http://doi.org/10.35784/iapgos.6955 received: 17.12.2024 | revised: 01.07.2025 | accepted: 21.08.2025 | available online: 30.09.2025

OPTIMIZING DEEP LEARNING TECHNIQUES WITH STACKING BILSTM AND BIGRU MODELS FOR GOLD PRICE PREDICTION

Iqbal Kharisudin[1], Nike Yustina Oktaviani[2]
[1]Universitas Negeri Semarang, Department of Statistics and Data Science, Semarang, Indonesia, [2]Universitas Negeri Semarang, Department of Mathematics, Semarang, Indonesia

Abstract. *Gold, essential in various sectors from jewelry to financial reserves, plays a crucial role as a financial asset and investment. Accurate prediction of gold prices is vital for informed investment decisions and economic management. This study utilizes a Stacking Ensemble approach to enhance gold price prediction accuracy by combining Bidirectional Long Short-Term Memory (BiLSTM) and Bidirectional Gated Recurrent Unit (BiGRU) as base learners, with Random Forest serving as the meta-learner. This method capitalizes on BiLSTM and BiGRU's strengths in processing sequential data in both directions, capturing intricate patterns in gold price fluctuations. The dataset spans from January 1, 2020, to May 31, 2024. The results reveal that the Stacking Ensemble model with BiLSTM-BiGRU consistently outperforms other models, achieving the lowest Mean Squared Error (MSE) of 0.000, Root Mean Squared Error (RMSE) of 0.0067, Mean Absolute Error (MAE) of 0.0050, Mean Absolute Percentage Error (MAPE) of 0.0083, and a high R-squared value of 0.9984 across various lookback periods (7, 15, and 30 days). These metrics underscore the method's effectiveness in accurately capturing and predicting gold price trends. This confirms that the Stacking Ensemble approach significantly enhances gold price prediction accuracy.*

Keywords: stacking ensemble, deep learning, BiLSTM, BiGRU, gold price forecasting

OPTYMALIZACJA TECHNIK GŁĘBOKIEGO UCZENIA SIĘ POPRZEZ ŁĄCZENIE MODELI BILSTM I BIGRU DO PRZEWIDYWANIA CEN ZŁOTA

Streszczenie. *Złoto, istotne w różnych sektorach, od biżuterii po rezerwy finansowe, odgrywa kluczową rolę jako aktywo finansowe i inwestycyjne. Dokładne przewidywanie cen złota ma zasadnicze znaczenie dla podejmowania świadomych decyzji inwestycyjnych i zarządzania gospodarką. Niniejsze badanie wykorzystuje podejście Stacking Ensemble w celu poprawy dokładności prognoz cen złota, łącząc dwukierunkowe modele Long Short-Term Memory (BiLSTM) i Gated Recurrent Unit (BiGRU) jako modele podstawowe, z Random Forest jako meta-modelem. Metoda ta wykorzystuje zdolność BiLSTM i BiGRU do przetwarzania danych sekwencyjnych w obu kierunkach, umożliwiając uchwycenie złożonych wzorców w zmianach cen złota. Zbiór danych obejmuje okres od 1 stycznia 2020 roku do 31 maja 2024 roku. Wyniki pokazują, że model Stacking Ensemble z BiLSTM-BiGRU konsekwentnie przewyższa inne modele, osiągając najniższy błąd średniokwadratowy (MSE) wynoszący 0,000, średnią kwadratową błędów (RMSE) 0,0067, średni błąd bezwzględny (MAE) 0,0050, średni bezwzględny błąd procentowy (MAPE) 0,0083 oraz wysoką wartość współczynnika determinacji (R^2) wynoszącą 0,9984 w różnych okresach wstecznych (7, 15 i 30 dni). Te metryki podkreślają skuteczność tej metody w dokładnym uchwyceniu i przewidywaniu trendów cen złota. Potwierdza to, że podejście Stacking Ensemble znacząco zwiększa dokładność prognoz cen złota.*

Słowa kluczowe: stacking ensemble, uczenie głębokie, BiLSTM, BiGRU, prognozowanie cen złota

Introduction

Gold, beyond its use for adorning jewellery, is a crucial raw material for the manufacturing industry, playing an indispensable role in various sectors. It combines the characteristics of a commodity, a precious metal, and a currency, making it a highly valuable asset in the global economy that serves multiple functions. Its significance is underscored by the large amounts of gold included in the international reserves of most central banks, which often reflect the economic stability and strength of a nation [13]. Central banks worldwide hold gold reserves not only to protect the value of deposits and ensure the confidence of currency holders but also to safeguard the interests of foreign debt creditors. They also utilize these reserves as a strategic tool to control inflation, a phenomenon that occurs when too much money chases too few goods, thereby destabilizing economies. This mechanism is crucial for strengthening a country's financial position and maintaining overall economic health [15]. Notably, gold prices remained stable during five significant pandemics from 1957 to 2009, highlighting gold's reliability as a safe-haven asset during times of uncertainty [22]. As financial markets have evolved, the significance of the gold market has steadily grown, establishing itself as a vital component of the global investment landscape. Gold is now recognized as a financial asset comparable to other major markets, such as stocks, futures, and bonds. This growing importance highlights gold's dual role as both a safe-haven asset and a profitable investment vehicle. Its appeal extends to a wide range of investors, from those seeking to hedge against market volatility to those pursuing steady returns. This evolution solidifies gold's reputation as a versatile and resilient financial instrument, capable of providing stability and long-term value in the face of economic fluctuations [5]. Given the widespread popularity and importance of gold, there is a growing interest in forecasting its price, as accurate predictions can significantly influence investment strategies and economic planning.

Over the decades, extensive research on gold price prediction, its movements, and the influencing variables has been conducted, leading to the development of various strategies. Classical time series techniques, such as multilinear regression and Auto-Regressive Integrated Moving Average (ARIMA), have been employed for gold price prediction, providing foundational models that capture historical price movements [24]. Besides conventional time series approaches, several machine learning methods have been employed to capture the complexities of gold prices, offering advanced analytical capabilities that traditional methods may not achieve [3]. Besides forecasting with machine learning, deep learning techniques have proven effective in addressing a range of complex real-world prediction challenges, such as time series forecasting, demonstrating their capacity to handle extensive datasets and reveal underlying patterns [19]. Deep learning methods are particularly well-suited for handling chaotic time series forecasting problems and producing more accurate predictions due to their sophisticated architectures and training techniques.

A significant part of deep learning is Artificial Neural Networks (ANNs), which have seen rapid advancements and applications in various fields such as forecasting, classification, and regression, mimicking the human brain's data processing mechanisms in an innovative way. Recurrent Neural Networks (RNNs), an extension of ANNs, are specifically designed for sequential data processing and are particularly effective at handling long-term dependencies that traditional models struggle with [27]. Key developments in RNNs include Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, which are tailored to tackle the vanishing gradient issue, thereby making them particularly effective for predicting complex sequential data with greater accuracy [27]. Furthermore, Bidirectional LSTM (BiLSTM) and Bidirectional GRU (BiGRU), which process data in both forward and backward directions, allow models to leverage contextual information from both

directions, thereby enhancing prediction accuracy by considering more comprehensive data inputs [26].

The integration of BiLSTM and BiGRU in prediction tasks is expected to yield more accurate results compared to methods that process data in a single direction, thus offering a more holistic approach to forecasting. Additionally, Ensemble Learning methods, particularly Stacking Ensemble, have been employed to enhance prediction accuracy by combining multiple model predictions [14]. Stacking Ensemble involves integrating predictions from several base learners to produce a superior final prediction compared to individual models, capitalizing on the strengths of diverse approaches. In the context of gold price prediction, a Stacking Ensemble using BiLSTM and BiGRU can effectively leverage the strengths of both methods to improve predictive accuracy, creating a robust framework for analysis [25]. This method entails independently training several base learners and subsequently merging their predictions with a meta-learner to produce the final prediction, enhancing the overall effectiveness of the predictive model. Studies by Ye et al. [25] and Gu et al. [7] have demonstrated that the Stacking Ensemble method produces better predictive outcomes compared to the individual models it combines, reinforcing the efficacy of this approach. In this study, the Stacking Ensemble of BiLSTM and BiGRU is used to predict gold prices, emphasizing its value as a critical resource for investors and policymakers alike.

This approach aims to contribute significantly to the development of more accurate gold price prediction methods, supporting decision-making in investment and economic policy by providing deeper insights into market dynamics and trends.

1. Related works

Predicting gold prices has become increasingly essential for financial analysts and investors due to the complex, multifactorial, and non-linear nature of the gold market. This section reviews recent studies on gold price prediction to provide context. Research from 2021 demonstrated that deep learning techniques, including Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), and Gated Recurrent Units (GRU), significantly enhance prediction accuracy by capturing non-linear trends [26]. In a comparative analysis of these models, LSTM outperformed with a Mean Absolute Percentage Error (MAPE) of 3.48, Root Mean Squared Error (RMSE) of 61,728, and Mean Absolute Error (MAE) of 48.85, indicating that LSTM effectively handles price volatility. These findings emphasize the importance of temporal dependency capture in gold price prediction, suggesting that univariate deep learning models remain competitive when external factors are carefully integrated.

Beyond individual models, ensemble learning has emerged as a robust approach to enhance prediction accuracy by leveraging the strengths of multiple models. In recent research, regression-based ensemble methods were explored, integrating models such as linear regression, polynomial regression, decision tree regression, and random forest regression [9]. A stacking ensemble technique was applied to predict gold prices per ounce in U.S. dollars, resulting in a high-performing stacking regressor with a MAPE of 2.2036. This result underscores the potential of stacking ensembles to model complex patterns in gold prices, highlighting the limitations of single-model approaches in capturing intricate non-linear patterns within financial markets.

The application of stacking ensemble techniques is increasingly popular in various financial forecasting fields, such as cryptocurrency and stock price prediction, reinforcing the relevance of ensemble methods for gold price forecasting. For example, stacking ensemble techniques have shown considerable promise in predicting Bitcoin prices, which share volatility characteristics with gold. In one study on Bitcoin price prediction using LSTM and GRU within a stacking ensemble framework, researchers combined price data, technical indicators, and sentiment indexes to enhance short-term predictions [25]. The ensemble outperformed standalone LSTM and GRU models, achieving an MAE of 88.74, RMSE of 173.4, and MAPE of 0.69763%. These findings suggest that integrating various neural network architectures into an ensemble can enhance accuracy and support more reliable decision-making in volatile markets, highlighting the relevance of stacking LSTM and GRU models for assets with rapid price fluctuations.

Aligned with gold price forecasting, similar research in stock markets demonstrates the effectiveness of deep learning ensemble methods. A novel approach combining LSTM and GRU models in an ensemble for stock price prediction significantly reduced the mean squared error from 438.94 to 186.32 compared to traditional models [12]. This ensemble approach is advantageous because it captures diverse price patterns and interactions among financial indicators, which are often overlooked by single models. These results indicate the potential of ensemble models to improve accuracy in predicting not only stock prices but also other volatile assets, like gold, where capturing complex interactions is crucial.

The use of BiLSTM and BiGRU architectures is especially beneficial for capturing sequential dependencies, particularly where past price directions influence future predictions [20]. A stacking approach combining BiLSTM and BiGRU models enables more effective temporal learning in highly volatile markets, which is essential for accurate gold price forecasting. Unlike single models, stacking ensembles utilize the complementary strengths of each neural network, blending BiLSTM's bidirectional capabilities with BiGRU's efficiency for improved results. This approach enhances adaptability to changes in gold price trends, particularly in univariate data contexts where individual contributions of each network can be maximized without external features.

In practical applications, stacking BiLSTM and BiGRU models offers unique advantages in high-frequency trading or short-term forecasting, where quick and precise predictions are critical for real-time decisions. By using BiLSTM and BiGRU as base models in a stacking framework, recurring patterns in gold price data can be effectively managed. This combination utilizes BiLSTM's bidirectional ability and BiGRU's efficiency, creating an ensemble that delivers stable and accurate predictions amid market fluctuations. Additionally, incorporating a robust meta-learner, such as Random Forest, as the final layer in the stacking model refines the predictions by synthesizing base model outputs into cohesive estimates, further improving stability and accuracy in dynamic markets.

Further research highlights the benefits of ensemble learning in capturing diverse feature representations and minimizing prediction errors through model integration. Unlike traditional models that may emphasize short-term or long-term trends, stacking models capture both through a layered approach, providing greater flexibility and enhancing prediction accuracy even during unusual market conditions. This advantage is particularly beneficial for gold price forecasting, as it accommodates sudden shifts or anomalies that may not be captured by single models. Stacking BiLSTM and BiGRU models for gold price forecasting is a forward-thinking approach that combines the strengths of individual recurrent networks to address market uncertainties.

Overall, advancements in deep learning ensemble stacking methods, particularly those that integrate BiLSTM and BiGRU architectures, represent significant progress in gold price forecasting. Research in cryptocurrency, stock, and commodity markets demonstrates that stacking frameworks have the potential to deliver more accurate and reliable predictions by capturing complex dependencies in time-series data. The ability to accurately predict gold prices using BiLSTM-BiGRU stacking models offers strategic advantages, especially amid increasing global market uncertainty. Applying stacking ensembles with strong meta-learners, like Random Forest, is expected to yield optimal predictive outcomes while maintaining model stability. This research underscores the necessity of advanced, adaptive deep learning methods to address the growing complexities in gold price prediction.

2. Methodology

2.1. Dataset

The dataset utilized in this study comprises daily gold prices collected from January 1, 2020, to May 31, 2024. This data was meticulously obtained from Yahoo Finance (https://finance.yahoo.com/), a reliable source for financial market information. The primary variable analyzed in this study is the daily closing price of gold, which reflects the value of gold at the end of each trading day. To prepare the data for analysis, a thorough cleaning process was performed to remove any anomalies or inconsistencies. Furthermore, the data was normalized to ensure that all values were on a comparable scale. Finally, the dataset was segmented into two subsets: training data, representing 80% of the total, and testing data, accounting for the remaining 20%. This division allows for a robust evaluation of the predictive models developed in this research.

2.2. Long Short-Term Memory (LSTM)

LSTM was first proposed by Hochreiter and Schmidhuber in 1997 and has since evolved into a crucial method in deep learning. LSTM is an enhancement of Recurrent Neural Networks (RNNs) designed to address long-term dependency issues and the vanishing gradient problem [18]. It effectively stores past information and utilizes it in predictions without loss of data [1]. When a notable disconnect exists between the nodes responsible for processing information and the pertinent data, RNNs face difficulties in retaining essential information, resulting in the challenge referred to as "long-term dependence". Hochreiter and Bengio explored this issue in detail, revealing that it arises from vanishing gradients during the training phase of RNNs [23]. The introduction of LSTM, specifically created to address issues related to long-term dependencies, has led to numerous enhancements and expansions in later research, demonstrating its effectiveness in tasks like classification, processing, and predicting time series data.

Fig. 1 demonstrates that the LSTM model architecture comprises three main gates: the forget gate, the input gate, and the output gate. The forget gate identifies which information needs to be discarded from the memory cell, the input gate controls the incorporation of new information, and the output gate determines which information will be utilized for the subsequent output [23]. The LSTM model operates through these gates and a cell state to manage and preserve information over long sequences.

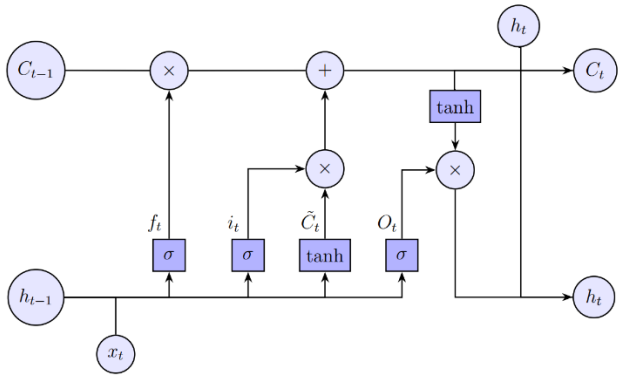

Fig. 1. The architecture of LSTM

The LSTM mechanism involves four main steps, namely Forget Gate, Input Gate, Update Cell State, and Output Gate.

1. Forget Gate: The forget gate identifies which information from the previous cell state should be discarded. This process is carried out using a sigmoid function that produces an output value ranging from 0 to 1. The formula is:
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, X_t] + b_f\right) \quad (1)$$
2. Input Gate: The input gate oversees the integration of new information into the cell state. It consists of two parts: a sigmoid layer that identifies which values need updating, and a tanh layer that produces new candidate values. The relevant formulas are:
$$i_t = \sigma(W_i \cdot [h_{t-1}, X_t] + b_i) \quad (2)$$
$$\widetilde{C}_t = tanh(W_C \cdot [h_{t-1}, X_t] + b_C) \quad (3)$$
3. Update Cell State: The cell state is modified by combining the previous cell state, adjusted by the forget gate, with the new candidate values, adjusted by the input gate. This process is represented by the formula:
$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C}_t \quad (4)$$
4. Output Gate: The output gate is responsible for determining the next hidden state, which plays a role in predictions and serves as input for the subsequent time step. A sigmoid function is used to determine which elements of the cell state should be passed to the output, while a tanh function is applied to scale the resulting output. The corresponding equations are:
$$o_t = \sigma(W_o \cdot [h_{t-1}, X_t] + b_o) \quad (5)$$
$$h_t = o_t \odot tanh(C_t) \quad (6)$$

2.3. Bidirectional LSTM (BiLSTM)

BiLSTM represents an advanced variant of conventional LSTM networks, capable of processing sequential data in both forward and backward directions [17]. This dual processing capability is especially beneficial for tasks that necessitate a comprehension of the current context by incorporating information from both past and future events. Unlike standard neural networks that predict the next output based only on previous time steps, BiLSTM uses two LSTM layers to capture data from both directions. One layer processes the sequence from start to finish, whereas the other processes it from finish to start. The outputs of these layers are combined to generate a single prediction, significantly improving performance on tasks that depend on contextual information, compared to conventional LSTM models [11]. The architecture of BiLSTM is illustrated in Fig. 2.

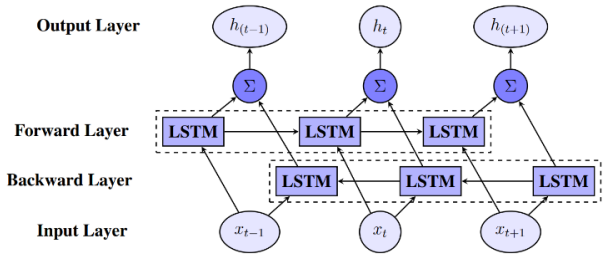

Fig. 2. The architecture of BiLSTM

In Fig. 2, X_t and h_t correspond to the input data and output data at the time t, respectively. The "Forward" part indicates the activation vector from the LSTM memory unit in the forward direction, whereas the "Backward" part represents the activation vector from the LSTM memory unit in the reverse direction [11]. The internal neuron states in BiLSTM at the time t are represented as follows:

1. Forward LSTM
$$\bar{y}_t = \overrightarrow{\text{LSTM}}\,(X_t, \bar{y}_{t-1}) \quad (7)$$
here, $\bar{y}_t$ is the hidden state at time t from the forward LSTM, X_t is the input at time t, and $\bar{y}_{t-1}$ is the hidden state from the previous time step.
2. Backward LSTM
$$\hat{y}_t = \overleftarrow{\text{LSTM}}\,(X_t, \hat{y}_{t+1}) \quad (8)$$
here, $\hat{y}_t$ is the hidden state at time t from the forward LSTM, X_t is the input at time t, and $\hat{y}_{t+1}$ is the hidden state from the previous time step.
3. Output State
$$h_t = [\bar{y}_t; \hat{y}_t] \quad (9)$$
here, h_t represents the final output state at time t, and $[;]$ denotes concatenation of the forward and backward states.

2.4. Gated Recurrent Unit (GRU)

GRU, introduced in 2014, is designed to handle sequences of varying lengths and retain past information [2]. Unlike LSTM, which uses multiple gates and memory cells, GRU simplifies the architecture by using only an update gate and a reset gate [21]. This makes GRU simpler and easier to train while achieving similar performance. GRU merges the forget and input gates of LSTM into a single update gate, which reduces the number of tensor operations and leads to faster training [21]. While GRU and LSTM perform similarly in many practical cases, GRU is computationally less complex, which makes it easier to implement.

As illustrated in Fig. 3, the GRU model operates through a series of gates to manage and preserve information over long sequences. The update gate merges the functions of LSTM's input and forget gates, determining whether the previous information needs to be updated. The reset gate, similar to LSTM's forget gate, decides whether to reset the previous information. GRU blends long-term and short-term memory by combining past information with present data. This is controlled by the reset gate and the current input, with both gates outputting values between 0 and 1 to regulate the information flow [23].

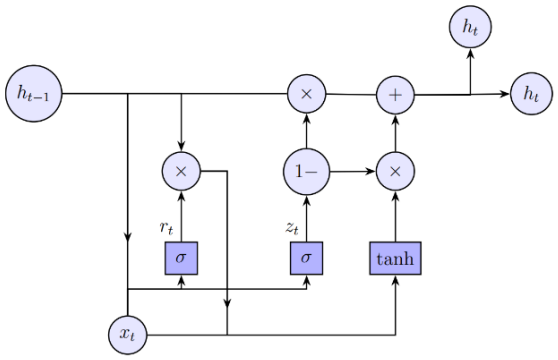

Fig. 3. The architecture of GRU

This mechanism involves four main steps, namely Reset Gate, Update Gate, Candidate Memory Vector, and Final Output.

1. Reset Gate: The reset gate regulates the impact of the previous hidden state. It determines which part of the previous hidden state to forget, using the sigmoid function to produce values between 0 and 1. The formula is:
$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \quad (10)$$
2. Update Gate: The update gate controls the amount of new input that should be added to the cell state. It dictates the degree to which previous information is retained for future use. The formula is:
$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \quad (11)$$
3. Candidate Memory Vector: The candidate memory vector is revised according to the activation of the reset gate and the current input. This process establishes the new candidate values for the cell state, employing a tanh activation function. The formula is:
$$\widetilde{h_t} = \tanh(W \cdot [r_t \odot h_{t-1}, x_t]) \quad (12)$$
4. Final Output: The final output is obtained by combining the candidate memory vector with the previous hidden state, controlled by the update gate. The formula is:
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h_t} \quad (13)$$

2.5. Bidirectional GRU (BiGRU)

BiGRU is an enhanced version of traditional GRU networks that handles sequential data in both forward and backward directions [17]. This dual processing ability is particularly useful for tasks where understanding the current context requires information from both past and future events. Unlike standard neural networks that predict the next output based only on previous time steps, BiGRU uses two GRU layers to capture data from both directions. One layer processes the sequence from beginning to end, while the other goes from end to beginning. The outputs of these layers are combined to generate a single prediction, significantly improving performance on tasks that depend on contextual information, compared to conventional GRU models [11]. The architecture of BiGRU is illustrated in Fig. 2.

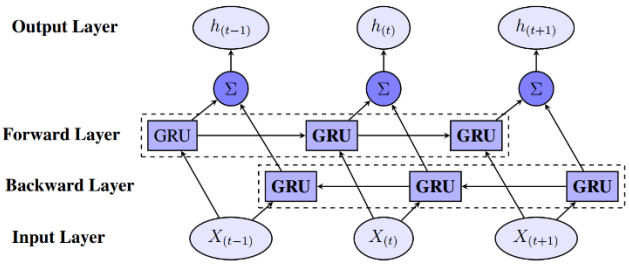

Fig. 4. The architecture of BiGRU

In Fig. 4, X_t and h_t correspond to the input and output data at time t, respectively. The "Forward" part indicates the activation vector from the GRU memory unit in the forward direction, whereas the "Backward" part represents the activation vector from the GRU memory unit in the reverse direction [11]. The internal neuron states in BiGRU at time t are represented as follows:

1. Forward GRU
$$\bar{y}_t = \overrightarrow{\mathrm{GRU}}\,(X_t, \bar{y}_{t-1}) \quad (14)$$
here, $\bar{y}_t$ is the hidden state at time t from the forward GRU, X_t is the input at time t, and $\bar{y}_{t-1}$ is the hidden state from the previous time step.
2. Backward GRU
$$\hat{y}_t = \overleftarrow{\mathrm{GRU}}\,(X_t, \hat{y}_{t+1}) \quad (15)$$
here, $\hat{y}_t$ is the hidden state at time t from the forward GRU, X_t is the input at time t, and $\hat{y}_{t+1}$ is the hidden state from the previous time step.
3. Output State
$$h_t = [\bar{y}_t; \hat{y}_t] \quad (16)$$
here, h_t represents the final output state at time t, and $[;]$ denotes concatenation of the forward and backward states.

2.6. Stacking ensemble

Ensemble learning integrates outputs from several models to generate predictions, typically improving performance by minimizing overfitting, circumventing local minima, and broadening the search space [4]. Stacking is an ensemble technique that utilizes base learners to create metadata from a problem dataset, followed by the application of a different learner, referred to as a meta-learner, to analyze this metadata [10]. The base learners serve as level 0 learners, whereas the meta-learner operates as a level 1 learner. In essence, the meta-learner is positioned above the base learners, which is the origin of the term "stacking". In a stacking ensemble, the predictions from each base learner are typically combined using a meta-learner or averaged [27]. The Stacking Ensemble model consists of two levels: the first level includes two RNNs; sub-model 1 is a BiLSTM, and sub-model 2 is a BiGRU.

The dataset is divided into two parts: training data (80%) and test data (20%). The training data is used to train the level 1 sub-models: BiLSTM and BiGRU. After the initial training phase, the trained level 1 models are used to generate predictions that will be used by the level 2 model. The test data is then used for final predictions and accuracy calculations.

First, training data is used to train sub-model 1, a BiLSTM model that consists of a single LSTM layer processing data bidirectionally, with 50 neurons in the layer. A dropout rate of 0.2 is applied, and the model is trained for 100 epochs. Next, sub-model 2 is trained, which is a BiGRU model, also with

a single GRU layer processing data in both directions, having 50 neurons, a dropout rate of 0.2, and 100 epochs. After training the BiLSTM and BiGRU models, predictions are produced using the training data for each model.

The predictions generated by BiLSTM and BiGRU are merged into a new training dataset structured as p×m (with p denoting the number of predictions and m indicating the number of models). This new dataset is used to train the meta-learner, which is a Random Forest model with 100 decision trees. Random Forest is a typical algorithm of the bagging approach that builds multiple decision trees on bootstrap samples of the data and combines their predictions, effectively reducing model variance and overfitting [28]. Integrating predictions from these sub-models using majority voting for classification or averaging for regression, Random Forest improves the robustness and accuracy of the stacking ensemble, making it well-suited for managing high-dimensional data [6]. Once the meta-learner has been trained, the test data is input into the sub-models to produce intermediate test data. This intermediate data is then utilized by the meta-learner to arrive at the final prediction.

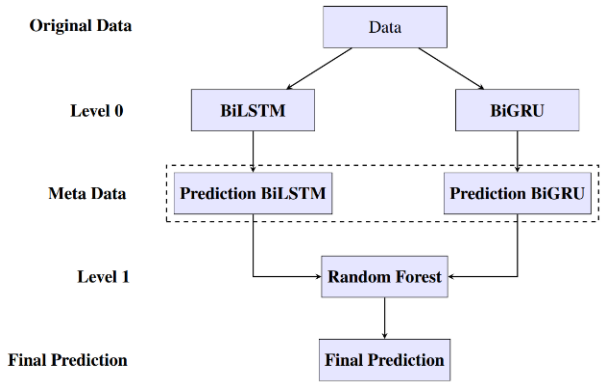

Fig. 5. The architecture of stacking BiLSTM-BiGRU

2.7. Metrics for evaluating predictive performance

In this research, several evaluation metrics were employed, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R-squared (R^2), to assess and compare the performance of the proposed models against traditional benchmarks for time series prediction [8]. These metrics provide a comprehensive view of model accuracy [16], with MSE and RMSE highlighting error magnitude, MAE indicating an average error, MAPE offering a percentage-based measure, and R^2 reflecting the model fit. Together, they allowed for a thorough evaluation of the models' accuracy and robustness

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{17}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{18}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \tag{19}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{20}$$

$$\mathrm{R}^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \tag{21}$$

3. Result and discussion

This section presents the experimental results of the proposed models based on daily gold price data gathered over a defined timeframe. The evaluation of these models' performance is conducted rigorously using several important metrics, such as MSE, RMSE, MAE, MAPE, and R2. These metrics provide a comprehensive assessment of how well the models predict daily gold prices, facilitating a detailed comparison of different models to determine their relative effectiveness. Additionally, the models were tested with various lookback periods of 7 days, 15 days, and 30 days to evaluate the impact of the lookback period on model performance, enabling a deeper understanding of how temporal context influences prediction accuracy.

The dataset was carefully split into training data (80%) and test data (20%) to guarantee a thorough validation of the model's performance. The benchmark models used for comparison include RNN, LSTM, BiLSTM, GRU, BiGRU and Stacking LSTM-GRU. All these models were trained with the following parameters to ensure a fair and consistent comparison: 1 layer, 50 neurons per layer, 100 epochs, a learning rate of 0.0005, a batch size of 32, the MSE loss function, and the Adam optimizer. This approach ensures that the training settings for each model are consistent with those used for the Stacking BiLSTM-BiGRU model, allowing for an equitable evaluation of performance across all models. Below is the comparison table of the Stacking BiLSTM-BiGRU model with the benchmark model, highlighting the results of the experimental analysis.

Table 1. Experimental result

Evaluation Metrics	Methods	Lookback Period		
		7 days	15 days	30 days
MSE	RNN	0.0009	0.0008	0.0007
	LSTM	0.0005	0.0008	0.0004
	BiLSTM	0.0011	0.0010	0.0004
	GRU	0.0005	0.0005	0.0004
	BiGRU	0.0006	0.0005	0.0003
	Stacking LSTM-GRU	0.0001	0.0001	0.0001
	Stacking BiLSTM-BiGRU	0.0000	0.0000	0.0000
RMSE	RNN	0.0301	0.0288	0.0257
	LSTM	0.0234	0.0274	0.0189
	BiLSTM	0.0331	0.0316	0.0206
	GRU	0.0213	0.0217	0.0211
	BiGRU	0.0242	0.0222	0.0169
	Stacking LSTM-GRU	0.0075	0.0075	0.0072
	Stacking BiLSTM-BiGRU	0.0069	0.0069	0.0067
MAE	RNN	0.0250	0.0227	0.0198
	LSTM	0.0174	0.0210	0.0135
	BiLSTM	0.0230	0.0237	0.0154
	GRU	0.0162	0.0168	0.0164
	BiGRU	0.0189	0.0172	0.0121
	Stacking LSTM-GRU	0.0055	0.0055	0.0054
	Stacking BiLSTM-BiGRU	0.0050	0.0050	0.0050
MAPE	RNN	0.0388	0.0348	0.0295
	LSTM	0.0293	0.0354	0.0223
	BiLSTM	0.0332	0.0363	0.0252
	GRU	0.0253	0.0267	0.0253
	BiGRU	0.0291	0.0265	0.0193
	Stacking LSTM-GRU	0.0094	0.0092	0.0091
	Stacking BiLSTM-BiGRU	0.0086	0.0083	0.0083
R^2	RNN	0.9668	0.9699	0.9765
	LSTM	0.9800	0.9727	0.9872
	BiLSTM	0.9599	0.9640	0.9848
	GRU	0.9834	0.9829	0.9841
	BiGRU	0.9785	0.9821	0.9898
	Stacking LSTM-GRU	0.9979	0.9980	0.9981
	Stacking BiLSTM-BiGRU	0.9982	0.9983	0.9984

Previous research indicates that deep learning techniques, including LSTM, BiLSTM, and GRU, are highly effective for the complex task of predicting gold prices. These models have received significant attention for their capability to capture temporal dependencies and fluctuations in gold price data, which frequently display nonlinear and intricate patterns [26]. Such temporal patterns are critical in financial data, where prices are influenced by various dynamic factors. The deep learning models, particularly LSTM and its variants, have proven adept at handling these complexities by learning from sequential data over time, making them well-suited for forecasting financial trends.

In addition to these individual deep learning models, the adoption of ensemble methods, specifically Stacking, has gained significant attention in recent research. Stacking involves combining the predictive power of multiple models to improve accuracy and robustness. Recent studies have provided compelling evidence that Stacking methods significantly outperform traditional standalone models, particularly in financial forecasting contexts like gold price prediction [9]. The unique strength of Stacking lies in its ability to leverage the complementary advantages of different algorithms, thereby enhancing model generalization and reducing overfitting.

Moreover, the superiority of ensemble approaches has been confirmed in research on financial market predictions, such as Stock Market Indices, where ensemble methods consistently outperform traditional hybrid models. These findings indicate that ensemble techniques, which combine multiple models, are better suited for capturing the complex and volatile nature of financial markets, leading to higher prediction accuracy [21]. The results from these studies strongly align with the findings of this research, where the Stacking BiLSTM-BiGRU model consistently exhibited superior performance across various lookback periods (7 days, 15 days, and 30 days), further validating the effectiveness of Stacking ensemble methods in time series prediction. These advanced techniques hold immense potential for improving predictive analytics in finance.

Table 1 shows that the Stacking BiLSTM-BiGRU model attains a remarkably low MSE value of 0.000 across all lookback periods, which underscores its exceptional capability in minimizing prediction errors in comparison to other models tested in this study. This remarkable performance not only reflects the model's sophisticated architecture but also its ability to learn complex patterns effectively. Furthermore, this model demonstrates superior performance in terms of RMSE, yielding values of 0.0069, 0.0069, and 0.0067 for the respective lookback periods. These low RMSE values indicate highly accurate predictions with minimal error magnitudes, affirming the model's reliability in delivering precise outputs.

The MAE results further bolster the model's accuracy credentials. The Stacking BiLSTM-BiGRU model records the lowest MAE value of 0.0050 across all lookback periods, showcasing its consistent prediction accuracy with minimal absolute errors throughout the dataset. For MAPE, the Stacking BiLSTM-BiGRU model continues to demonstrate outstanding performance with the lowest MAPE values of 0.0086, 0.0083, and 0.0083. This indicates highly accurate predictions relative to the scale of the data, emphasizing the model's efficacy in providing insightful forecasts.

Moreover, the high R^2 values of 0.9982, 0.9983, and 0.9984 across different lookback periods suggest that the Stacking BiLSTM-BiGRU model effectively captures a significant portion of the variability present in daily gold price data. This reflects its robustness and effectiveness in understanding and predicting the underlying patterns that govern price movements.

To offer a more comprehensive comparison of the model performance metrics, Fig. 6, Fig. 7, Fig. 8, Fig. 9, and Fig.10 present bar charts that compare the MSE, RMSE, MAE, MAPE, and R^2 values for each model across the various lookback periods. These figures vividly illustrate the superior performance of the Stacking BiLSTM-BiGRU model, clearly highlighting its advantages over other models included in this analysis. Additionally, the charts indicate that the Stacking LSTM-GRU model also performs competitively, emphasizing its strengths in specific scenarios where it excels. These visual comparisons enhance the understanding of each model's effectiveness and substantiate the quantitative results detailed in the evaluation metrics.

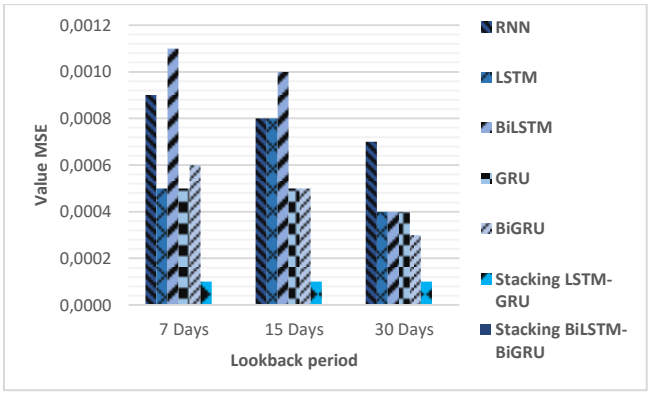

Fig. 6. MSE comparison for various models and lookback periods

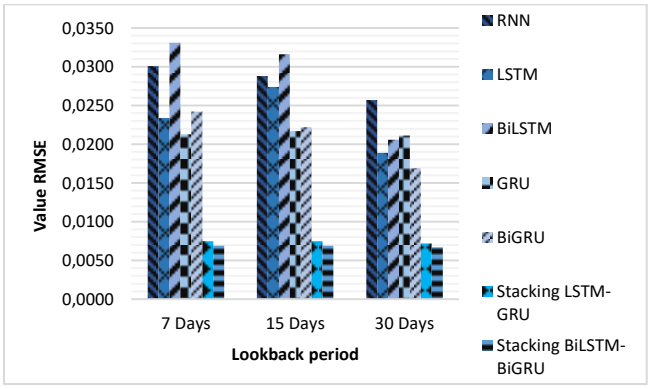

Fig. 7. RMSE comparison for various models and lookback periods

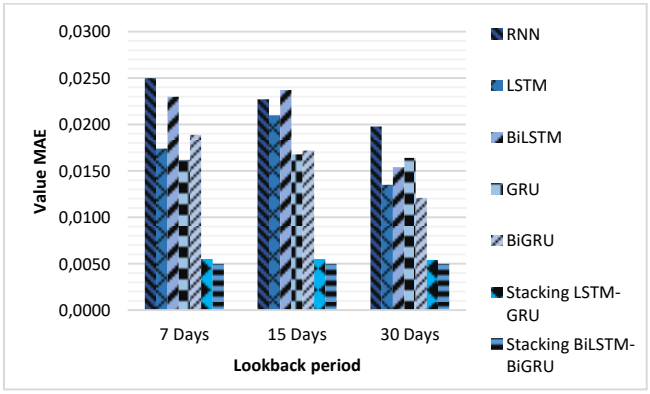

Fig. 8. MAE comparison for various models and lookback periods

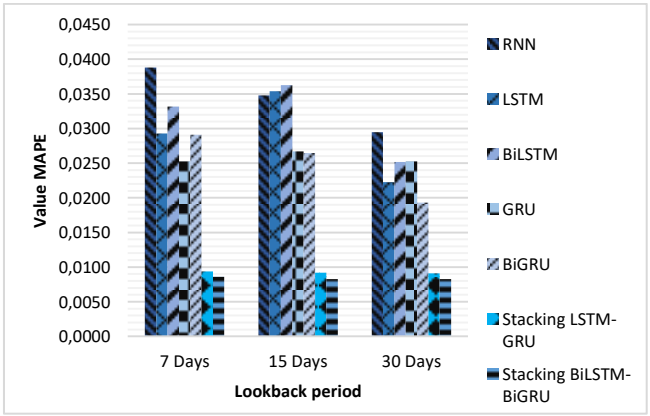

Fig. 9. MAPE comparison for various models and lookback periods

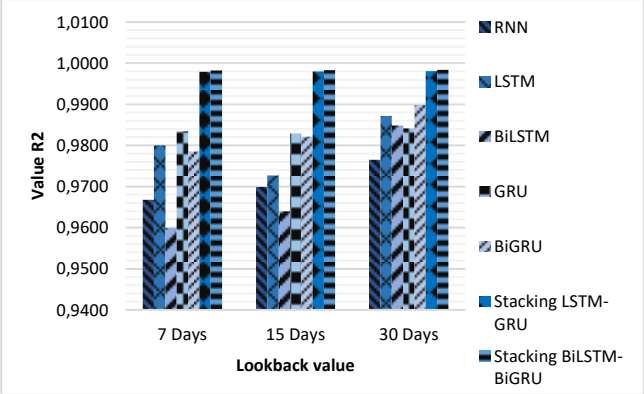

Fig. 10. R^2 comparison for various models and lookback periods

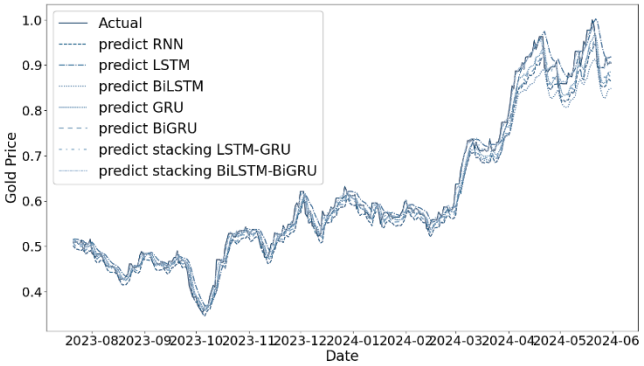

Fig. 11. Actual vs. predicted gold price with 7-day lookback period, showing model prediction accuracy

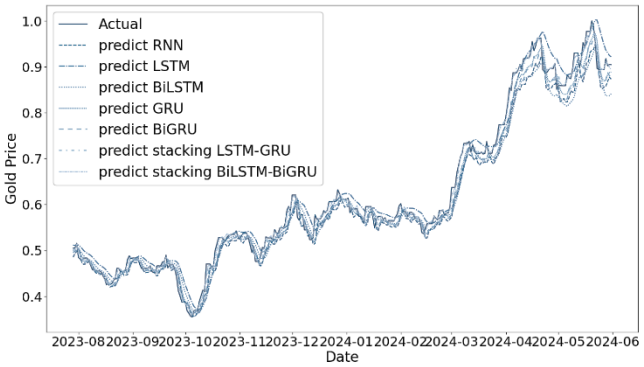

Fig. 12. Actual vs. predicted gold price with 15-day lookback period, showing model prediction accuracy

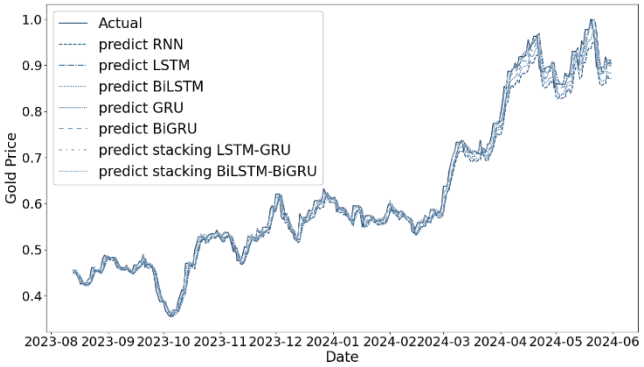

Fig. 13. Actual vs. predicted gold price with 30-day lookback period, showing model prediction accuracy

Fig. 11, Fig. 12, and Fig. 13 present a detailed and comprehensive comparison of the actual versus predicted gold prices for each model at different lookback periods of 7 days, 15 days, and 30 days, respectively. These graphs visually represent the performance of the models, demonstrating the degree to which each model corresponds to the actual gold price data over time and offering insights into their predictive abilities. This visual comparison allows for an intuitive assessment of prediction accuracy, making it easier to identify which models are more effective at tracking fluctuations in gold prices and adapting to market changes.

An increase in the lookback period correlates with a reduction in the number of available prediction data points. This reduction can slightly impact the visual representation of the plots, as having fewer data points may alter the observed trends. The diminished availability of prediction data arises from the inherent nature of lookback periods, which necessitate a longer historical dataset for generating accurate predictions. As a result, this requirement constrains the total number of forecasts that can be produced.

This phenomenon aligns with findings from previous research [21], which highlights the considerable influence of lookback period length on the volume of data that can be effectively predicted. This relationship underscores the complex interplay between historical data and prediction accuracy. As the lookback period extends, while it may provide a richer context for understanding underlying trends, it also limits the immediacy of available predictions. This balance between the depth of historical data and the recency of predictions illustrates a critical aspect of model performance, emphasizing the importance of selecting an optimal lookback period to enhance forecasting accuracy.

The graphical representations in these figures not only highlight the models' tracking abilities but also provide deeper insights into their overall effectiveness in different scenarios. By visually contrasting the predicted values against the actual values, it becomes evident which models consistently outperform others, showcasing their reliability. Furthermore, these comparisons reinforce the quantitative results presented in the evaluation metrics, underscoring the importance of selecting appropriate lookback periods for enhancing prediction accuracy. Overall, these visualizations serve as a valuable tool for understanding the strengths and weaknesses of each model in the context of gold price prediction, facilitating more informed decision-making based on visual data analysis.

4. Conclusion

The Stacking BiLSTM-BiGRU model has demonstrated exceptional performance in predicting gold prices, particularly when utilizing a 30-day lookback period. This model not only outperforms all other models in terms of critical metrics such as MSE, RMSE, MAE, MAPE, and R^2 values, but it also sets a new standard for accuracy in this domain. Its outstanding accuracy and stability underscore its effectiveness in the intricate and volatile market of gold price prediction, making it a highly reliable tool for analysts and investors alike. The results of this study emphasize the substantial benefit of using a stacking ensemble method that integrates BiLSTM and BiGRU, paired with a strong meta-learner like Random Forest, to attain high predictive accuracy in this setting. Moreover, future research may delve into further optimizations of this model, exploring various hyperparameter settings and architectures. It may also investigate the applicability of the Stacking BiLSTM-BiGRU model to different climatic regions, thereby validating its effectiveness in diverse contexts and enhancing its generalizability across various market conditions.

References

[1] Brownlee J.: Long Short-Term Memory Networks With Python: Develop Sequence Prediction Models with Deep Learning. Machine Learning Mastery. Jason Brownlee, 2017.

[2] Chung J., et al.: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. ArXiv Preprint ArXiv:1412.3555, 2014, 1–9.

[3] Cohen G., Aiche A.: Forecasting Gold Price Using Machine Learning Methodologies. Chaos, Solitons and Fractals 175(P2), 2023, 114079 [https://doi.org/10.1016/j.chaos.2023.114079].

[4] Dang Y., et al.: A Comparative Study of Non-Deep Learning, Deep Learning, and Ensemble Learning Methods for Sunspot Number Prediction. Applied Artificial Intelligence 36(1), 2022 [https://doi.org/10.1080/08839514.2022.2074129].

[5] Gbadamosi S. L., et al.: Exploring the Effectiveness of a Multilayer Neural Network Model for Gold Price Prediction. Przegląd Elektrotechniczny 3, 2024, 157–161 [https://doi.org/10.15199/48.2024.03.28].

[6] Genuer R., Poggi J. M.: Random Forests with R. Springer Cham 2020 [https://doi.org/10.1007/978-3-030-56485-8].

[7] Gu J., et al.: A Stacking Ensemble Learning Model for Monthly Rainfall Prediction in the Taihu Basin, China. Water 14(3), 2022, 1–20 [https://doi.org/10.3390/w14030492].

[8] Hyndman R. J., Athanasopoulos G.: Forecasting: Principles and Practice. OTexts 2018.

[9] Kilimci Z. H.: Ensemble Regression-Based Gold Price (XAU/USD) Prediction. Journal of Emerging Market Finance 2(1), 2022, 7–12.

[10] Kyriakides G., Margaritis K. G.: Hands-On Ensemble Learning with Python. Build Highly Optimized Ensemble Machine Learning Models Using Scikit-Learn and Keras. Packt 2019.

[11] Li X., et al.: Resource Usage Prediction Based on BILSTM-GRU Combination Model. IEEE International Conference on Joint Cloud Computing (JCC). 2022, 9–16 [https://doi.org/10.1109/JCC56315.2022.00009].

[12] Li Y., Pan Y.: A Novel Ensemble Deep Learning Model for Stock Prediction Based on Stock Prices and News. International Journal of Data Science and Analytics 13(2), 2022, 139–49 [https://doi.org/10.1007/s41060-021-00279-9].

[13] Livieris I. E., et al.: A CNN–LSTM Model for Gold Price Time-Series Forecasting. Neural Computing and Applications 32(23), 2020, 17351–17360 [https://doi.org/10.1007/s00521-020-04867-x].

[14] Mulaab et al.: Air Temperature Forecasting Based on Stacking Machine Learning Model with Multi-Step Time Series. IEEE 9th Information Technology International Seminar (ITIS). 2023, 1–4 [https://doi.org/10.1109/ITIS59651.2023.10420402].

[15] Nwokike C., et al.: Forecasting Monthly Prices of Gold Using Artificial Neural Network. Journal of Statistical and Econometric Methods 9(3), 2020, 19–28.

[16] Rivas-Perea P., et al.: Support Vector Machines for Regression: A Succinct Review of Large-Scale and Linear Programming Formulations. International Journal of Intelligence Science 3(1), 2013, 5–14 [https://doi.org/10.4236/ijis.2013.31002].

[17] Schuster M., Paliwal K. K.: Bidirectional Recurrent Neural Networks. IEEE Transactions on Signal Processing 45(11), 1997, 2673–81 [https://doi.org/10.1109/78.650093].

[18] Seabe Phumudzo L., et al.: Forecasting Cryptocurrency Prices Using LSTM, GRU, and Bi-Directional LSTM: A Deep Learning Approach. Fractal and Fractional 7(2), 2023, 203 [https://doi.org/10.3390/fractalfract7020203].

[19] Sezer O. B., et al.: Financial Time Series Forecasting with Deep Learning: A Systematic Literature Review: 2005–2019. Applied Soft Computing Journal 90, 2020, 2005–2019 [https://doi.org/10.1016/j.asoc.2020.106181].

[20] Shaikh Z. M., Ramadass S.: Unveiling Deep Learning Powers: LSTM, BiLSTM, GRU, BiGRU, RNN Comparison. Indonesian Journal of Electrical Engineering and Computer Science 35(1), 2024, 263–273 [https://doi.org/10.11591/ijeecs.v35.i1.pp263-273].

[21] Song H., Choi H.: Forecasting Stock Market Indices Using the Recurrent Neural Network Based Hybrid Models: CNN-LSTM, GRU-CNN, and Ensemble Models. Applied Sciences 13(7), 2023, 4644 [https://doi.org/10.3390/app13074644].

[22] Wesley G. Y. Y., Sufahani S.: Research Review on Time Series Forecasting of Gold Price Movement. International Journal of Multidisciplinary Research and Development 5(5), 2018, 44–49.

[23] Xu L., et al.: Deep Heuristic Evolutionary Regression Model Based on the Fusion of BiGRU and BiLSTM. Cognitive Computation 15(5), 2023, 1672–1686 [https://doi.org/10.1007/s12559-023-10135-6].

[24] Yang X.: The Prediction of Gold Price Using ARIMA Model. 2nd International Conference on Social Science, Public Health and Education – SSPHE 2018. 2019, 273–276 [https://doi.org/10.2991/ssphe-18.2019.66].

[25] Ye Z., et al.: A Stacking Ensemble Deep Learning Model for Bitcoin Price Prediction Using Twitter Comments on Bitcoin. Mathematics 10(8), 2022, 1307 [https://doi.org/10.3390/math10081307].

[26] Yurtsever M.: Gold Price Forecasting Using LSTM, Bi-LSTM and GRU. Avrupa Bilim ve Teknoloji Dergisi 31(31), 2021, 341–347 [https://doi.org/10.31590/ejosat.959405].

[27] Zhang C., Sjarif N. N. A., Ibrahim R: Deep Learning Models for Price Forecasting of Financial Time Series: A Review of Recent Advancements: 2020–2022. WilREs Data Mining and Knowledge Discovery 14(1), 2024, e1519 [https://doi.org/10.1002/widm.1519].

[28] Zhao Y., Deng W.: Prediction in Traffic Accident Duration Based on Heterogeneous Ensemble Learning. Applied Artificial Intelligence 36(1), 2022 [https://doi.org/10.1080/08839514.2021.2018643].

M.Sc. Iqbal Kharisudin
e-mail: iqbalkharisudin@mail.unnes.ac.id

He is a lecturer in the statistics and data science study program at Universitas Negeri Semarang. He teaches time series analysis, multivariate statistics, and machine learning. His research interests are related to statistical modeling and simulation including time series forecasting, multivariate analysis, and spatial statistics. Some research involves deep learning models for image classification and text mining. He is also interested in mathematics education and modeling skills in problem-solving.

https://orcid.org/0000-0002-1156-4974

Nike Yustina Oktaviani
e-mail: nikeyustina12.o@gmail.com

She is a graduate of Universitas Negeri Semarang, majoring in Mathematics at the Faculty of Mathematics and Natural Sciences. She completed her undergraduate thesis titled "Stacking Ensemble Modeling with Bidirectional LSTM and Bidirectional GRU for Air Temperature Prediction in Ngawi Regency". Her research interests include machine learning, deep learning, and their applications in data prediction. Through her work, Nike aspires to contribute to educational advancements and provide meaningful benefits to the broader community.

https://orcid.org/0009-0003-9209-1105