Scientific Journal of Informatics
Vol. 12, No. No. 4, November 2025

p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Performance of SARIMA, LSTM, GRU and Ensemble Methods for Forecasting Nickel Prices

Irdayanti[1], Khairil Anwar Notodiputro[2]*, Sachnaz Desta Oktarina[3]
[1,2,3] Department of Statistics and Data Science, IPB University, Indonesia

Abstract.
Purpose: There are several forecasting methods, including SARIMA, LSTM, and GRU, which are often claimed to exhibit strong performance in capturing patterns in time series data. However, few studies have conducted direct comparisons among these methods. Therefore, it is necessary to conduct a performance evaluation using empirical data, particularly nickel prices data. This study also aims to improve forecasting performance by combining prediction outputs from deep learning-based models.
Methods: This study utilized data on monthly global nickel prices from January 1990 to May 2025. The models developed include SARIMA, LSTM, GRU, and two ensemble approaches: Weighted Averaging and Bayesian Model Averaging (BMA). Model validation was conducted using walk-forward validation with a sliding window approach to evaluate each model's generalization performance on out-of-sample validation data. The performance was evaluated using MAPE, RMSE, and MAE.
Result: The BMA Ensemble approach shows the best performance in forecasting nickel prices, with a MAPE value of 5.39%, RMSE of 1897.84, and MAE of 1133.96. Prediction validation produces MAPE values below 10%, which indicates that the forecasting results are accurate. The ensemble BMA approach is able to produce more accurate and stable predictions compared to other models.
Novelty: This study offers a novel approach combining LSTM and GRU through ensemble methods to forecast global nickel prices using monthly historical data from 1990 to 2025. In contrast to previous studies that relied on single models, the proposed method with the ensemble BMA approach demonstrates improved forecasting accuracy and stability.

Keywords: Decent work and economic growth, Deep learning, Ensemble forecast, Volatility
Received August 2025 / **Revised** October 2025 / **Accepted** November 2025

INTRODUCTION

In recent development, the use of a seasonal model has been developed rapidly. One of the seasonal models is Seasonal Autoregressive Integrated Moving Average (SARIMA). SARIMA is a development of the ARIMA model designed to handle seasonal components in time series data by adding seasonal parameters to the ARIMA model. This method often assumes a linear relationship between variables, which may not be appropriate in complex systems. This limitation makes it difficult for the method to capture nonlinear patterns that are often found in time series data [1].

Along with the development of information technology, deep learning-based methods such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are currently being developed to analyze time series data more adaptively in capturing complex and nonlinear patterns in the data. Both methods were used to overcome the weakness of Recurrent Neural Network (RNN) in handling long-term dependency due to the vanishing gradient problem [2]. Study [3] compared the RNN, LSTM, and GRU methods for economic and financial data. The results show that GRU consistently provides the best performance. Although both methods have proven effective in modeling complex time series, each model's performance may differ according to the characteristics of the data. Therefore, a combination of these methods can be an alternative solution to improve prediction accuracy and precision. Combining forecasting results from different methods is often more effective than choosing the best one from a single model [4]. [5] Demonstrated that ensembles can reduce the variance and bias of individual models, thereby improving the robustness of predictions.

Forecast combination can be carried out using the ensemble weighted averaging approach, in which each model is assigned a weight based on its performance. Models with lower error rates are given higher weights. Referring to [6], there is another approach, namely Bayesian Model Averaging (BMA), which

[1]*Corresponding author.
Email addresses: irdairdayanti@apps.ipb.ac.id (Irdayanti), khairil@apps.ipb.ac.id (Notodiputro), sachnazdes@apps.ipb.ac.id (Oktarina)
DOI: 10.15294/sji.v12i4.32225

combines prediction results by considering the probability distribution of each model. The weights are based on the posterior probabilities, resulting in a more accurate and precise prediction combination. A study by [7] shows that ensemble BMA can improve prediction accuracy. This study tells that ensemble BMA is significantly more accurate in out-of-sample predictions compared to the individual models that comprise it.

Forecasting of time series data has been extensively conducted for various commodities, one of which is nickel prices. Nickel prices data is selected in this research due to its strong relevance to the Sustainable Development Goals (SDGs), particularly Goal 8: Decent Work and Economic Growth. Nickel is one of the world's most important metal commodities, widely used across various industries. It is chiefly employed in the manufacture of stainless steel and other high-temperature and corrosion-resistant alloys. As a strategic industrial metal, nickel also plays a vital role in the energy sector, electric vehicle (EV) battery production, power generation, coin manufacturing, aerospace, and military fields. Its broad range of uses highlights nickel's critical contribution to modern technology and industrial development [8], [9], [10]. Nickel prices tend to fluctuate sharply due to various influencing factors. Therefore, accurate forecasting of nickel prices is crucial to support sound decision-making by governments, industry players, investors, and market analysts.

Several previous studies have been conducted for nickel price forecasting. A study by [11] utilized an LSTM model optimized with the Improved Particle Swarm Optimization (PSO) algorithm. The results showed that the PSO-LSTM method achieved good performance, with an MAPE of 4.11%. Another study by [12] also, compared the performance of LSTM and GRU models for nickel price forecasting. The results indicated that both LSTM and GRU effectively predicted nickel price fluctuations, with MAPE values of 7.060% and 6.986%, respectively. Additionally, GRU is 33% more computationally efficient than LSTM. Beyond these single-model approaches, ensemble methods have also been applied in other commodities, such as crude oil, where [13] demonstrated that combining multiple models improved forecasting accuracy.

Previous studies mainly focused on comparing individual models. They have not examined the comparison among statistical, deep learning, and ensemble models for forecasting nickel prices. This study aims to fill that gap by comparing the performance of SARIMA, LSTM, GRU models, as well as the ensemble approaches of weighted averaging and BMA to combine LSTM and GRU forecasts, to identify the most accurate approach for forecasting nickel prices. The use of BMA in this context is a relatively unexplored approach, making it a novel contribution of this study. [14] showed that BMA can enhance forecast accuracy in exchange rate prediction. This method is expected to provide more accurate forecasts for nickel prices. The best-performing model based on the evaluation results will then be used to forecast nickel prices for the next 12 months, which can support stakeholders in anticipating market fluctuations and making better investment and policy decisions.

METHODS

Data Description

Global nickel prices on a monthly basis ($/mt) are used in this study, starting from January 1990 to May 2025, with a total of 425 observations. The data were obtained from the World Bank and sourced from the London Metal Exchange (LME), representing nickel prices in the global market.

Flowchart

In general, the workflow of this study includes data exploration, splitting the data into training data (January 1990 - December 2019) and testing data (January 2020 - May 2025). Modeling is carried out using SARIMA, LSTM, GRU, and ensemble methods. The best model is determined based on the smallest error values, namely Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Finally, the best-performing model is subsequently used to produce a 12-month forecast for June 2025–May 2026, which extends beyond the available dataset. The research workflow is illustrated in Figure 1.

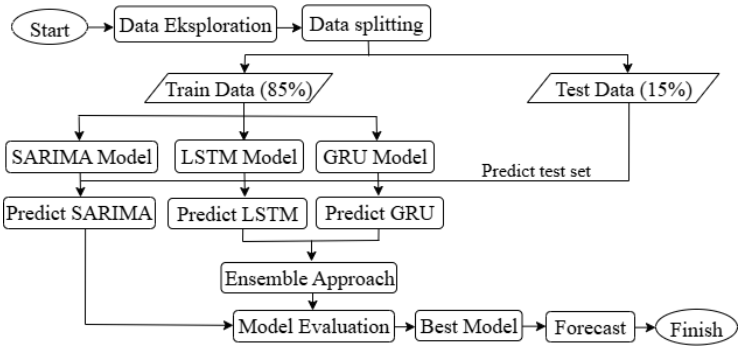


Figure 1. Research Flowchart

Seasonal Autoregressive Integrated Moving Average (SARIMA)

SARIMA is denoted as ARIMA$(p, d, q)(P, D, Q)^S$. The general form of the SARIMA model is expressed in Equation 1.

$$\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D Y_t = \theta_q(B)\Theta_Q(B^S)\varepsilon_t \quad (1)$$

where (p, d, q) represent the orders of the non-seasonal Autoregressive (AR), non-seasonal differencing, and non-seasonal Moving Average (MA) components, respectively. (P, D, Q) denote the seasonal components of the model, with S as the number of seasonal periods and B as the backshift operator [15].

The SARIMA modelling procedure consists of the following steps:

1) Checking the stationarity of variance and mean. If the data are not stationary in variance and mean, a Box-Cox transformation and differencing are applied.
2) Identifying the model by examining the plots of the Autocorrelation Function (ACF), Partial Autocorrelation Function (PACF), and Extended Autocorrelation Function (EACF).
3) Estimating parameters and testing the significance of the estimated parameters from the tentative model.
4) Conducting a series of model diagnostic tests to assess the model's adequacy for the data. Diagnostic tests include testing the independence of residuals, the homogeneity of residual variance, and the normality of residuals.
5) Performing overfitting by increasing the orders of p, q, P, and Q from the tentative model while still ensuring that the model meets parameter significance and residual assumptions.
6) Using the lowest Akaike's Information Criterion (AIC) value to determine which SARIMA model is the best.

Long Short-Term Memory (LSTM)

LSTM was first introduced by Hochreiter and Schmidhuber in 1997 as an improvement to the RNN architecture [16]. It replaces the update mechanism in the hidden layer of a standard RNN with a specialized structure known as the memory cell or cell state. This component can store and retain information over extended time periods [17]. Furthermore, the three gates of the LSTM, the forget, input, and output gates, control the information flow in the network. LSTM architecture is illustrated in Figure 2.

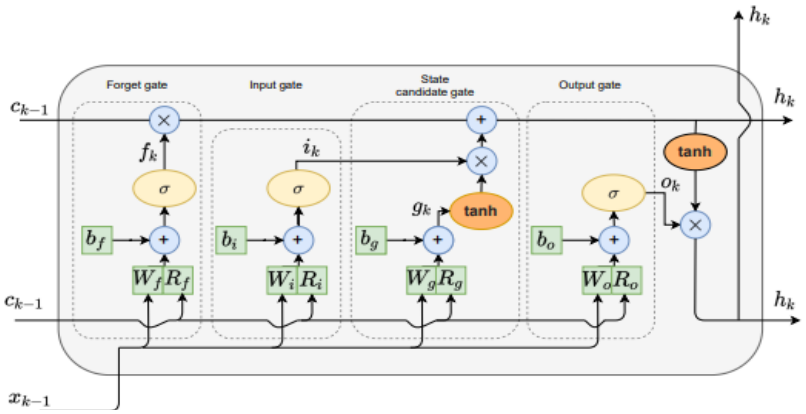


Figure 2. LSTM Architecture [18]

$$f_k = \sigma\left(\mathrm{W}_f x_k + \mathrm{R}_f h_{k-1} + b_f\right) \quad (2)$$

$$i_k = \sigma(\mathrm{W}_i x_k + \mathrm{R}_i h_{k-1} + b_i) \quad (3)$$

$$g_k = tanh\left(\mathrm{W}_g x_k + \mathrm{R}_g h_{k-1} + b_g\right) \quad (4)$$

$$c_k = f_k \circ c_{k-1} + i_k \circ g_k \quad (5)$$

$$o_k = \sigma(\mathrm{W}_o x_k + \mathrm{R}_o h_{k-1} + b_o) \quad (6)$$

$$h_k = o_k \circ tanh\,(c_k) \quad (7)$$

The LSTM architecture processes sequential information through several key components. The forget gate (Equation 2) determines which information from the previous cell state (c_{k-1}) should be retained or discarded. The input gate (Equation 3), together with the candidate memory vector g_k (Equation 4), regulates the new information to be stored. The cell state (c_k) is updated by combining the information from the forget gate and the input gate, as shown in Equation 5. The output gate (Equation 6) is computed using a sigmoid activation function and determines the proportion of the cell state to be realesed at the current time step. The hidden state (h_k) in Equation 7 is then obtained by multiplying the output gate value (o_k) by the cell state (c_k) after it has been activated with the tanh function. The value of c_k and h_k become the outputs of the LSTM model to be used in the next step. In practice, the gate mechanisms allow the model to balance long-term patterns with short-term fluctuations when applied to nickel price data. The input and output gates make sure that recent market signals are included in the prediction process, while the forget gate lessens the effect of transient shocks. This allows the LSTM to capture both overall trends and sudden changes, which are common in commodity prices such as nickel.

Gated Recurrent Unit (GRU)

GRU was first introduced by Kyunghyun Cho in 2014 [19]. GRU has two main gates: the reset gate and the update gate. In the GRU architecture, the update gate combines the functions of the input gate and forget gate found in LSTM. Meanwhile, the reset gate replaces the role of the output gate in LSTM. GRU architecture is illustrated in Figure 3.

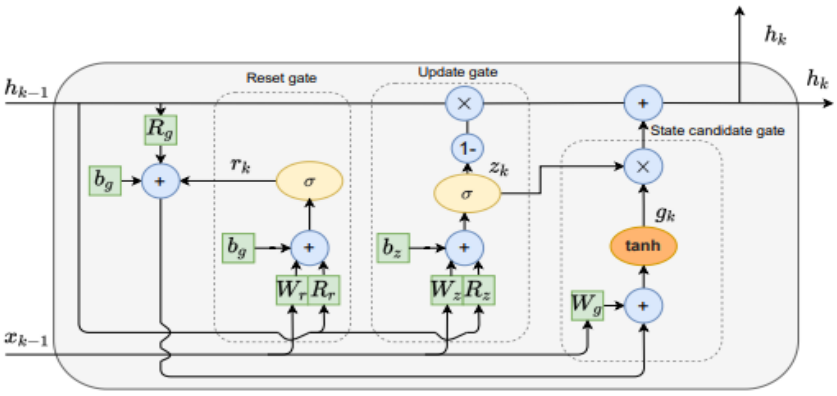


Figure 3. GRU Architecture [18]

$$r_k = \sigma(\mathrm{W}_r x_k + \mathrm{R}_r h_{k-1} + b_r) \quad (8)$$

$$z_k = \sigma(\mathrm{W}_z x_k + \mathrm{R}_z h_{k-1} + b_z) \quad (9)$$

$$g_k = tanh\left(\mathrm{W}_g x_k + r_k \circ \mathrm{R}_g h_{k-1} + b_g\right) \quad (10)$$

$$h_k = \left(1_{n_N \times 1} - z_k\right) \circ g_k + z_k \circ h_{k-1}\right) \quad (11)$$

$$y_k = \mathrm{W}_y h_k + b_y \quad (12)$$

The information update process in GRU begins with the reset gate (r_k), which determines the extent to which information from the previous (h_{k-1}) is retained (Equation 8). The update gate (z_k) controls the extent to which the previous hidden state (h_{k-1}) is retained (Equation 9). Subsequently, the candidate hidden state (g_k) is computed using Equation 10. The updated hidden state (h_k) is then obtained based on Equation 11. Finally, the model output (y_k) is determined by multiplying the hidden state by the corresponding weights, as described in Equation 12. In practice, GRU can balance long-term patterns with short-term fluctuations. The reset gate reduces the impact of historical data during abrupt market

movements, and the update gate takes into account recent price changes, so that GRU is effective for capturing both long-term patterns and sudden changes in erratic commodity markets.

Ensemble Forecast

The ensemble forecasting approach is applied to combine the prediction results from the developed models. Two ensemble methods utilized in this study are weighted averaging and BMA. Weighted averaging produces a combined output by calculating the mean of the individual model outputs using assigned weights, where each model is given a weight that reflects its level of importance or contribution. This importance indicates the extent to which a model's predictions are trusted or influential in generating the final forecast. More accurate models are assigned larger weights, thereby exerting a greater influence on the overall result. In this study, the weights were determined based on the RMSE calculated on data validation, with lower RMSE values corresponding to higher weights. Specifically, the combined prediction $H(x)$ using weighted averaging is given in Equation 13 [20].

$$Y(x) = \sum_{i=1}^{T} w_i y_i(x) \tag{13}$$

where $Y(x)$ represents the combined prediction of all models, $y_i(x)$ denotes the prediction of the i-th individual model, T is the total number of models used, and w_i is the weight assigned to the i-th model. The weights w_i are subject to the following constraints: $w_n \geq 0$ and $\sum_{n=1}^{N} w_n = 1$.

BMA was originally developed as a method to combine predictions and inferences from multiple statistical models. [21] extended the application of BMA to post-analysis statistical processing for forecast ensembles. BMA integrates various forecasting results by accounting for prior uncertainty regarding the best model.

Let y^* denote the quantity to be predicted at time $t^* \in T^*$. The previous predictions for y^* at period $t \in T$ are generated by K models: $M_1, M_2, \ldots, M_K$. Each model M_K is assumed to follow a prior probability distribution $M_K \sim \pi(M_k)$ and the Probability Density Function (PDF) for y^* is $p(y^*|M_k)$. By applying Bayes' rule, the posterior probability $p(M_k|y^*)$ is calculated as shown in Equation 14.

$$p(M_k|y^*) = \frac{p(y^*|M_k)\pi(M_k)}{\sum_{k=1}^{K} p(y^*|M_k)\pi(M_k)} \tag{14}$$

The marginal predictive PDF is expressed in Equation 15.

$$p(y^*) = \sum_{k=1}^{K} p(y^*|M_k)\, p(M_k|y^*) \tag{15}$$

The BMA PDF can be interpreted as a weighted average of the component PDFs, where the weights are determined by the performance of each model during the previously observed period T. In a dynamic setting, the data are partitioned into three periods: training, validation, and testing. The combined forecast is then computed as shown in Equation 16.

$$p\left(y^{s|t^*} \middle| f_1^{s|t^*}, \ldots, f_K^{s|t^*}\right) = \sum_{k=1}^{K} w_k g_k \left(y^{s|t^*} \middle| f_K^{s|t^*}\right) \tag{16}$$

where $w_k \in [0,1]$ represents the weight of model k and satisfies $\sum_{k=1}^{K} w_k = 1$. In simple terms, these weights are associated with the predictive performance of each component model during the validation period, while also considering the extent to which each model provides distinct or unique information (e.g., predictions from one model differ from those of others) [7].

Walk Forward Validation (WFV)

Model validation was conducted using the Walk Forward Validation (WFV) approach, which evaluates the model's ability to generalize previously unseen data by using sequential training and validation windows. This method employs a fixed-length sliding window strategy, in which both the training and test sets maintain a constant size while moving forward in time [22]. In this study, a 24-month shift was applied at each iteration to simulate periodic model re-training and evaluation. The training data were divided into six

subsets (windows), each consisting of sequentially organized training and validation sets that followed the temporal structure of the dataset. There were 24 months (2 years) of validation data and 216 months (18 years) of training data in each window. To make sure that the validation periods in each window did not overlap, a 24-month shift was selected. The details of the data split for each window are presented in Table 1. Meanwhile, data from January to May 2025 were kept separate as the final testing set for out-of-sample prediction.

Table 1. Splitting data for walk forward validation

Window	Training Data (January – December)	Validation Data (January – December)
1	1990 – 2007	2008 – 2009
2	1992 – 2009	2010 – 2011
3	1994 – 2011	2012 – 2013
4	1996 – 2013	2014 – 2015
5	1998 – 2015	2016 – 2017
6	2000 – 2017	2018 – 2019

Model Evaluation

Model evaluation is a critical stage for assessing the performance of the forecasting models. Forecasting accuracy serves as a key measure to determine how well the methods perform in predicting data. One common approach to evaluate models is by calculating error metrics such as MAPE, RMSE, and MAE. MAPE measures the average prediction error as a percentage. RMSE focuses on the square root of the mean squared differences between actual and predicted values, assigning greater weight to larger errors. MAE calculates the average error based on the absolute differences between actual and predicted values [23]. The formulas for MAPE, RMSE, and MAE are presented in Equation 17.

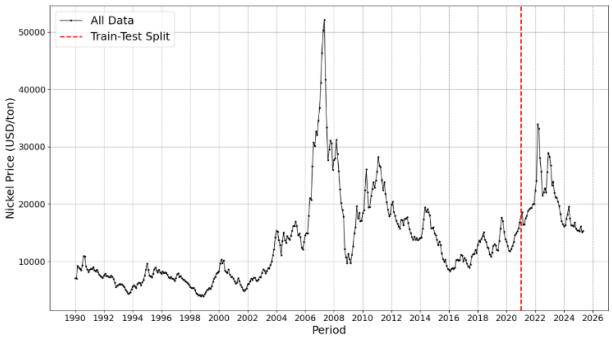
$$MAPE = \frac{1}{n}\sum_{t=1}^{N}\left|\frac{Y_t - \hat{Y}_{(t)}}{Y_t}\right| \times 100\% \; ; \quad RMSE = \sqrt{\frac{\sum_{t-1}^{n}\left(Y_t - \hat{Y}_{(t)}\right)^2}{n}} ; \quad MAE = \frac{\sum_{t-1}^{n}|Y_t - \hat{Y}_{(t)}|}{n} \tag{17}$$

where Y_t denotes the actual nickel price at period t, $\hat{Y}_{(t)}$ is the predicted value of nickel price at period t, and n is the number of predicted observations [24].

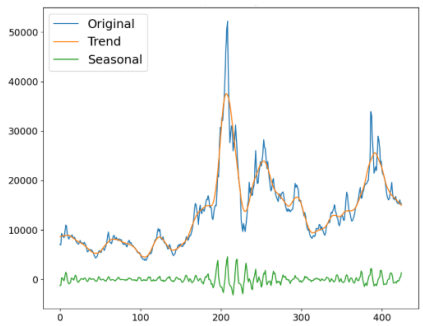
RESULT AND DISCUSSION

Data Exploration

The global nickel prices from January 1990 to May 2025 exhibit a fluctuating pattern with several periods of sharp spikes and declines, as illustrated in Figure 4(a). The most significant surge occurred in May 2007, when prices reached over USD 50,000 per ton. This extreme increase was driven by massive demand from China, fueled by the rapid growth of its stainless-steel industry [25], [26]. In 2008, nickel prices experienced a steep decline due to the global financial crisis. Between 2009 and 2012, the global economy began to recover, industrial capacity increased, and nickel prices rose again [27]. Indonesia's ban on nickel ore exports in 2014 further contributed to rising global nickel prices by reducing supply for the world's nickel processing industries [28]. Prices climbed again in 2021, driven by a significant surge in nickel demand, which was also supported by increased stainless steel production and soaring demand for electric vehicle batteries [29], [30]. Russia's invasion of Ukraine triggered concerns over global supply, leading to a sharp price spike in early 2022 [31].



(a)



(b)

Figure 4. (a) Plot of nickel prices data from January 1990 to May 2025; (b) Decomposition result of nickel prices

The decomposition of nickel prices data using Seasonal-Trend decomposition based on Loess (STL) is presented in Figure 4(b). This method separates a time series into three main components: trend, seasonal, and residual. In the displayed graph, only trend and seasonal components are presented to visualize the underlying tendencies and seasonal fluctuations in the data. The decomposition results indicate that the trend component exerts a dominant influence on the dynamics of nickel prices, while the seasonal component appears relatively small but still reveals a consistent seasonal pattern within the data.

SARIMA Model

An essential step before building the SARIMA model is to check data stationarity. This model assumes that the data are stationary, that is, they have constant mean and variance over time. The initial check can be done visually through the rolling mean and standard deviation plots shown in Figure 5. The data can be regarded as stationary if the rolling mean and rolling standard deviation plots show little variation over time or appear to be flat. On the other hand, the data is probably non-stationary if the graphs exhibit significant fluctuations. The visualization result shows a change in the mean and standard deviation over time, which indicates that the data is not yet stationary. This result is validated through statistical testing, namely the ADF test (p-value > 0.05) and the Bartlett test (p-value < 0.05), which means that the data is not stationary in mean and variance.

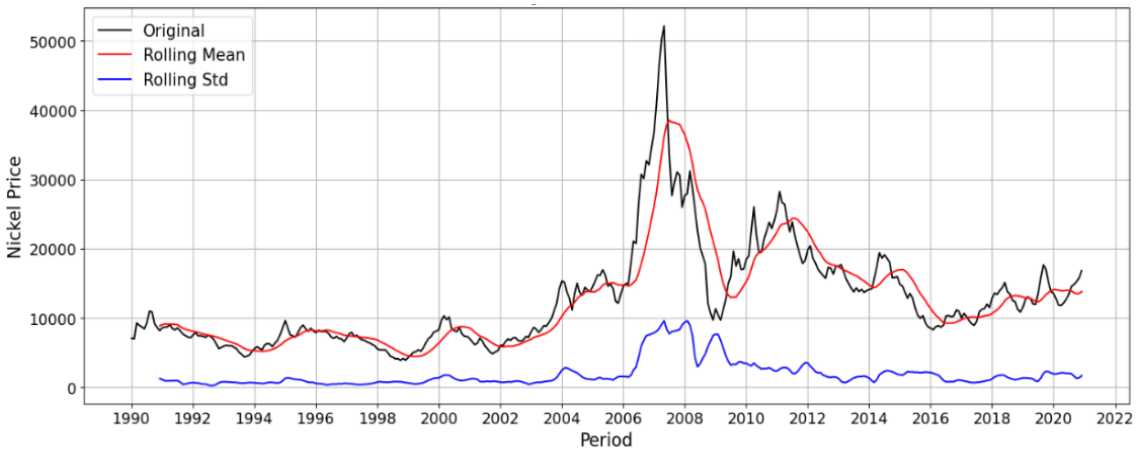
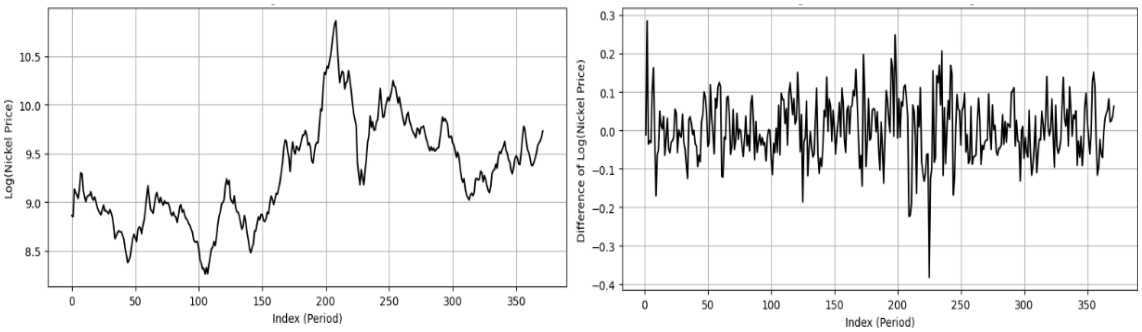


Figure 5. Plot of the rolling mean and rolling standard deviation of nickel prices

Transformation is applied to make the data stationary in variance, and differencing is used to make the data stationary in mean. The visualization of the data after transformation and first-order differencing is presented in Figure 6. Following these steps, the data pattern appears more stable, indicating that the stationarity assumption has been satisfied.



(a) (b)

Figure 6. Log-transformed (a) and differenced (b) plots of nickel prices

Identification of order $p, q,$ and P, Q The SARIMA model was carried out using ACF, PACF, and EACF plots applied to the first differenced series. Based on the identification results from the stationary data, the tentative SARIMA model is presented in Table 2.

Table 2. Parameter estimates of the tentative SARIMA model

Model	Parameter	Coefficient	p-value	AIC	Model	Parameter	Coefficient	p-value	AIC
ARIMA(1,1,0)(1,1,0)^{12}	$\hat{\phi}_1$	0.322	0.000*	-631.28	ARIMA(0,1,1)(0,1,1)^{12}	$\hat{\theta}_1$	0.270	0.000*	-753.89
	$\hat{\Phi}_1$	-0.483	0.000*			$\hat{\Theta}_1$	-0.989	0.001*	
ARIMA(1,1,0)(3,1,0)^{12}	$\hat{\phi}_1$	0.270	0.000*	695.64	ARIMA(1,1,2)(1,1,0)^{12}	$\hat{\phi}_1$	0.699	0.000*	-628.51
	$\hat{\Phi}_1$	-0.763	0.000*			$\hat{\theta}_1$	-0.399	0.032*	
	$\hat{\Phi}_2$	-0.536	0.000*			$\hat{\theta}_2$	-0.086	0.319	
	$\hat{\Phi}_3$	-0.205	0.000*			$\hat{\Phi}_1$	-0.485	0.000*	
ARIMA(1,1,0)(2,1,0)^{12}	$\hat{\phi}_1$	0.271	0.000*	683.68	**ARIMA(0,1,2)(0,1,1)^{12}**	$\hat{\theta}_1$	0.307	0.000*	**-755.82**
	$\hat{\Phi}_1$	-0.675	0.000*			$\hat{\theta}_2$	0.109	0.036*	
	$\hat{\Phi}_2$	-0.392	0.000*			$\hat{\Theta}_1$	-0.991	0.010*	

*Parameter estimates are significant at 5% significance level

The ARIMA(0,1,2)(0,1,1)^{12} model is identified as the optimal model based on having the lowest AIC value compared to other models. Each parameter estimate in the model is significant as indicated by a p-value that is smaller than the 5% significance level. The model diagnostic tests, including tests for residual independence, homoscedasticity, and normality, show p-values greater than the 5% significance level, indicating no assumption violations. Thus, the model can be considered valid for forecasting. The ARIMA(0,1,2)(0,1,1)^{12} model was further overfitted by alternately increasing the orders of $p, q, P,$ and Q from the initial model to explore the possibility of a better-fitting model. The results of overfitting model are presented in Table 3.

Table 3. Parameter estimates of the overfitting model

Model	Parameter	Coefficient	z	p-value	AIC
ARIMA(0,1,3)(0,1,1)^{12}	$\hat{\theta}_1$	0.3027	5.679	0.000*	-754.781
	$\hat{\theta}_2$	0.0981	1.865	0.062	
	$\hat{\theta}_3$	-0.0511	-1.044	0.296	
	$\hat{\Theta}_1$	-0.9910	-2.618	0.009*	
ARIMA(0,1,2)(0,1,2)^{12}	$\hat{\theta}_1$	0.3054	5.647	0.000*	-755.755
	$\hat{\theta}_2$	0.1135	2.194	0.028	
	$\hat{\Theta}_1$	-1.0757	-4.091	0.000*	
	$\hat{\Theta}_2$	0.0869	1.631	0.103	

*Parameter estimates are significant at 5% significance level

Based on the results of the model overfitting, it can be observed that both models have parameter estimates that are not significant at the 5% significance level. This indicates that, the most appropriate SARIMA model is ARIMA(0,1,2)(0,1,1)^{12}. The final form of this model, including the estimated coefficients, is presented in Equation 18.

$$(1-B)(1-B^{12})Y_t = (1+0.307B+0.109B^2)(1-0.991B^{12})\varepsilon_t \quad (18)$$

LSTM and GRU Models

To make sure compatibility with the network architecture, LSTM and GRU modeling started with a data preparation step. This step involved normalizing the data, changing the input dimensions, and converting the time series into a supervised learning format. Both models employed a similar architecture, consisting of a primary layer (either LSTM or GRU) followed by a dropout layer to mitigate the risk of overfitting, and a dense layer serving as the final output.

The number of neurons, learning rate, number of epochs, and optimizer were among the parameter combinations that were tested in order to find the best hyperparameters. The number of neurons refers to the number of hidden units in the LSTM and GRU layers. The more neurons used, the higher the model representation capacity, although this may also increase the risk of overfitting. Learning rate is a parameter

that sets the step size in the weight update process during training. Epoch is the number of full iterations when all the training data is used to update the network weights. Batch size refers to the number of data samples used for one model parameter update. The hyperparameters were tuned using the following value ranges: neurons (32, 64, 100, 150, and 200), learning rates (0.01, 0.001, and 0.0001), epochs (50 and 100), batch size (32), and optimizers (Adam and RMSprop). The best hyperparameter combination is determined based on the smallest validation loss or error value during the model training process. The tuning process is performed through a grid search approach combined with a walk-forward validation scheme through a sliding window technique. At this stage, the training data is divided into six subsets, each consisting of training and validation data, as shown in Table 1. An early stopping mechanism was employed during training, in which the process was halted once the validation performance ceased to improve. This approach helps reduce the risk of overfitting while preserving the model's ability to generalize. The best hyperparameter combination for both LSTM and GRU models consists of 200 neurons, a learning rate of 0.001, batch size of 32, and the Adam optimizer. The only difference lies in the number of epochs 100 for LSTM and 50 for GRU.

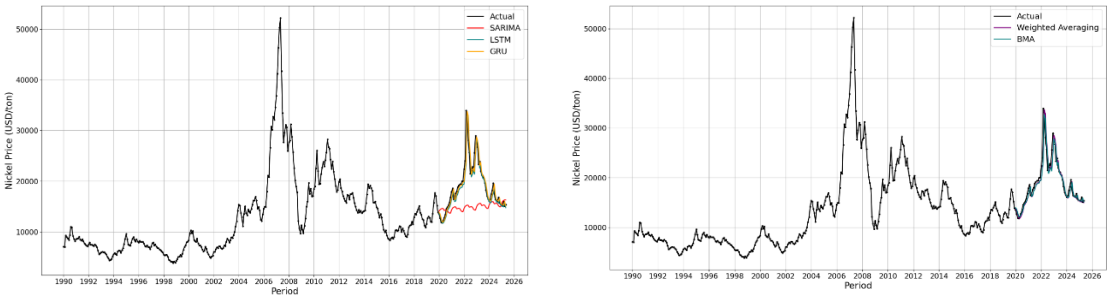
Model Evaluation

The performance of all models was evaluated using MAPE, RMSE, and MAE. The evaluation was conducted on the testing data covering the period from January 2020 to May 2025. The evaluation results are presented in Table 4. Lower values for each metric indicate better predictive performance of the model.

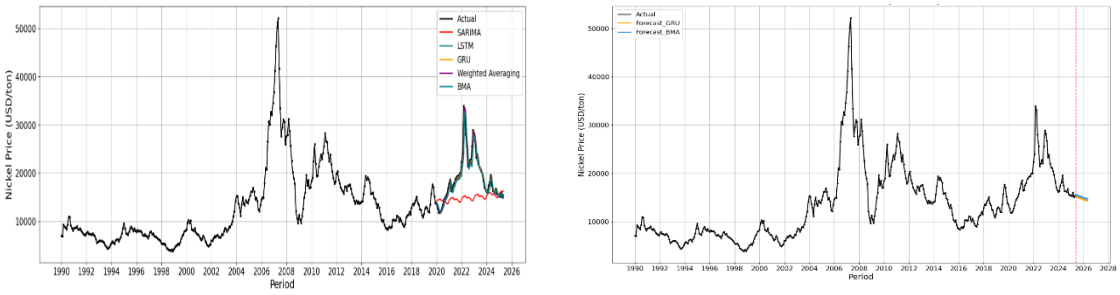
Table 4. Model evaluation

	Model	MAPE	RMSE	MAE
Single Model	$ARIMA(0,1,2)(0,1,1)^{12}$	20.7%	6275.23	4545.33
	LSTM	5.93%	1977.11	1238.61
	GRU	5.45%	1891.80	1150.82
Ensemble Model (LSTM-GRU)	Weighted Averaging	5.63%	1924.34	1186.49
	BMA	5.39%	1897.84	1133.96

Based on the evaluation results, the SARIMA model as a baseline shows the lowest prediction performance, with a MAPE value of 20.7%, RMSE of 6275.23, and MAE of 4545.33. The LSTM and GRU models significantly improve performance compared to SARIMA. Among the two, the GRU model provides better results with a MAPE of 5.45%, RMSE of 1891.80, and MAE of 1150.82. The ensemble method is applied by combining predictions from LSTM and GRU through two approaches, namely weighted averaging and BMA. The evaluation results show that the single GRU model performs better than weighted averaging. The BMA method produces the lowest prediction error value (MAPE 5.39%, RMSE 1897.84, MAE 1133.96), with posterior weights of 0.49 for LSTM and 0.51 for GRU, although the difference is relatively small compared to the GRU single model. The visual results of predictions using single models, ensemble models, and the best-performing models for forecasting are presented in Figure 7.



(a) (b)



(c) (d)

Figure 7. Prediction plots on test data using (a) single models; (b) ensemble models; (c) both single and ensemble models; (d) a 12-month ahead forecast of monthly nickel prices

Figure 7(a) shows that GRU and LSTM models can follow the data patterns well. The GRU model is a slightly better fit to the actual data than the LSTM model, although the difference in performance is relatively small. In contrast, the SARIMA model struggles to capture the data patterns, particularly during significant spikes. Meanwhile, Figure 7(b) shows that the two ensemble methods weighted averaging and BMA also produce smoother predictions that are closer to the actual data. Among all the methods, BMA demonstrates higher accuracy in capturing nickel price fluctuations, as shown visually in Figure 7(c) and based on evaluation metrics in Table 4. Based on these evaluation results, nickel price forecasting for the next 12 months (June 2025–May 2026) was conducted using the two best-performing models, namely GRU and BMA, as illustrated in Figure 7(d). The forecasting period extends beyond the available historical dataset, meaning that the models generate future values based solely on the learned patterns and dependencies from past data, without having access to actual observations for that period. In the BMA method, the forecasts are generated through a weighted combination of predictions from the LSTM and GRU models, with the weights determined by the posterior probabilities of each model. The forecasting results indicate that both the GRU and BMA models forecast a consistent decrease in nickel prices throughout the forecasting period.

CONCLUSION

The performance of SARIMA, LSTM, GRU, and ensemble methods for nickel price forecasting was compared in this study. Two ensemble methods were used: weighted averaging and BMA. These ensemble approaches were applied to combine the predictive outputs of the LSTM and GRU models. The results showed that GRU was the best-performing single model compared to SARIMA and LSTM. Meanwhile, the ensemble BMA produced the most accurate and stable predictions, although its performance margin over GRU was relatively small. A Diebold–Mariano test indicated that the difference in forecast accuracy was not statistically significant (p-value > 0.05). Despite these findings, the study has some constraints. The forecasting models were built only on past price data and did not include other economic indicators such as global steel production, EV sales, or exchange rates, which might improve accuracy. In addition, applying deep learning and BMA requires considerably more computational effort than SARIMA. With these points in mind, future research could test the inclusion of exogenous variables in LSTM and GRU, and examine other ensemble frameworks like stacking.

REFERENCES

[1] D. C. Montgomery, C. L. Jennings, and M. Kulahci, *Wiley Series in Probability and Statistics*, 7th ed. Wiley, 2015.

[2] S.-H. Noh, "Analysis of Gradient Vanishing of RNNs and Performance Comparison," *Information*, vol. 12, no. 11, p. 442, 2021, doi: 10.3390/info12110442.

[3] C. Alkahfi, A. Kurnia, and A. Saefuddin, "Perbandingan Kinerja Model Berbasis RNN pada Peramalan Data Ekonomi dan Keuangan Indonesia," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 4, pp. 1235–1243, Jul. 2024, doi: 10.57152/malcom.v4i4.1415.

[4] X. Wang, R. J. Hyndman, F. Li, and Y. Kang, "Forecast combinations: an over 50-year review," 2022.

[5] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *Proceedings of the First International Workshop on Multiple Classifier Systems*, Springer, 2000, pp. 1–15. doi: 10.1007/3-540-45014-9_1.

[6] K. Duan, X. Wang, B. Liu, T. Zhao, and X. Chen, "Comparing Bayesian Model Averaging and Reliability Ensemble Averaging in Post-Processing Runoff Projections under Climate Change," *Water (Basel)*, vol. 13, no. 15, p. 2124, Aug. 2021, doi: 10.3390/w13152124.
[7] J. M. Montgomery, F. M. Hollenbach, and M. D. Ward, "Improving Predictions using Ensemble Bayesian Model Averaging," *Political Analysis*, vol. 20, no. 3, pp. 271–291, Jan. 2012, doi: 10.1093/pan/mps002.
[8] N. Syarifuddin, "Pengaruh Industri Pertambangan Nikel Terhadap Kondisi Lingkungan Maritim di Kabupaten Morowali," *Jurnal Riset & Teknologi Terapan Kemaritiman,* vol. 1, no. 2, pp. 19–23, 2022.
[9] Y. Zhu, D. Xu, S. H. Ali, and J. Cheng, "A hybrid assessment model for mineral resource availability potentials," *Resources Policy*, vol. 74, p. 102283, 2021, doi: 10.1016/j.resourpol.2021.102283.
[10] British Geological Survey, "Nickel," www.mineralsuk.com.
[11] B. Shao, M. Li, Y. Zhao, and G. Bian, "Nickel Price Forecast Based on the LSTM Neural Network Optimized by the Improved PSO Algorithm," *Math Probl Eng*, vol. 2019, no. 1, 2019, doi: 10.1155/2019/1934796.
[12] A. C. Ozdemir, K. Buluş, and K. Zor, "Medium- to long-term nickel price forecasting using LSTM and GRU networks," *Resources Policy*, vol. 78, p. 102906, 2022, doi: 10.1016/j.resourpol.2022.102906.
[13] M. Hasan, M. Z. Abedin, P. Hajek, K. Coussement, Md. N. Sultan, and B. Lucey, "A blending ensemble learning model for crude oil price forecasting," *Ann Oper Res*, 2024, doi: 10.1007/s10479-023-05810-8.
[14] M. P. Bonel, "Combination of theoretical models for exchange rate forecasting," *Cuadernos de Economía*, vol. 43, no. 92, Oct. 2024, doi: 10.15446/cuad.econ.v43n92.98393.
[15] K. E. ArunKumar, D. V. Kalaga, Ch. M. Sai Kumar, G. Chilkoor, M. Kawaji, and T. M. Brenza, "Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA)," *Appl Soft Comput*, vol. 103, p. 107161, May 2021, doi: 10.1016/j.asoc.2021.107161.
[16] S. M. Al-Selwi *et al.*, "RNN-LSTM: From applications to modeling techniques and beyond—Systematic review," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 5, p. 102068, Jun. 2024, doi: 10.1016/j.jksuci.2024.102068.
[17] P. L. Seabe, C. R. B. Moutsinga, and E. Pindza, "Forecasting Cryptocurrency Prices Using LSTM, GRU, and Bi-Directional LSTM: A Deep Learning Approach," *Fractal and Fractional*, vol. 7, no. 2, p. 203, Feb. 2023, doi: 10.3390/fractalfract7020203.
[18] K. Zarzycki and M. Ławryńczuk, "LSTM and GRU Neural Networks as Models of Dynamical Processes Used in Predictive Control: A Comparison of Models Developed for Two Chemical Reactors," *Sensors*, vol. 21, no. 16, p. 5625, Aug. 2021, doi: 10.3390/s21165625.
[19] A. Islam, "Cryptocurrency Price Prediction: A Comparative Study using LSTM, GRU and Stacking Ensemble Algorithm for Time Series Forecasting," Feb. 2022, pp. 465–477. doi: 10.54389/NTPV9785.
[20] I. Kalinina, P. Bidyuk, A. Gozhyj, P. Malchenko, and P. M. Black, "Combining Forecasts Based on Time Series Models in Machine Learning Tasks."
[21] A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski, "Using Bayesian Model Averaging to Calibrate Forecast Ensembles," *Mon Weather Rev*, vol. 133, no. 5, pp. 1155–1174, May 2005, doi: 10.1175/MWR2906.1.
[22] M. Mailoc. López de Prado, *Advances in financial machine learning*. John Wiley & Sons, Inc., 2018.
[23] Ma. del R. C. Estrada, M. E. G. Camarillo, M. E. S. Parraguirre, M. E. G. Castillo, E. M. Juárez, and M. J. C. Gómez, "Evaluation of Several Error Measures Applied to the Sales Forecast System of Chemicals Supply Enterprises," *International Journal of Business Administration*, vol. 11, no. 4, p. 39, Jun. 2020, doi: 10.5430/ijba.v11n4p39.
[24] W. Zha *et al.*, "Forecasting monthly gas field production based on the CNN-LSTM model," *Energy*, vol. 260, p. 124889, Dec. 2022, doi: 10.1016/j.energy.2022.124889.
[25] J. Stepan, "An Analysis of Nickel Price Variation and Its Impact on the Global Economy," AGH Faculty of Management - XVII International Scientific Conference, AGH University of Krakow, 2015.
[26] "Asian Metal Ltd., 2007 Annual Report on Chinese Nickel Market [Online]," 2007.

[27] X. Zhou *et al.*, "Risk Transmission of Trade Price Fluctuations from a Nickel Chain Perspective: Based on Systematic Risk Entropy and Granger Causality Networks," *Entropy*, vol. 24, no. 9, p. 1221, Aug. 2022, doi: 10.3390/e24091221.
[28] B. Lim, H. S. Kim, and J. Park, "Implicit Interpretation of Indonesian Export Bans on LME Nickel Prices: Evidence from the Announcement Effect," *Risks*, vol. 9, no. 5, p. 93, May 2021, doi: 10.3390/risks9050093.
[29] International Nickel Study Group (INSG), "INSG Press Release October 2021," https://insg.org/INSG_Press_Release_Oct_2021.pdf.
[30] Reuters, "Nickel booms on short squeeze while other metals retreat," 2022.
[31] London Metal Exchange (LME), "Independent Review of the March 2022 Nickel Market Events," https://www.lme.com/Nickel-independent-Final-Report.pdf.