1/2}} \right) \right| + \sum_{j=1}^{w} \phi_j \left(\frac{u_{t-j}}{(h_{t-j})^{1/2}} \right) + \sum_{i=1}^{p} \delta_i h_{t-i} \qquad (12)$$

The eGARCH model helps us assess the shock's impact by measuring the magnitude and sign of the shocks on the conditional volatility. A positive value for γ_j and ϕ_j means that positive and negative news have a similar impact. A positive value for γ_j and a negative value for ϕ_j indicates the negative shock has a more significant influence on the time series and vice versa.

2.3.4 tGARCH (p, q) model

The major drawback with GARCH models is that they cannot model asymmetric responses to positive and negative shocks. This primarily arises because the conditional variance present in the GARCH equation is dependent on the magnitude of the residuals and not on the sign. The tGARCH variant, however, can model such asymmetries in the news. The following equation describes the variance equation of a tGARCH process.

$$tGARCH\ (p,q): h_t = \alpha + \sum_{i=1}^{p} \delta_i h_{t-i} + \sum_{j=1}^{q} \gamma_j {u_t}^2_{-j} + \sum_{j=1}^{q} \phi_j {u_t}^2_{-j} D_{t-j} \quad (13)$$

tGARCH models employ an additional dummy variable that allows the model to react to positive and negative shocks. The sign of the residual terms is integrated into the model, which allows for asymmetric modeling. D_{t-j} takes the value of 1 if $u_{t-j} < 0$. In other cases, D_{t-j} is equal to 0.

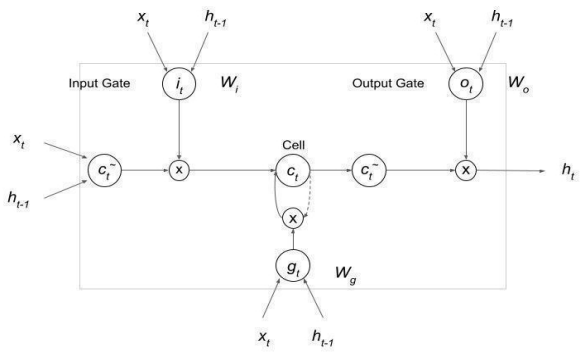
2.3.5 LSTM model

Recurrent neural networks (RNN) can forecast sequential data through internal loops derived from input sequences. RNNs produce results based on past observations by maintaining a cell state which is updated every step. A significant drawback of RNNs is that they cannot model long-run dependencies due to their inability to factor in error signals from older observations while training. The slopes obtained while training is too small (or large), leading to vanishing (or exploding gradients). The LSTM (Hochreiter and Schmidhuber, 1997) is a class of RNN that can avoid the problems mentioned above by using memory cells or states that can remember information in the long run or conveniently forget past data not necessary. LSTM is a repetitive network construction network with a gradient-based learning algorithm.

$$g_t = \sigma(U_g x_t + W_g h_{t-1} + b_f) \quad (14)$$
$$i_t = \sigma(U_i x_t + W_i h_{t-1} + b_i) \quad (15)$$
$$\tilde{c_t} = \tanh(U_c x_t + W_c h_{t-1} + b_c) \quad (16)$$
$$c_t = g_t * c_{t-1} + i_t * \tilde{c}_t \quad (17)$$
$$o_t = \sigma(U_o x_t + W_o h_{t-1} + b_o) \quad (18)$$
$$h_t = o_t * \tanh(c_t) \quad (19)$$

As Figure 3 illustrates, the LSTM consists of a memory cell (c_t) and three gates: an input gate (i_t), a forget gate (g_t), and an output gate (o_t). Equations (14) to (19) depict the calculations of the cell state (c_t), the hidden state (h_t), and the three gates. The equations are initialized by $c_0 = 0$ and $h_0 = 0$. The input is represented by h_t and the hidden state by h_t, at a given time t. The value $\tilde{c_t}$ (input modulate gate) determines the amount of new information received in the cell state for every time step. U and W are weight matrices in these equations, b is a bias term, σ (·) is a sigmoid function, tanh (.) is the hyperbolic tangent function, and the symbol * denotes element-wise multiplication.

Figure 3: The structure of a standard LSTM



2.3.6 LSTM and xGARCH hybrids:

We utilize the predictions from the GARCH, eGARCH, and tGARCH models to create our hybrid models. Firstly, we combine each model individually with the LSTM to make our first hybrid class of models: the S-LSTM, E-LSTM, and T-LSTM models. Next, we combine two of the three GARCH models with the LSTM to create our second-class models: SE-LSTM, ST-LSTM, and ET-LSTM. Finally, we use all three GARCH types to develop the SET-LSTM hybrid. The unique contribution of this study in terms of methodology is our proposed multi-component hybrid-LSTM systems that combine information from forecasts of multiple GARCH-type models using a novel ensemble learning-based approach to model and predict the volatility of the given series.

The basis of our ensemble learning-based system is as follows. For the single GARCH hybrid models, we train two networks. The first is a simple LSTM model that takes the actual volatility for the past N days rolling window as the only input. Three different window sizes of lengths 5, 11, and 22 days are used to account for historical data ranging from one week, two weeks, and one month, respectively. The other network is a single-input LSTM that takes the forecasts of the GARCH type models as its input in rolling window style. We develop the double and triple GARCH input hybrid models similarly. The double GARCH hybrid model is developed by initially training three networks; the first is the plain LSTM network. The other two are single-input LSTM networks trained on one of the series of associated GARCH-type forecasts. Similarly, the triple GARCH input hybrid has four single-input LSTM networks trained in parallel. One network works with the R.V., while the rest are trained on the forecasts obtained from the GARCH models.

The final step involves combining the results of the individual neural networks trained on the R.V. and the forecasts of the GARCH-type models. There are several methods to combine predictions, including majority vote, Bayesian bagging, and boosting. Lahmiri and Boukadoum (2015) employ a backpropagation neural network to learn and forecast the individual neural networks' inputs. This study uses a simple artificial neural network (ANN) that learns to merge the predictions of the individual LSTM to forecast the values closest to the actual volatility values. We thus arrive at the multi-component neural network ensemble system. Table 2 shows the abbreviations for each volatility model used in the study for all five metals. See figures 5, 6, and 7 in the Appendix for a diagrammatic representation of the hybrid model.

Table 2: Abbreviations used to represent models

Model type	**Model Abbreviation**	**Model Name**
GARCH models	GARCH	Standard GARCH
	eGARCH	Exponential GARCH
	tGARCH	Threshold GARCH
Hybrid Deep	LSTM	Long Short-Term Memory (RNN)
	S-LSTM	GARCH + LSTM (hybrid)
	E-LSTM	eGARCH + LSTM (hybrid)

Learning Models	T-LSTM	tGARCH + LSTM (hybrid)
	SE-LSTM	GARCH + eGARCH+LSTM (hybrid)
	ST-LSTM	GARCH +tGARCH+ LSTM (hybrid)
	ET-LSTM	eGARCH + tGARCH +LSTM (hybrid)
	SET-LSTM	GARCH +eGARCH+tGARCH+LSTM (hybrid)

3. Experiment:

The log return series of metal prices consists of 3575 data points from 3/30/2006 to 3/20/2020. We first generate the volatility forecasts for each metal using the individual time series forecasting models: GARCH, eGARCH, and tGARCH. Following Kristjanpoller. et al. (2014), we generate the GARCH forecasts using a moving rolling window of 252 days. We fit the GARCH model every 252 days and obtain one-step-ahead estimates for the same. The start of the forecasted volatility series corresponds to the 253rd day and has 3323 data points.

The models' performances are compared by measuring the closeness of forecasted volatility to the R.V. by calculating the RMSE, MAE, and MAPE metrics.

The LSTM model's architecture is determined after continuous experimentation (see Fig. 5, 6, and 7 in the Appendix) and fine-tuned for the inputs. We design the plain LSTM networks with a single hidden LSTM layer with 20 nodes. The dropout is set at 0.2 to prevent overfitting, and an Adam optimizer (Kingma, 2014) is used to train the network using MSE as the loss function. Each time series is split into train and test, consisting of 70% and 30% of the data, respectively. The one day ahead R.V. is set as the target variable. We train the model using a total of 100 epochs along with an early stopping mechanism with the patience set at 5 to avoid overfitting on in-sample data. Each hybrid model was trained twenty-five times independently, and the average of the forecast over all the iterations was used to compute the accuracy. We aim to produce a more reliable and consistent prediction accuracy by averaging all the models.

Once the plain LSTM models are run, we measure the performance of the hybrid ensemble system models. We first consider the single GARCH-type class of hybrid ensemble models. As mentioned in the previous section, we use the rolling window GARCH forecasts and the R.V. as the two main inputs to the single hybrid. The architecture of each LSTM network is similar to the plain LSTMs. The feedforward network structure used to combine the forecasts of the individual LSTMs is experimentally determined to have two fully connected layers having 128 and 64 nodes each. The single GARCH type hybrids thus obtained are the S-LSTM, E-LSTM, and T-LSTM models.

We subsequently design the two and three GARCH hybrid ensemble models described in the methodology, namely the SE-LSTM, ET-LSTM, ST-LSTM, and SET-LSTM models. The structure of these models follows from the initially developed single

GARCH-type models. The proposed models are run with each LSTM and hybrid model using inputs taken in three separate rolling windows of sizes 5, 11, and 22 days.

4. Empirical results:

We run the simple GARCH, plain LSTM, and hybrid GARCH ensemble models and generate the errors in prediction on the test data using three loss functions: RMSE, MAE, and MAPE, as shown in the Tables. We obtain the model predictions' errors for the one-day ahead forecasts of log price volatility over the out-of-sample period. The following section details the comparisons between the performances of the models across all metals. The LSTM outperforms the single GARCH-type models for all metals for every error metric across all window sizes. See Appendix (see Table 18 in the Appendix) for the error results of the GARCH models. This study investigates how combining these GARCH-type forecasts into hybrid LSTM models can improve forecasting accuracy. Therefore, the LSTM model is usedus as the baseline for evaluating all proposed hybrid models for the remainder of the analysis.

Table 3: Percentage improvement in Loss function values of Nickel for 5,11 & 22-day window sizes for LSTM.

METAL	MODEL	WIN_SZ	RMSE	MAE	MAPE
Nickel	S-LSTM	5	12.18%	19.82%	25.04%
		11	13.82%	29.31%	39.56%
		22	13.63%	29.82%	41.52%
	E-LSTM	5	11.59%	21.89%	23.84%
		11	10.92%	23.91%	27.29%
		22	13.45%	29.02%	38.02%
	T-LSTM	5	12.53%	19.94%	18.86%
		11	13.62%	28.40%	28.93%
		22	13.85%	30.33%	39.00%
	SE-LSTM	5	16.29%	25.46%	35.30%
		11	15.39%	29.72%	35.54%
		22	17.32%	33.58%	45.31%
	ST-LSTM	5	15.60%	22.92%	39.32%
		11	17.18%	32.53%	46.27%
		22	15.30%	30.42%	42.03%
	ET-LSTM	5	14.17%	21.55%	28.77%
		11	19.25%	35.94%	46.82%
		22	16.05%	32.23%	46.71%
	SET-LSTM	5	13.98%	19.02%	47.08%

		11	17.46%	32.06%	53.14%
		22	16.52%	30.78%	44.40%

Table 3 shows the results for the models run on the price volatility of Nickel. Results show that S-LSTM, E-LSTM & T-LSTM outperform the LSTM in all the loss functions across all three window sizes. The S-LSTM RMSE error is 12.18%, 13.82 %, 13.63% lower than the LSTM for the window sizes of 5, 11, and 22, respectively. A similar trend is observed for the E-LSTM and T-LSTM models. Furthermore, we see that the double GARCH hybrid LSTM models generally have lower forecasting errors than the single GARCH input model. The ET-LSTM model outperforms the plain LSTM by 14.17%, 19.25%, and 16.05% across the window sizes for the RMSE. This performance also exceeds the improvement observed in the E-LSTM and T-LSTM models. For the SET-LSTM model in the case of Nickel, the forecasting power is the best among all hybrid models for the MAPE, with errors of 0,2213, 0.2504, and 0.2976 (see Table 13 in the Appendix). However, the SET-LSTM performance is comparable with the lesser input models for the RMSE and MAE. The results support the hypothesis that the GARCH predictions contain additional explanatory power about Nickel price volatility, allowing the ensemble learning-based LSTM hybrid models to incorporate this and produce relatively better forecasts than the standalone GARCH type and LSTM models.

Table 4: Percentage improvement in Loss function values of Copper for 5,11 & 22-day window sizes for LSTM.

METAL	MODEL	WIN_SZ	RMSE	MAE	MAPE
Copper	S-LSTM	5	18.71%	14.35%	17.15%
		11	20.41%	17.73%	19.55%
		22	22.59%	17.72%	29.73%
	E-LSTM	5	10.39%	4.00%	55.20%
		11	16.86%	14.29%	54.69%
		22	17.17%	13.29%	61.31%
	T-LSTM	5	18.63%	12.03%	60.45%
		11	19.01%	14.49%	65.24%
		22	23.20%	19.59%	67.21%
	SE-LSTM	5	21.25%	19.57%	27.19%
		11	25.75%	27.07%	39.56%
		22	27.91%	26.67%	69.38%
	ST-LSTM	5	20.52%	17.70%	21.63%
		11	22.10%	20.20%	34.45%
		22	26.38%	23.89%	49.83%
	ET-LSTM	5	22.76%	20.20%	33.99%

		11	25.48%	25.94%	35.84%
		22	28.55%	28.05%	59.76%
	SET-LSTM	5	22.65%	20.42%	69.19%
		11	34.35%	35.23%	57.59%
		22	29.41%	29.81%	74.41%

In the case of Copper, as shown in Table 4, all three single GARCH hybrid models: S-LSTM, E-LSTM & T-LSTM outperform the LSTM in terms of all the loss functions across all three window sizes. However, the improvement in RMSE is more significant for T-LSTM and S-LSTM models than E-LSTM. For the double GARCH hybrid models, all three GARCH models outperform the LSTM for all three window sizes. All three models also have comparable RMSE errors. The reduction in RMSE error values of the ET-LSTM model is 22.76%, 25.48%, and 28.55% from the LSTM for the RMSE values. The two input hybrid models, on average, perform better than the single input hybrid models on all three window sizes, with the exception in terms of MAPE, where the E-LSTM and T-LSTM give better results. The triple GARCH input hybrid model, i.e., SET-LSTM, has the best forecasting power with a reduction of 22.65%, 34.35%, and 29.41% in RMSE values (see Table 14 in the Appendix).

Table 5: Percentage improvement in Loss function values for Tin of 5,11 & 22-day window sizes concerning LSTM.

METAL	MODEL	WIN_SZ	RMSE	MAE	MAPE
Tin	S-LSTM	5	7.70%	17.36%	26.70%
		11	11.63%	22.26%	35.47%
		22	12.03%	30.63%	36.31%
	E-LSTM	5	10.23%	27.54%	43.24%
		11	8.37%	23.65%	31.35%
		22	3.50%	13.35%	25.92%
	T-LSTM	5	6.97%	16.03%	30.90%
		11	11.40%	23.91%	26.08%
		22	14.49%	32.50%	41.06%
	SE-LSTM	5	7.99%	8.95%	46.70%
		11	10.78%	13.28%	52.45%
		22	10.73%	16.80%	45.62%
	ST-LSTM	5	2.08%	-1.42%	45.82%
		11	0.20%	-7.05%	48.02%
		22	8.31%	15.78%	50.93%
	ET-LSTM	5	4.16%	14.74%	38.81%
		11	12.76%	26.02%	31.80%

		22	13.46%	32.34%	41.56%
	SET-LSTM	5	6.25%	17.02%	43.45%
		11	6.70%	24.60%	50.19%
		22	13.15%	25.10%	58.87%

Table 5 confirm that the single hybrids S-LSTM, E-LSTM & T-LSTM outperform the single LSTM counterpart in terms of the error functions for every window size. The two GARCH hybrid ensemble models also perform better compared to the LSTM. Their results are mainly comparable to the single hybrid models for the RMSE and MAE. We observe two exceptions in ST-LSTM for the window sizes of 5 and 11 while considering MAE, where the LSTM performs slightly better. The three GARCH input hybrid SET-LSTM shows definite improvement compared to the performance of the LSTM. In terms of RMSE and MAE, this model is comparable to the lesser GARCH input hybrid ensembles. However, based on MAPE, the SET-LSTM is seen to be the best performing amongst all model variants, across all window sizes with errors of 0.2738, 0.2459, and 0.2397 (see Table 15 in the Appendix).

Table 6: Percentage improvement in Loss function values of Lead for 5,11 & 22-day window sizes concerning LSTM.

METAL	MODEL	WIN_SZ	RMSE	MAE	MAPE
Lead	S-LSTM	5	6.85%	18.35%	15.50%
		11	9.78%	19.16%	21.91%
		22	2.86%	11.71%	1.11%
	E-LSTM	5	8.15%	23.23%	26.88%
		11	11.26%	21.97%	24.28%
		22	4.48%	13.60%	16.39%
	T-LSTM	5	9.42%	26.95%	38.63%
		11	12.54%	26.12%	32.28%
		22	5.47%	20.22%	28.87%
	SE-LSTM	5	11.77%	33.72%	33.20%
		11	13.74%	33.08%	34.36%
		22	6.59%	24.85%	25.96%
	ST-LSTM	5	11.22%	30.77%	49.12%
		11	14.48%	33.78%	44.88%
		22	9.15%	30.76%	41.21%
	ET-LSTM	5	10.72%	30.26%	67.54%
		11	14.52%	37.34%	69.06%
		22	9.63%	32.81%	53.95%

	SET-LSTM	5	14.12%	36.57%	52.86%
		11	17.41%	37.68%	57.94%
		22	11.47%	35.40%	54.33%

As seen in Table 6, the results for Lead are in line with those for the other metals. Consistent with previous results, all the single-input hybrid LSTM models outperform the LSTM across all window sizes for every chosen error measure. The S-LSTM, E-LSTM, and T-LSTM models perform better than the plain LSTM in MAE with an error reduction of 18.35%, 23.23%, and 26.95%, respectively, for the window size of 5. This performance improvement also extends to the double GARCH hybrids, where significant error reduction in MAE and MAPE is observed compared to the performance of both the plain and the single GARCH hybrid models. For example, the ET-LSTM shows a forecast accuracy improvement of 67.54%, 69.06%, and 53.95% in MAPE compared to the LSTM for the three window sizes. The triple GARCH-type hybrid SET model's performance is not significantly better than its double and single-input input models.

Table 7: Percentage improvement in Loss function values of Gold for 5,11 & 22-day window sizes concerning LSTM.

METAL	MODEL	WIN_SZ	RMSE	MAE	MAPE
Gold	**S-LSTM**	5	12.10%	3.13%	81.62%
		11	18.30%	10.92%	83.78%
		22	21.04%	18.26%	85.48%
	E-LSTM	5	4.11%	2.99%	79.19%
		11	13.51%	2.95%	82.43%
		22	23.40%	20.43%	86.34%
	T-LSTM	5	10.31%	-2.41%	80.98%
		11	25.35%	24.47%	86.64%
		22	18.65%	11.78%	84.31%
	SE-LSTM	5	13.28%	0.48%	81.70%
		11	22.06%	17.20%	85.04%
		22	27.97%	27.40%	87.70%
	ST-LSTM	5	12.80%	1.57%	81.63%
		11	21.94%	16.87%	85.18%
		22	29.42%	31.60%	89.42%
	ET-LSTM	5	15.45%	4.36%	82.58%
		11	8.64%	9.40%	80.56%
		22	26.29%	23.75%	86.81%
	SET-LSTM	5	33.81%	33.16%	89.96%
		11	22.11%	13.44%	85.16%

		22	30.14%	32.91%	89.51%

Table 7 displays the improvements in forecasting accuracy for the price volatility of Gold. The trend in performance improvements is significant and broadly consistent with the preceding metals. The single GARCH input hybrid models show vast improvements in forecasting volatility than the baseline plain LSTM, except for the T-LSTM. It is noteworthy that the single hybrids outperform the baseline by more than 80% when MAPE is used to measure the improvement. A similar gain in performance is observed for the double GARCH input hybrid ensembles. The SE-LSTM shows RMSE reductions of 13.28%, 22.06%, and 27.97% across the window sizes, which are substantially better than the S-LSTM and E-LSTM models. The results support the hypothesis that forecasting accuracy can be improved by combining the predictions of individual GARCH models.

Figure 4: Comparison of prediction errors of the hybrid models with the LSTM model for the 5-day rolling window size

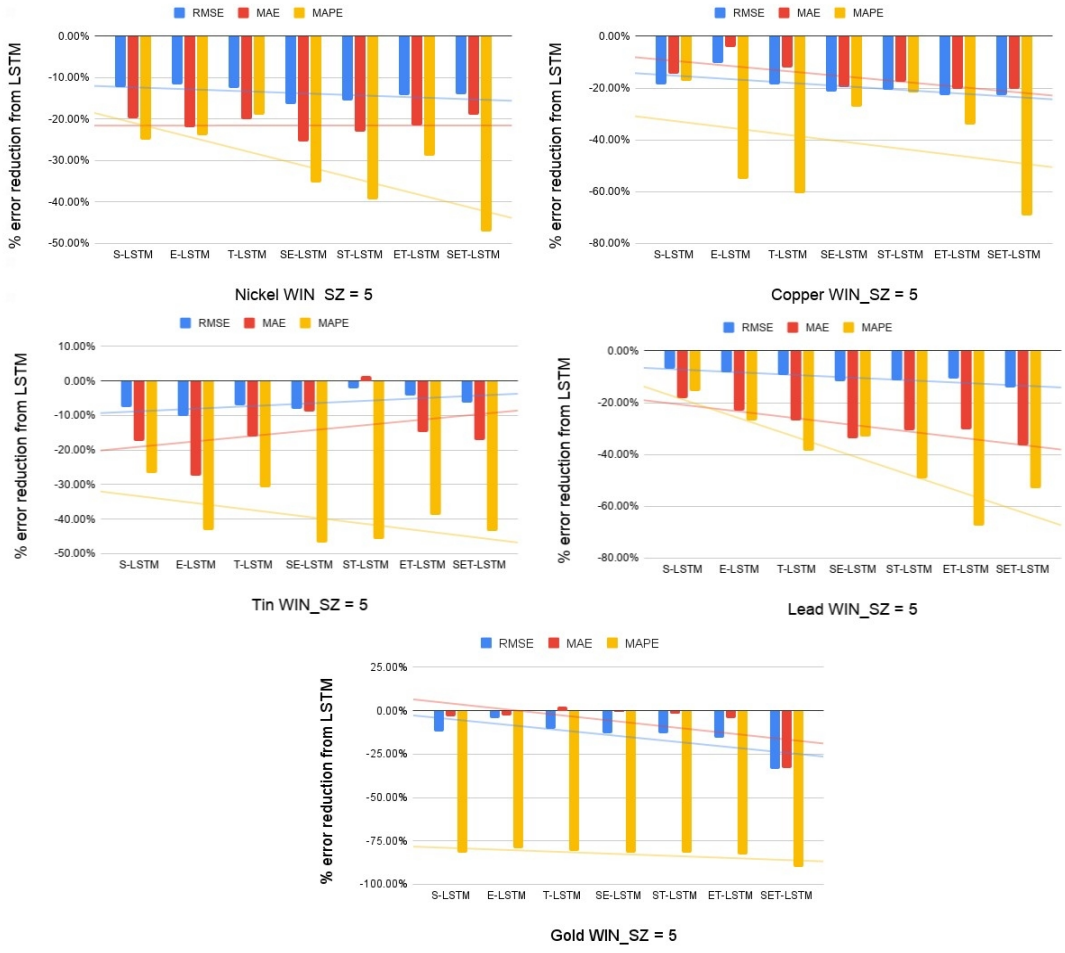


Figure 4 shows a graphical representation of the general improvement in one-step-ahead forecast performance considering the 5-day window size for all metals as we move from the plain LSTM to the single and multiple input GARCH LSTM hybrids. The trend is the most visible while considering MAPE as the error metric. This trend is consistent, even if the rolling window size increases to 11 & 22 days (see Figure 9 & Figure 10 in the Appendix). We see that there is no significant change in performance depending on the choice of window size. In line with previous studies such as

Kristjanpoller. et al. (2014), Kristjanpoller and Hernández. (2017) and Hu et al. (2020) this study verifies that incorporating information from simple statistical GARCH models into neural network models can significantly improve forecast predictions. The improvement can arise as economic characteristics captured in the forecasts of the multiple GARCH type models are different (Kim and Won, 2018) and can add to the predictive capabilities of the LSTM constructively. In some instances, the three-input SET-LSTM does not show vast predictive capability improvements due to possible overlap in information from the GARCH forecasts. In general, however, the performance of the SET-LSTM does exhibit slightly superior forecasting accuracy across all metals.

Tables 8, 9, 10, 11, and 12 show the Diebold-Mariano (D.M.) and Wilcoxon signed-rank (W.S.) test results for equal forecast accuracy. The tests' null hypothesis suggests that the forecasts are of comparable accuracy, and therefore comparisons made between them are not significant. The p values associated with the D.M. tests are reported above the diagonal, while those of the W.S. tests are below. At the 95% significance level, the null hypothesis is rejected for a majority of the cases, indicating that the out-of-sample forecast accuracy obtained from each model is significantly different from the other. There are very few exceptions, such as Nickel, where S-LSTM cannot be significant compared with SE-LSTM and ST-LSTM. Overall, the comparison of the out-of-sample performance of most models is statistically significant. This result, combined with the observation that the hybrid ensemble models show an improved forecast accuracy of price volatility, suggests that GARCH type forecasts to neural network models can significantly improve performance.

Table 8: Diebold-Mariano & Wilcoxon signed-rank test results for Nickel

Nickel	LSTM	S-LSTM	E-LSTM	T-LSTM	SE-LSTM	ST-LSTM	ET-LSTM	SET-LSTM
LSTM		0.00	0.00	0.00	0.00	0.00	0.00	0.00
S-LSTM	0.00		**0.28**	**0.50**	0.00	**0.40**	0.00	0.03
E-LSTM	0.00	0.00		0.00	0.00	0.01	0.00	**0.13**
T-LSTM	0.00	0.00	0.00		0.00	0.00	0.00	0.00
SE-LSTM	0.00	**0.35**	0.00	0.00		0.00	0.00	0.00
ST-LSTM	0.00	0.00	0.00	0.00	0.00		0.00	0.00
ET-LSTM	0.00	0.00	0.00	0.00	0.00	0.00		0.00
SET-LSTM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

Table 9: Diebold-Mariano & Wilcoxon signed-rank test results for Copper

Copper	LSTM	S-LSTM	E-LSTM	T-LSTM	SE-LSTM	ST-LSTM	ET-LSTM	SET-LSTM
LSTM		0.00	0.00	0.00	0.00	0.00	0.00	0.00
S-LSTM	0.00		0.01	0.04	0.00	0.00	0.00	0.00
E-LSTM	0.00	0.00		0.03	0.00	0.00	0.00	0.00
T-LSTM	0.00	0.00	0.00		0.00	0.00	0.00	0.00

SE-LSTM	0.00	0.00	0.00	0.00		0.00	**0.10**	0.00
ST-LSTM	0.00	0.00	0.00	0.00	0.00		0.00	0.00
ET-LSTM	0.00	0.00	0.00	0.00	0.00	0.00		0.01
SET-LSTM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

Table 10: Diebold-Mariano & Wilcoxon signed-rank test results for Tin

Tin	**LSTM**	**S-LSTM**	**E-LSTM**	**T-LSTM**	**SE-LSTM**	**ST-LSTM**	**ET-LSTM**	**SET-LSTM**
LSTM		0.00	0.00	0.00	0.00	0.00	0.00	0.00
S-LSTM	0.00		0.00	0.00	0.00	0.00	**0.10**	0.00
E-LSTM	0.00	0.00		0.00	0.00	**0.60**	0.00	0.00
T-LSTM	0.00	0.00	0.00		0.00	0.00	**0.70**	0.00
SE-LSTM	0.00	0.00	0.00	0.00		0.20	0.00	0.00
ST-LSTM	0.00	0.00	0.00	0.00	0.00		0.00	0.00
ET-LSTM	0.00	0.00	0.00	0.00	0.00	0.00		0.00
SET-LSTM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

Table 11: Diebold-Mariano & Wilcoxon signed-rank test results for Lead

Lead	**LSTM**	**S-LSTM**	**E-LSTM**	**T-LSTM**	**SE-LSTM**	**ST-LSTM**	**ET-LSTM**	**SET-LSTM**
LSTM		0.00	0.00	0.00	0.00	0.00	0.00	0.00
S-LSTM	0.00		0.00	0.00	0.00	0.00	0.00	0.00
E-LSTM	0.00	0.00		0.00	0.00	0.00	0.00	0.00
T-LSTM	0.00	0.00	0.00		0.00	0.00	0.00	0.00
SE-LSTM	0.00	0.00	0.00	0.00		0.00	0.00	0.00
ST-LSTM	0.00	0.00	0.00	0.00	0.00		0.02	0.00
ET-LSTM	0.00	0.00	0.00	0.00	0.00	0.00		0.00
SET-LSTM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

Table 12: Diebold-Mariano & Wilcoxon signed-rank test results for Gold

Gold	**LSTM**	**S-LSTM**	**E-LSTM**	**T-LSTM**	**SE-LSTM**	**ST-LSTM**	**ET-LSTM**	**SET-LSTM**
LSTM		0.00	0.00	0.00	0.00	0.00	0.00	0.00
S-LSTM	0.00		0.00	0.00	0.00	0.00	0.00	0.00
E-LSTM	0.00	0.00		0.00	0.00	0.00	0.04	0.00
T-LSTM	0.00	0.00	0.00		0.00	0.00	0.00	0.00
SE-LSTM	0.00	0.00	0.00	0.00		0.00	0.00	0.00

ST-LSTM	0.00	0.00	0.00	0.00	0.00		**0.22**	0.00
ET-LSTM	0.00	0.00	0.00	0.00	0.00	0.00		0.00
SET-LSTM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

Note: In Tables 8, 9, 10, 11, and 12, values above the diagonal are p values for the D.M. test, and values below the diagonal are p values for the W.S. test with the bold values indicating cases where the p-value exceeds 0.05, leading to rejection of the null hypothesis.

5. Conclusions

This study seeks to model and predict the volatility dynamics of the commodity markets by proposing advanced hybrid deep learning models that combine the forecasting abilities of multiple well-known GARCH-type models, namely the GARCH, eGARCH, and tGARCH. This paper builds on previous studies (Kristjanpoller and Minutolo, 2015; Kristjanpoller and Hernández, 2017; Kim and Won, 2018; Hu et al., 2020) that suggest hybrid models to model and forecast the price volatility of critical metal commodities (Nickel, Copper, Tin, Lead and Gold) in the Indian market. The LSTM is the choice of neural network architecture due to its notable ability to remember long-term dependencies in data, which gives it the capability to perform exceptionally well while forecasting time series data. This study verifies the hypothesis that adding the forecasts of the three GARCH-type models, based on distinct economic characteristics, as inputs to the proposed LSTM based model can significantly improve the one-step-ahead forecasting accuracy of volatility.

After rigorous experimentation, this study proposes the following single GARCH (S-LSTM, E-LSTM, and T-LSTM), double GARCH hybrid (SE-LSTM, ET-LSTM, and ST-LSTM), and the triple GARCH (SET-LSTM) LSTM models. The performance is compared against standalone LSTM and GARCH models on forecasting the one-step-ahead volatilities of metal prices. Three different window sizes are used for testing the LSTM and hybrid LSTM models. We find that the LSTM outperforms the GARCH models for all metals for every error metric across all window sizes. To verify whether adding GARCH-type forecasts can improve the model performance of the hybrids, we focus the remainder of our analysis on evaluating the performance for all proposed hybrid models taking LSTM model performance as the baseline. The SET-LSTM, created by incorporating the forecasts of all three GARCH-type models into the hybrid LSTM model, is the best performing ensemble learning-based hybrid across all metals, with certain exceptions. The Diebold-Mariano and Wilcoxon Signed-rank tests verify that most performance comparisons are statistically significant. This study's findings may be of particular relevance to policymakers and risk management professionals who seek to understand the risk associated with the commodity markets. Multiple days ahead forecasts of the asset price volatility can be a further extension to this study. This paper could also help motivate other studies on the applications of ensemble learning techniques and hybrid neural network models in enhancing the performance of financial time series models.

Data Availability Statement:
The data supporting this study's findings are available from the corresponding author upon reasonable request.

References

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

GHOSH, D., LEVIN, E. J., MACMILLAN, P., & WRIGHT, R. E. (2004). GOLD AS AN INFLATION HEDGE? *Studies in Economics and Finance, 22(1), 1–25*. doi:10.1108/eb043380

Dhamija, A. K., & Bhalla, V. K. (2010). Exchange rate forecasting: Comparison of various architectures of neural networks. *Neural Computing and Applications, 20*(3), 355-363. doi:10.1007/s00521-010-0385-5.

Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics, 13*(3), 253. doi:10.2307/1392185

Donaldson, R. G., & Kamstra, M. (1996). Forecast combining with neural networks. *Journal of Forecasting, 15*(1), 49-61. doi:10.1002/(sici)1099-131x(199601)15:13.0.co;2-2

Dooley, G., & Lenihan, H. (2005). An assessment of time series methods in metal price forecasting. *Resources Policy, 30*(3), 208-217. DOI: 10.1016/j.resourpol.2005.08.007

Ederington, L. H., & Lee, J. H. (1993). How Markets Process Information: News Releases and Volatility. *The Journal of Finance, 48*(4), 1161-1191. doi:10.1111/j.1540-6261. 1993.tb04750.x

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica, 50*(4), 987. doi:10.2307/1912773

Fuertes, A., Izzeldin, M., & Kalotychou, E. (2009). On forecasting daily stock volatility: The role of intraday information and market conditions. *International Journal of Forecasting, 25*(2), 259-281. DOI: 10.1016/j.ijforecast.2009.01.006

Gil-Alana, L. A., & Tripathy, T. (2014). Modeling volatility persistence and asymmetry: A Study on selected Indian non-ferrous metals markets. *Resources Policy, 41*, 31-39. DOI: 10.1016/j.resourpol.2014.02.004

Hajizadeh, E., Seifi, A., Zarandi, M. F., & Turksen, I. B. (2012). A hybrid modeling approach for forecasting the volatility of S&P 500 index return. *Expert Systems with Applications*, *39*(1), 431-436.

Hamid, S. A., & Iqbal, Z. (2004). Using neural networks for forecasting volatility of S&P 500 Index futures prices. *Journal of Business Research, 57*(10), 1116-1125. doi:10.1016/s0148-2963(03)00043-2

Hamilton, J. D. (1994). Time Series Analysis. doi:10.1515/9780691218632

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation, 9*(8), 1735-1780. doi:10.1162/neco.1997.9.8.1735

Hu, Y., Ni, J., & Wen, L. (2020). A hybrid deep learning approach by integrating LSTM-ANN networks with the GARCH model for copper price volatility prediction. *Physica A: Statistical Mechanics and Its Applications, 557*, 124907. DOI: 10.1016/j.physa.2020.124907

Jondeau, E., & Rockinger, M. M. (2003). The Allocation of Assets Under Higher Moments. *SSRN Electronic Journal*. doi:10.2139/ssrn.410743

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. DOI: 10.1016/j.eswa.2018.03.002

Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

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. DOI: 10.1016/j.eswa.2015.04.058

Kristjanpoller,W., Fadic, A., & Minutolo, M. C. (2014). Volatility forecast using hybrid Neural Network models. *Expert Systems with Applications, 41*(5), 2437-2442. DOI: 10.1016/j.eswa.2013.09.043

Lahmiri, S., & Boukadoum, M. (2014). An Ensemble System Based on Hybrid EGARCH-ANN with Different Distributional Assumptions to Predict S&P 500 Intraday Volatility. *Fluctuation and Noise Letters, 14*(01), 1550001. doi:10.1142/s0219477515500017

Lim, C. M., & Sek, S. K. (2013). Comparing the Performances of GARCH-type Models in Capturing the Stock Market Volatility in Malaysia. *Procedia Economics and Finance, 5*, 478-487. doi:10.1016/s2212-5671(13)00056-7

Lu, X., Que, D., & Cao, G. (2016). Volatility Forecast Based on the Hybrid Artificial Neural Network and GARCH-type Models. *Procedia Computer Science, 91*, 1044-1049. doi:10.1016/j.procs.2016.07.145

Markowitz, H. (1959). Portfolio selection.doi:10.2307/2975974.

Mcmillan, D. G., & Speight, A. E. (2001). Non-ferrous metals price volatility: A component analysis. *Resources Policy, 27*(3), 199-207. doi:10.1016/s0301-4207(01)00019-8

Meade, N. (1995). Neural network time series forecasting of financial markets. *International Journal of Forecasting, 11*(4), 601-602. doi:10.1016/s0169-2070(95)90005-5

Meissner, G., & Kawano, N. (2001). Capturing the volatility smile of options on high-tech stocks—A combined GARCH-neural network approach. *Journal of Economics and Finance, 25*(3), 276-292. doi:10.1007/bf02745889

Musunuru, N. (2014). Modeling Price Volatility Linkages between Corn and Wheat: A Multivariate GARCH Estimation. *International Advances in Economic Research, 20*(3), 269-280. doi:10.1007/s11294-014-9477-9

Namugaya, J., Waititu, A. G., & Diongue, A. K. (2019). Forecasting stock returns volatility on Uganda securities exchange using TSK fuzzy-GARCH and GARCH models. *Reports on Economics and Finance, 5*(1), 1-14. doi:10.12988/ref.2019.81022

Nelson, D. M., Pereira, A. C., & Oliveira, R. A. (2017). Stock market's price movement prediction with LSTM neural networks. *2017 International Joint Conference on Neural Networks (IJCNN)*. doi:10.1109/ijcnn.2017.7966019

Ni, H., & Yin, H. (2009). Exchange rate prediction using hybrid neural networks and trading indicators. *Neurocomputing, 72*(13-15), 2815-2823. DOI: 10.1016/j.neucom.2008.09.023

Ormoneit, D. & Neuneier, R. (1996, March). Experiments in predicting the German stock index DAX with density estimating neural networks. In *IEEE/IAFE 1996 Conference on Computational Intelligence for Financial Engineering (CIFEr)* (pp. 66-71). IEEE.doi:10.1109/cifer.1996.501825

Pratt, J. W. (1959). Remarks on Zeros and Ties in the Wilcoxon Signed Rank Procedures. *Journal of the American Statistical Association, 54*(287), 655-667. doi:10.1080/01621459.1959.10501

Ranco, G., Bordino, I., Bormetti, G., Caldarelli, G., Lillo, F., & Treccani, M. (2015). Coupling News Sentiment with Web Browsing Data Improves Prediction of Intra-Day Price Dynamics. *SSRN Electronic Journal*. doi:10.2139/ssrn.2699167

R, W. K., & P, E. H. (2017). Volatility of main metals forecasted by a hybrid ANN-GARCH model with regressors. Expert Systems with Applications, 84, 290-300. DOI: 10.1016/j.eswa.2017.05.024

Rosa, R., Maciel, L., Gomide, F., & Ballini, R. (2014). Evolving hybrid neural fuzzy network for realized volatility forecasting with jumps. *2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*. doi:10.1109/cifer.2014.6924112

Roh, T. H. (2007). Forecasting the volatility of the stock price index. *Expert Systems with Applications*, *33*(4), 916-92

Walther, B. A., & Moore, J. L. (2005). The concepts of bias, precision, and accuracy,

and their use in testing the performance of species richness estimators, with a literature review of estimator performance. Ecography, 28(6), 815-829. doi:10.1111/j.2005.0906-7590. 04112.x

Watkins, C., & Mcaleer, M. (2006). Pricing of non-ferrous metals futures on the London Metal Exchange. *Applied Financial Economics, 16*(12), 853-880. doi:10.1080/09603100600756514

Wei, Y., Liu, J., Lai, X., & Hu, Y. (2016). Which determinant is the most informative in forecasting crude oil market volatility: Fundamental, speculation, or uncertainty? *Energy Economics, 68*, 141-150. DOI: 10.1016/j.eneco.2017.09.016

Wu, Y., Wu, Q., & Zhu, J. (2019). Improved EEMD-based crude oil price forecasting using LSTM networks. *Physica A: Statistical Mechanics and Its Applications, 516*, 114-124. DOI: 10.1016/j.physa.2018.09.120

Zhu, X., Zhang, H., & Zhong, M. (2017). Volatility forecasting using high-frequency data: The role of after-hours information and leverage effects. *Resources Policy, 54*, 58-70. DOI: 10.1016/j.resourpol.2017.09.006

Appendix

Figure 5: Single GARCH Hybrid Model Structure

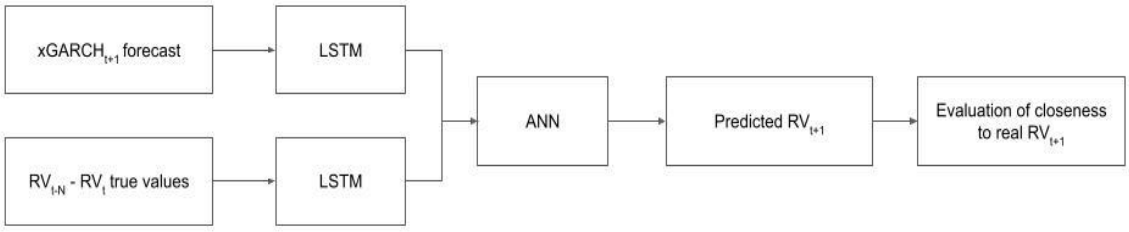


Figure 6: Double GARCH Hybrid Model Structure

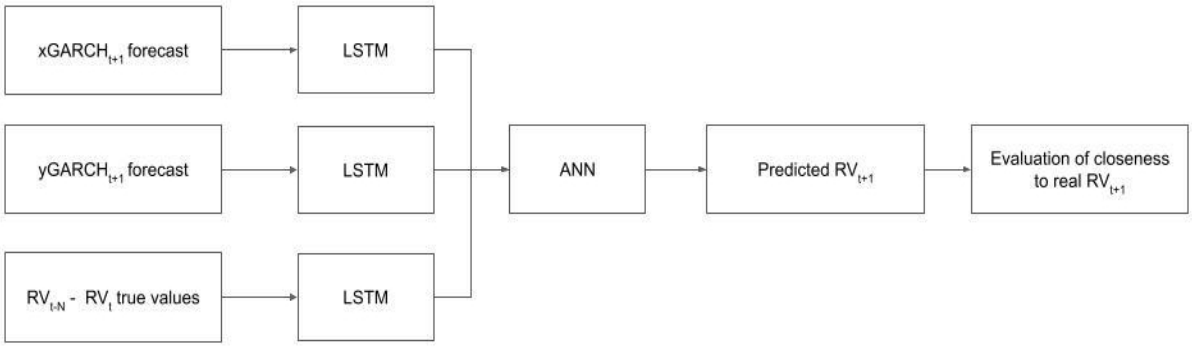


Figure 7: Triple GARCH Hybrid Model Structure

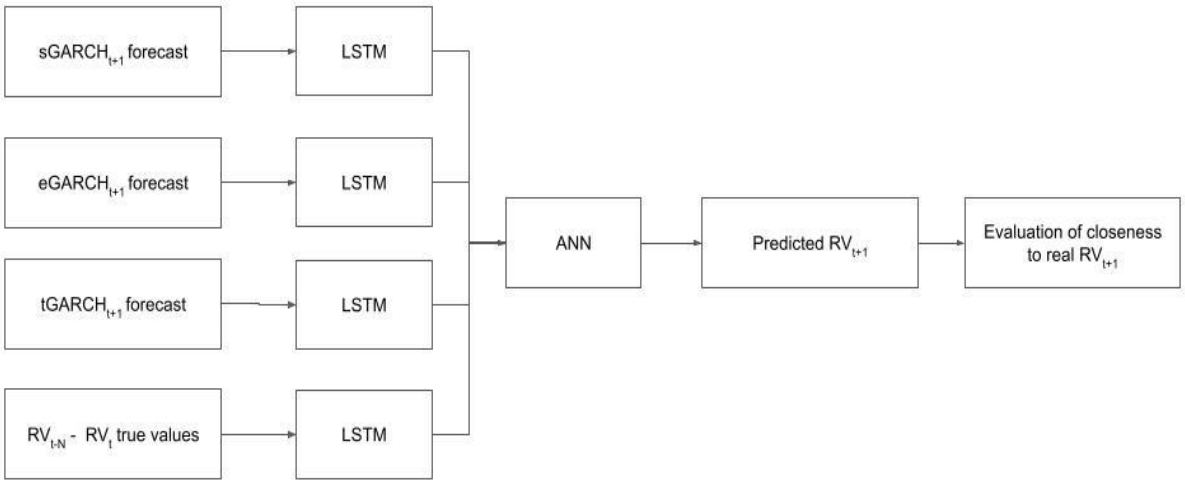


Figure 8: Realized Volatility with 22 days as the window size

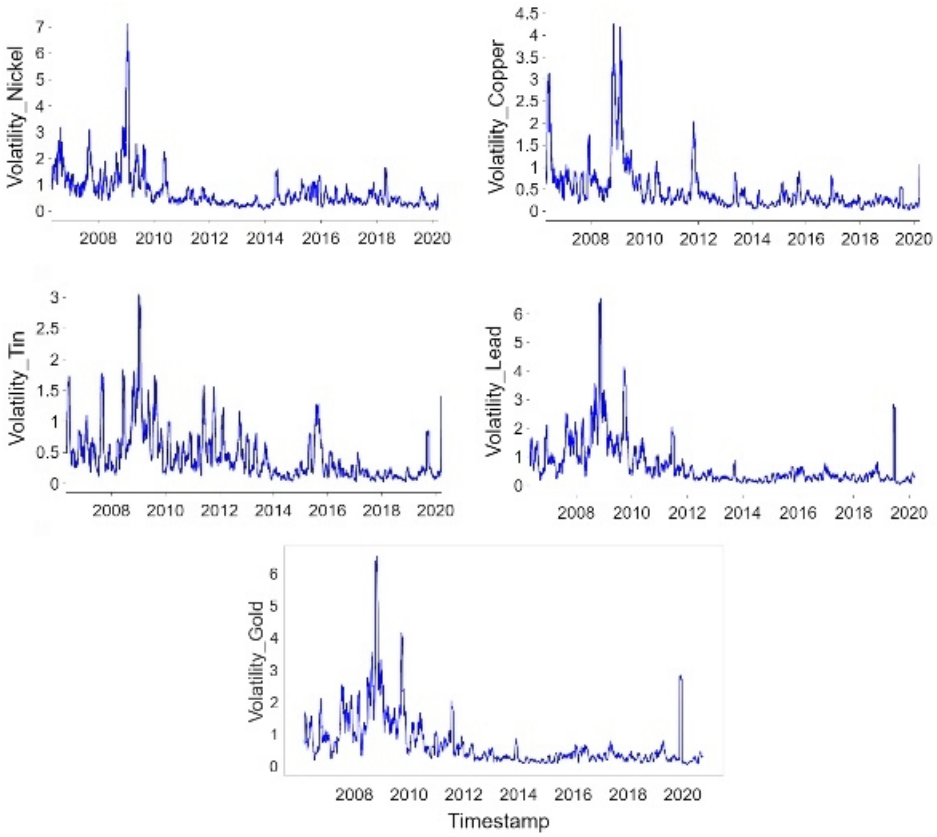


Table 13: Loss function values for Nickel for 5,11 & 22-day window sizes.

METAL	MODEL	WIN_SZ	RMSE	MAE	MAPE
Nickel	LSTM	5	0.01017	0.00699	0.41824
		11	0.01067	0.00815	0.53426
		22	0.01055	0.00805	0.53530
	S-LSTM	5	0.00893	0.00560	0.31353
		11	0.00920	0.00576	0.32292
		22	0.00911	0.00565	0.31306
	E-LSTM	5	0.00899	0.00546	0.31852
		11	0.00951	0.00620	0.38848
		22	0.00913	0.00572	0.33175
	T-LSTM	5	0.00889	0.00560	0.33936
		11	0.00922	0.00584	0.37972
		22	0.00909	0.00561	0.32654
	SE-LSTM	5	0.00851	0.00521	0.27058
		11	0.00903	0.00573	0.34440
		22	0.00872	0.00535	0.29277

	ST-LSTM	5	0.00858	0.00539	0.25377
		11	0.00884	0.00550	0.28708
		22	0.00893	0.00560	0.31031
	ET-LSTM	5	0.00873	0.00548	0.29790
		11	0.00862	0.00522	0.28410
		22	0.00885	0.00546	0.28528
	SET-LSTM	5	0.00875	0.00566	0.22133
		11	0.00881	0.00554	0.25036
		22	0.00880	0.00557	0.29763

Table 14: Loss function values for Copper for 5,11 & 22-day window sizes

METAL	MODEL	WIN_SZ	RMSE	MAE	MAPE
Copper	LSTM	5	0.01009	0.00580	1.13690
		11	0.01080	0.00655	1.35505
		22	0.01119	0.00668	1.45458
	S-LSTM	5	0.00820	0.00497	0.94191
		11	0.00860	0.00539	1.09015
		22	0.00866	0.00550	1.02220
	E-LSTM	5	0.00904	0.00557	0.50934
		11	0.00898	0.00561	0.61399
		22	0.00927	0.00579	0.56277
	T-LSTM	5	0.00821	0.00510	0.44961
		11	0.00875	0.00560	0.47107
		22	0.00860	0.00537	0.47692
	SE-LSTM	5	0.00795	0.00467	0.82779
		11	0.00802	0.00478	0.81905
		22	0.00807	0.00490	0.44540
	ST-LSTM	5	0.00802	0.00478	0.89097
		11	0.00842	0.00523	0.88821
		22	0.00824	0.00508	0.72975
	ET-LSTM	5	0.00779	0.00463	0.75049
		11	0.00805	0.00485	0.86940
		22	0.00800	0.00481	0.58533

	SET-LSTM	5	0.00781	0.00462	0.35033
		11	0.00709	0.00424	0.57469
		22	0.00790	0.00469	0.37223

Table 15: Loss function values for Tin for 5,11 & 22-day window sizes.

METAL	MODEL	WIN_SZ	RMSE	MAE	MAPE
Tin	**LSTM**	5	0.01296	0.00704	0.48422
		11	0.01310	0.00734	0.49374
		22	0.01335	0.00792	0.58282
	S-LSTM	5	0.01196	0.00582	0.35491
		11	0.01158	0.00571	0.31860
		22	0.01174	0.00549	0.37120
	E-LSTM	5	0.01163	0.00510	0.27482
		11	0.01200	0.00561	0.33893
		22	0.01288	0.00686	0.43174
	T-LSTM	5	0.01205	0.00591	0.33458
		11	0.01161	0.00559	0.36495
		22	0.01141	0.00534	0.34350
	SE-LSTM	5	0.01192	0.00641	0.25809
		11	0.01169	0.00637	0.23479
		22	0.01192	0.00659	0.31695
	ST-LSTM	5	0.01269	0.00714	0.26235
		11	0.01307	0.00786	0.25667
		22	0.01224	0.00667	0.28600
	ET-LSTM	5	0.01242	0.00600	0.29632
		11	0.01143	0.00543	0.33673
		22	0.01155	0.00536	0.34061
	SET-LSTM	5	0.01215	0.00584	0.27384
		11	0.01222	0.00554	0.24593
		22	0.01159	0.00593	0.23969

Table 16: Loss function values for Lead for 5,11 & 22-day window sizes.

METAL	MODEL	WIN_SZ	RMSE	MAE	MAPE
	LSTM	5	0.02454	0.00816	3.09340
		11	0.02561	0.00862	3.15988

Lead		22	0.02407	0.00831	3.07686
	S-LSTM	5	0.02286	0.00666	2.61394
		11	0.02310	0.00697	2.46749
		22	0.02338	0.00734	3.04269
	E-LSTM	5	0.02254	0.00626	2.26204
		11	0.02272	0.00673	2.39271
		22	0.02299	0.00718	2.57242
	T-LSTM	5	0.02223	0.00596	1.89834
		11	0.02240	0.00637	2.13976
		22	0.02275	0.00663	2.18842
	SE-LSTM	5	0.02165	0.00541	2.06650
		11	0.02209	0.00577	2.07419
		22	0.02248	0.00625	2.27813
	ST-LSTM	5	0.02179	0.00565	1.57388
		11	0.02190	0.00571	1.74170
		22	0.02186	0.00576	1.80899
	ET-LSTM	5	0.02191	0.00569	1.00421
		11	0.02189	0.00540	0.97778
		22	0.02175	0.00559	1.41697
	SET-LSTM	5	0.02107	0.00517	1.45827
		11	0.02115	0.00538	1.32899
		22	0.02131	0.00537	1.40512

Table 17: Loss function values for Gold for 5,11 & 22-day window sizes.

METAL	MODEL	WIN_SZ	RMSE	MAE	MAPE
Gold	LSTM	5	0.00849	0.00583	1.51958
		11	0.00905	0.00663	1.77984
		22	0.00888	0.00644	1.71771
	S-LSTM	5	0.00746	0.00565	0.27928
		11	0.00694	0.0052	0.24649
		22	0.0067	0.00477	0.22071
	E-LSTM	5	0.00814	0.00566	0.31624
		11	0.00734	0.00566	0.26705
		22	0.0065	0.00464	0.20751

		5	0.00761	0.00597	0.28908
		11	0.00634	0.00441	0.20309
	T-LSTM	22	0.00691	0.00515	0.23844
		5	0.00736	0.0058	0.27807
		11	0.00662	0.00483	0.22732
	SE-LSTM	22	0.00612	0.00423	0.18696
		5	0.0074	0.00574	0.27911
		11	0.00663	0.00485	0.22525
	ST-LSTM	22	0.00599	0.00399	0.16078
		5	0.00718	0.00558	0.26467
		11	0.00776	0.00528	0.29535
	ET-LSTM	22	0.00626	0.00445	0.20051
		5	0.00562	0.0039	0.15261
		11	0.00661	0.00505	0.22556
	SET-LSTM	22	0.00593	0.00391	0.15933

Table 18: Loss function values for the GARCH models.

METAL	MODEL	RMSE	MAE	MAPE
	S-GARCH	0.03689	0.02441	0.71357
	E-GARCH	0.06362	0.02681	0.60728
Nickel	**T-GARCH**	0.03548	0.02308	0.6909
	S-GARCH	0.03266	0.02556	1.97953
	E-GARCH	0.03403	0.02683	1.17968
Copper	**T-GARCH**	0.03291	0.02503	1.16799
	S-GARCH	0.04981	0.03141	0.90387
	E-GARCH	0.06758	0.03322	0.93471
Tin	**T-GARCH**	0.05078	0.03112	0.93613
	S-GARCH	0.06523	0.02805	7.48198
	E-GARCH	0.06792	0.02752	4.0003
Lead	**T-GARCH**	0.06388	0.02581	3.82876
	S-GARCH	0.09452	0.06008	2.81921
	E-GARCH	0.081	0.05726	2.65833
Gold	**T-GARCH**	0.08126	0.06024	2.74645

Figure 9: Comparison of prediction errors of the hybrid models with the LSTM model for the 11-day rolling window size

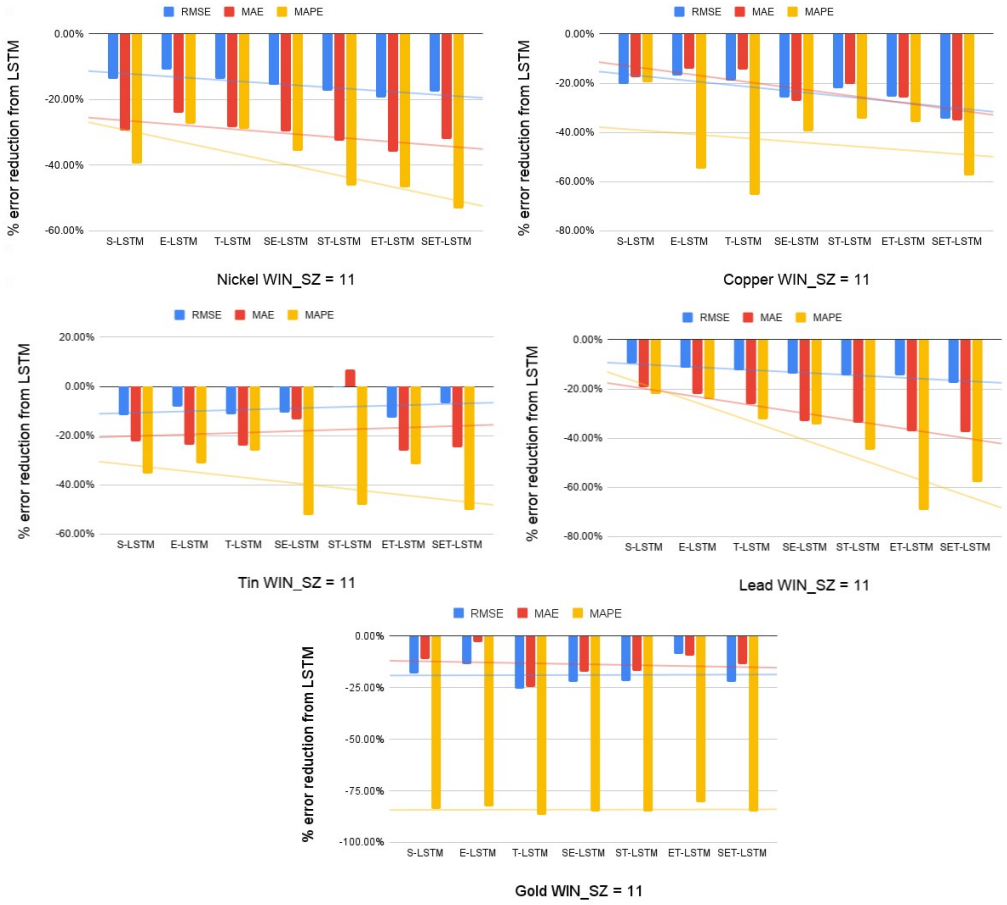
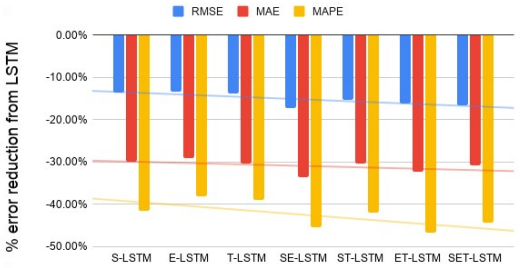
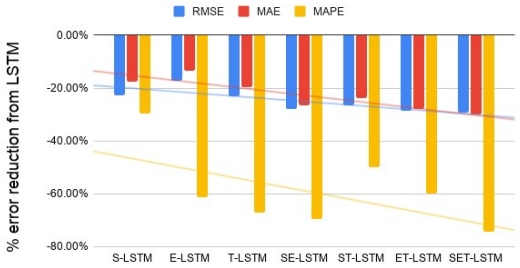


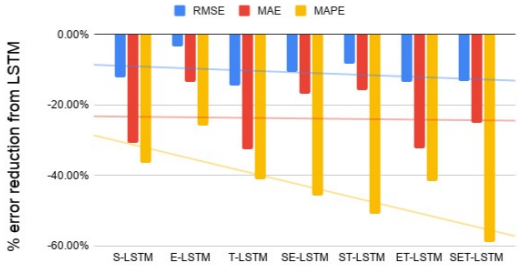
Figure 10: Comparison of prediction errors of the hybrid models with the LSTM model for the 22-day rolling window size



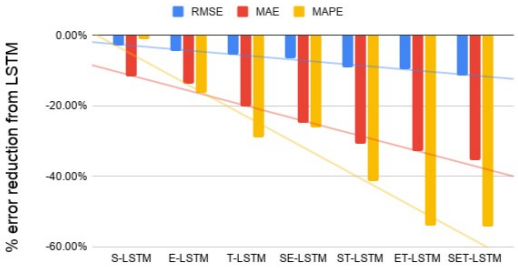
Nickel WIN SZ = 22



Copper WIN_SZ = 22



Tin WIN_SZ = 22



Lead WIN_SZ = 22

Gold WIN_SZ = 22