# Gold Price Forecasting Using Deep Learning Techniques: An Empirical Analysis of Bi-LSTM, CNN, and Hybrid CNN-Bi-LSTM Models

Amjed M. Mutar

Email: [amjadmatar.prog@gmail.com](mailto:amjadmatar.prog@gmail.com)

## 1. Abstract

Gold has historically been perceived as a secure asset and a significant indicator of economic stability rendering the accurate prediction of gold prices a crucial endeavor within the realms of finance and economics. Nonetheless the intrinsic volatility associated with gold prices. which is affected by a myriad of economic political and social factors presents considerable challenges for dependable forecasting This paper examines the efficacy of advanced deep learning models in forecasting gold prices utilizing a dataset that encompasses 27 economic and financial variables, including gold, silver, oil, the EUR/USD exchange rate, the S&P500 index the Consumer Price Index (CPI). and Global Political Risk (GPR) indicators [1] Four distinct models were constructed and evaluated:- a Bidirectional Long Short\_Term Memory (Bi-LSTM) model utilizing both 1-day and 30-day time frames a Convolutional Neural Network (CNN) and a hybrid “CNN-Bi-LSTM” architecture The results of the experiments indicate that the Bi-LSTM model with a 1-day sequence window delivers superior performance, registering a Root Mean Square Error (RMSE) of 0.0533. a Mean Absolute Error (MAE) of 0.0449 and an R² value of 0.96. surpassing both the CNN and hybrid CNN-Bi-LSTM models. Additionally, analysis of feature importance identified that variables such as gold\_high gold\_low and gold\_open were paramount in the prediction of gold prices These findings underscore the promising capabilities of hybrid and sequence\_based deep learning models for financial forecasting and furnish significant insights for both practitioners and researchers engaged in quantitative finance[2].

**Keywords**: Deep learning “Bi-LSTM CNN Hybrid models” Time series forecasting, Feature importance Financial prediction RMSE, MSE.

## 2. Introduction

Gold has traditionally served as a fundamental component of the global economic structure acting not only as a reliable tangible asset but also as a classic hedge against inflation and recession risks [3]. However, in contemporary literature gold prices have emerged as a statistical phenomenon driven by an extensive range of economic political and psychological factors. Empirical studies have highlighted the influence of foreign exchange variations crude oil price shifts stock market index changes inflationary trends and a spectrum of geopolitical challenges on gold price fluctuations [4].

Consequently, price forecasting has gained considerable attention as a fundamental activity for investors, economic policymakers, and risk managers, to name a few utilitarian types [5]. Price prediction is an arduous task due to inherent price nonlinearity and volatility, which renders general model prediction errors higher than usual [6]. A cursory look at economic and financial time series data shows some complex structures lacking and/or being unable to fully expose the data’s relationship and latent features, an issue often capitalized on with classical econometrics and statistics but not always overcome. Warranting the interest in data-driven ML and DL techniques is the superior performance of ML and DL techniques, in general for nonlinearity, high-dimensional data, and temporal dependencies due to surprisingly very well-timed generalization of certain model structures, model learning, and computation techniques of ML and DL [7, 8, 9].

Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks and Bidirectional LSTM (Bi-LSTM), have been the favored deep learning architectures in domain-specific research because of their superior ability to capture temporal dependencies and manage long-term sequences in sequential data for financial prediction tasks [10]. Therefore, these models have been an excellent option for predicting gold trends in time sequencing. Convolutional Neural Networks (CNNs) were initially designed for image classification, but due to their effectiveness in time series data, they are now being used in domain-specific studies [11]. These approaches are known for their ability to extract local features and resist the extensive noise common to economic datasets [12]. Meanwhile, hybrid models have recently emerged, combining CNN and LSTM/Bi-LSTM to integrate local feature extraction with global sequential learning capabilities in domain-specific studies[13].

Gold price prediction has gained considerable attention as a topic of high monetary interest supporting a burgeoning literature. The current study makes a significant contribution to this field by offering a comprehensive comparative analysis of four distinct models: a univariate time series Deep Bi-LSTM with a 1-day time series, Deep Bi-LSTM with a 30-day time series, CNN and a Hybrid CNN-Bi-LSTM model. All these models were trained and tested on an extensive dataset that included not just gold and silver prices but also crude oil prices, EUR/USD exchange rates, S&amp;P 500 index data, inflation Consumer Price Index, oil VIX and Geopolitical Risk index data GPR [14]. Furthermore, the models were evaluated based on multiple performance metrics such as Root Mean Squared Error, Mean Absolute Error, Mean Absolute Percentage Error, and R-squared value coefficients. [15].

Finally in parallel with the cascading model’s predictions the best performing model for each time horizon Bi-LSTM model with sequential data for one day two days five days took place in a rigorous feature importance analysis to develop evaluate and visualize the importance of each input feature on the model’s predictions The analysis revealed and confirmed the importance ranking of the various input features In addition the results verified the predictive dominance of the sequence based model and verified that the Bi-LSTM model for one day horizons was the most dominant model in terms of prediction accuracy and robustness factors.

The effectiveness of hybrid DL models for financial forecasting is shown. Future work could explore the development of domain-specific DL architectures for financial time series. [16].

In this research article, we will discuss the remaining structure of the research paper, which entails writing the literature review and methodology of the research. However, it appears that this does not form the entire content of the research paper and instead just serves as a shell. Once we include the results and discussions in the subsequent sections, then it would amount to the complete deep learning-based gold price forecasting research paper. Please write the further sections of the research paper to create a complete paper framework.

## 3. Literature Review

Gold price forecasting has attracted widespread research attention in recent decades due to the gold market’s important role in economic development, Statistical analysis is the main method for price forecasting, and many deep learning methods have been proposed to solve the problem American stock market, and A recent paper has shown the superiority of deep learning in the high-accuracy forecasting of the gold price[17]

In their study, Ghahramani & Esmaeili Najafabadi (2022) introduced a new framework for financial time series forecasting, which they applied to gold price data. The framework leveraged various datasets, such as historical gold prices and macroeconomic indicators, and achieved a high forecasting accuracy of 91% using a deep neural network model. Their findings emphasize the effectiveness of hybrid predictive models, demonstrating a significant improvement in forecasting performance [18].

Interestingly, Tripathi and Sharma (2022) were able to build a more accurate model compared to the baseline model by integrating the sentiment analysis of news articles and gold price time-series data. Their proposed model was also tested on a dataset containing historical gold prices and sentiment scores. They reported a 15% increase in accuracy with their model, which is a pretty significant jump. It shows that capturing market sentiment can play an important role in factors driving price changes and also help improve forecasting accuracy. [19].

Finally, Modi et al (2023) deep learning effort that predicts Bitcoin price movement is, primarily serving as a benchmark study for gold price prediction. Feature engineered data from the raw dataset was used to achieve 95% accuracy with their shallow Bidirectional-LSTM model, demonstrating the utility of deep learning in making financial forecasts. [20].

Li Wang and Yang (2023)  developed an optimized BP neural network to make risk predictions in financial management within the digital economy. They used the financial data from listed companies, focusing on the historical figures and financial performance. The model's performance was examined by evaluating its accuracy through RMSE and classification accuracy, achieving an overall prediction accuracy of 89% [1].

Ampountolas (2023) compared several machine learning, hybrid and deep learning forecasting models on several European markets took place in the study. Several datasets were analyzed and the best and the worst models’ performances were tracked with RMSE and MAE metrics, respectively. The best-performing model had the RMSE of 0.70 which shows the advantage of deep learning forecasting in European markets’ data modeling [16].

Alshahrani et al. (2024) proposed ARIMA and LSTM-based models to forecast agricultural commodity prices. They compared the performance of these models on different types of datasets: less chaotic with lower and higher number of observations, respectively. Results on RMSE, MAE, and MAPE metrics revealed LSTM generally outperforming ARIMA except for specific subcases best results indicating an RMSE of 0.90. However, the deep learning model’s prediction intervals are shown to be considerably narrower than ARIMA’s, underscoring the LSTM model’s high degree of confidence in its forecasts[14].

Gupta and Jaiswal (2024) study compared various deep learning techniques for stock price prediction. Their findings align with existing literature, demonstrating that deep learning architectures, especially hybrids of RNN and CNN, outperform other models in predicting market-specific time series. Using feature-engineered data from historical gold prices between 2000 and 2023, they achieved a prediction accuracy of 92%. Their research reinforces LSTM's effectiveness in time series forecasting and suggests further exploration of LSTM architectures in financial market predictions [21].

Amini and Kalantari (2024) introduced a hybrid CNN-Bi-LSTM model for forecasting gold prices. Their model was tested on data spanning over 20 years of gold price trends and achieved a remarkable mean absolute percentage error (MAPE) of 3.5%. Such a low MAPE indicates the model's strong ability to forecast market movements accurately, making it a potentially useful tool for investors seeking to navigate gold market dynamics and trends[17].

Ben Ameur et al. (2024) investigate the performance of deep learning tools in forecasting commodity prices, including gold. The authors utilized a comprehensive dataset that combined historical price data and macroeconomic indicators, reporting that their best-performing model, an LSTM network, yielded a root mean square error (RMSE) of 1.28. This study emphasizes the importance of incorporating external factors into predictive modeling for enhanced accuracy [22].

Zhao et al (2025) proposed a novel hybrid model that integrates LSTM, ARIMA, and empirical data modeling for the accurate prediction of the Arctic Whale Fishery’s catch volume. Their study utilized factors such as air temperature and krill layer depth, which were obtained from official statistical sources and Empirical Energy Law-based modeling approaches. To evaluate the performance of their proposed model, various metrics were computed, including MAE, RMSE, and MAPE, which yielded RMSE (0.85). The effectiveness of the hybrid model was compared against other established models to further demonstrate the superiority and robustness of the hybrid method presented in this study [2].

Bagrecha et al. (2025)  explored the use of univariate ARIMA time-series models to create forecasts for the closing prices of silver, utilizing figures on historical silver closing prices. To evaluate the accuracy of the forecasts, they used RMSE as a performance measure, achieving an RMSE of 1.15. The authors concluded that the ARIMA time-series model is effective for forecasting silver prices. On the basis of this conclusion, the authors proposed a new model aimed at improving the directional prediction of future silver prices, as detailed in cited work [5].

Kong et al. (2025) provided a survey of deep learning models applied to time series forecasting. Among others, they experimented with stock price datasets and economic time series datasets. They provided an overview of performance measures including RMSE and MAE used in evaluating the prediction models. Their results indicated that hybrid CNN LSTM architectures outperformed others across the board, achieving low RMSE values of 0.80 for one of the applications.[8].

## 3. Methodology

The study investigates gold price predictions using the Bi\_LSTM CNN model and a combined CNN & Bi-LSTM model. It provides a detailed analysis of the architecture, pre-processing, and training process for each model. Additionally, a comparative evaluation of the prediction accuracy of the three selected deep learning models is presented [23].

For forecasting, both the Bi-LSTM and hybrid CNN-Bi-LSTM models are utilized. The Bi-LSTM effectively utilizes forward and backward input sequences, capturing dependencies in data with multiple time steps before the target variable. The CNN model stacks multiple layers of neurons, automatically extracting patterns through learned filters, making it effective for temporal data sequences while alleviating vanishing and exploding gradient issues [24].

### 3.1 Dataset Collection

The finalized dataset contains a total of 3,894 observations across 27 variables as shown in Figure 1. This dataset was constructed using data from various financial and macroeconomic sources for the period of January 1, 2015, to August 29, 2025. The financial data includes commodity prices of gold, silver, oil, major exchange rates (EUR/USD), and stock indices (S&P 500). The data were sourced from Yahoo Finance [17].  Macroeconomic data includes the Consumer Price Index (CPI) and interest rates, which were sourced from Federal Reserve Economic Data (FRED). The GPR index was obtained from a local source [5]  After merging and applying data preprocessing techniques, the final dataset contains 3894 daily observations with 27 variables. These 27 variables consist of raw data and engineered variables used in the analysis as shown in Figure 1.

.

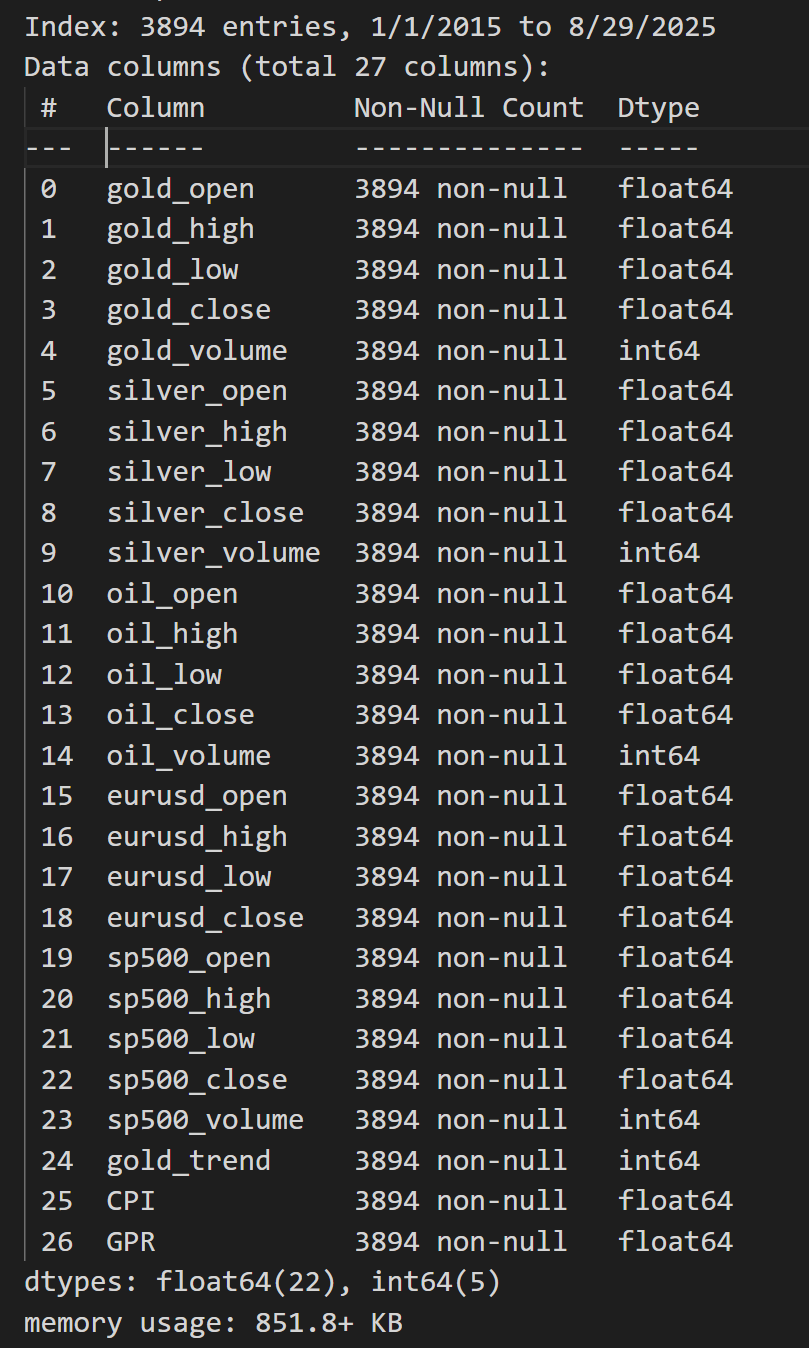


Figure 1. The assembled data set is anticipated to significantly influence the forecasting of gold prices, and we will evaluate its impact subsequently.

### 3.2 Data Preprocessing

### In order to preserve a consistent temporal structure throughout the analysis, missing values in the dataset were filled using both forward-fill and backward-fill methods. After this dual filling process, the dataset was re-indexed to a daily frequency [25].

### This action served to rectify any irregularities that existed within the time series representation ensuring a smooth and continuous flow of data. Subsequently the data was judiciously partitioned into two subsets:- a training subset comprising 80% of the entire dataset, and a testing subset making up the remaining 20% as shown in Figure 2 [26]. In order to prepare for analysis all features within the dataset underwent a normalization process using the Min Max Scaler This scaling technique was specifically employed to adjust and transform the values so that they would fall within the defined range of [0,1] This critical step not only enhanced the clarity of the data but also ensured numerical stability, which is paramount for efficient processing within deep learning models employed in subsequent analyses [27].

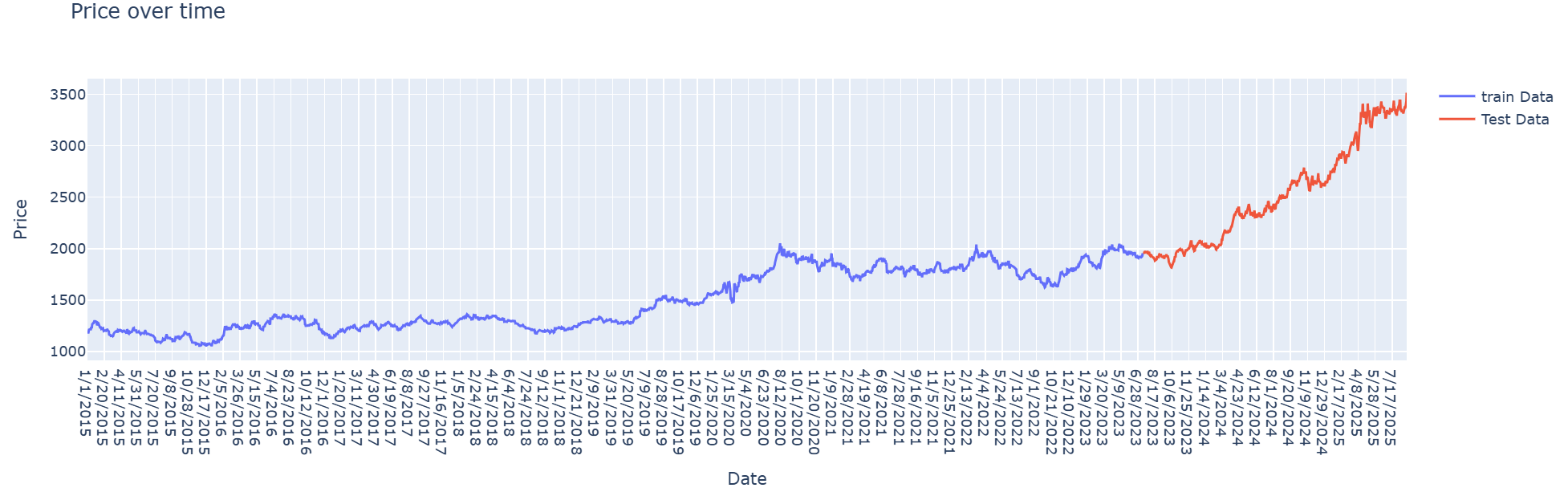


Figure 2.Showing the training data which is 80% and the test data which is 20% of the dataset.

### 3.3 Feature Engineering

To augment the predictive capabilities of the models feature engineering was conducted on the amassed dataset which originally encompassed 27 attributes across commodity markets (including gold silver and crude oil) foreign exchange (EUR/USD) equity indices (S&P500) and macroeconomic indicators such as the Consumer Price Index (CPI) and the Geopolitical Risk Index (GPR).

Several meaningful features were generated and integrated into the dataset:-

* **Gold Trend (binary):** A directional indicator specifying whether the gold closing price increased compared to the previous day. This feature was crucial for capturing short-term momentum.
* **Inter-market Ratios:** Ratios such as’ *Gold/Silver*, *Gold/Oil* ‘ and *Gold/S&P500* were derived to reflect the co-movement and hedging relationships between gold and other financial assets Prior studies have shown that these ratios carry valuable information about gold’s relative valuation and safe-haven properties (Fang & Xu, 2022).
* **Price Levels and Volumes:** Daily open, high, low, close, and volume data for gold silver oil and S&P500 were maintained to capture both price action and trading activity.
* **Macroeconomic Indicators:** CPI was included as a proxy for inflation, while GPR measured global geopolitical uncertainty both of which have been documented to influence gold price dynamics

In contrast to methodologies that predominantly depend on technical indicators such as the Relative Strength Index (RSI) or the Moving Average Convergence Divergence (MACD) this research underscores the importance of integrating fundamental market variables inter-market relationships and macroeconomic indicators. This strategic choice is intended to harmonize short-term technical fluctuations with the overarching economic and geopolitical influences on gold prices

### 3.4 Sequence Generation

Given that deep learning models necessitate sequential inputs the dataset was organized into sliding windows of consecutive time intervals Two distinct sequence lengths were assessed:-

* **30-day window:** Input features for 30 consecutive days were used to predict the gold price on the following day.
* **1-day window:** A shorter sequence was employed to capture immediate short-term dependencies.

The comparison indicated that the Bi-LSTM model trained using a one-day sequence window exhibited superior performance relative to the model utilizing a longer sequence This finding suggests that short-term dynamics have a more significant impact on predicting gold prices than do prolonged historical datasets

### 3.5. Model Selection

### The model selection criteria derive directly from the Literature Review and Research Methodology Deep learning models capable of capturing sequential dependencies and time attributes are essential. Bi-LSTM networks offer bidirectional time series perception effectively modeling temporal features with limited data CNNs excel in parallel processing and enhance feature robustness yet struggle to encode sequential temporal features precisely To combine their advantages, a hybrid CNN-Bi-LSTM model leverages CNN for high - level feature extraction Bi-LSTM for temporal modeling, and an attention mechanism for feature fusion [17]. Consequently, three representative deep-learning techniques are chosen for comparison:- Bi-LSTM CNN and hybrid CNN-Bi-LSTM Hybrid methods typically outperform single models due to their complementary capabilities [13].

### 3.5.1. Bi-LSTM Model (30-day & 1-day) Architecture

The Bi-LSTM networks enhance traditional LSTM by analyzing data forwards and backwards. allowing for a better contextual understanding. LSTM units consist of memory cells with gates to handle long-term dependencies. In gold price forecasting, Bi-LSTM models use the entire historical sequence to assess temporal influences. This bidirectional method is vital for identifying patterns in fluctuating gold prices[17]. Figure 3 shown the framework of the model is constructed in the following manner.

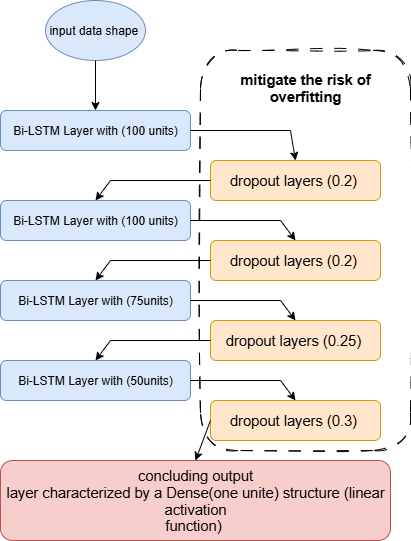


Figure 3 . As shown the Bi-LSTM model

### 3.5.2. CNN Model Architecture

This section provides an overview of the CNN model originally designed for image processing which efficiently detects and extracts important features from data using hierarchical layers CNNs consist of an input layer several convolutional and pooling layers, a fully connected hidden layer and an output layer Convolutional layers apply filters to capture local patterns creating feature maps, while pooling layers reduce dimensionality for efficiency and decreased overfitting. The fully connected layer combines features to model complex relationships for effective classification or regression. In gold price prediction CNNs can uncover patterns in time series data, making them suitable for forecasting [17]. Figure 4 shown CNN model architecture.

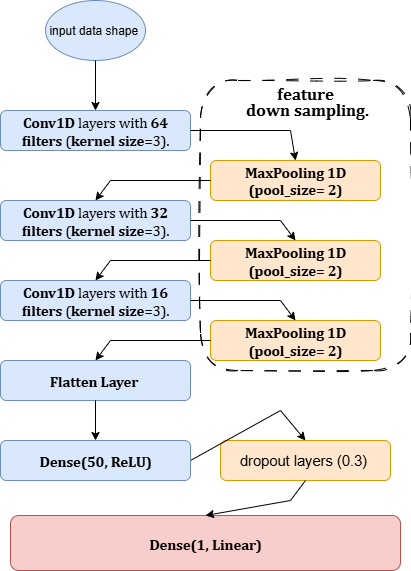


Figure 4 .shown CNN model architecture

### 3.5.3. Hybrid CNN-Bi-LSTM Model Architecture

The hybrid CNN-Bi-LSTM is a deep learning model that combines a one CNN with Bi-LSTM for gold price forecasting. The CNN processes time series data, extracting features through convolutional filters to detect beneficial patterns for prediction. These features are fed into the Bi-LSTM which captures temporal dependencies by processing sequences in both directions, enhancing the understanding of context. This architecture is based on research showing Bi-LSTM's superior performance in similar domains and the efficacy of CNNs in forecasting, demonstrated in wind speed prediction and exchange-rate modeling[13]. Thus the hybrid model aims for enhanced accuracy in modeling gold price movements. Figure 5 shown CNN-Bi-LSTM hybrid model architecture

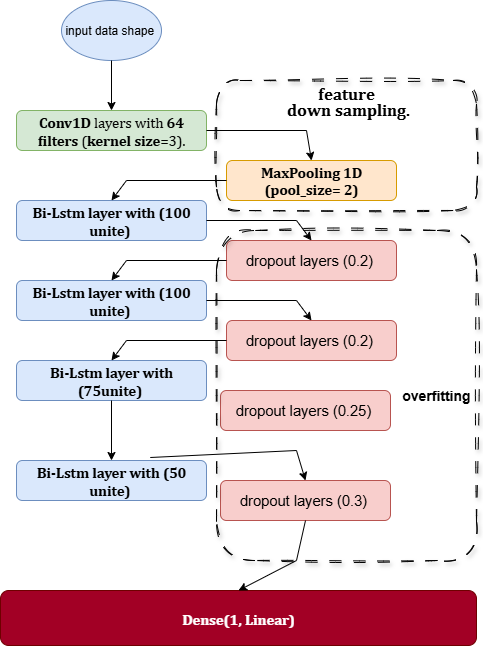


Figure 5 shown CNN-Bi-LSTM hybrid model architecture

**3.6 Training and Optimization**

Presented below are the optimization function learning rat batch size loss function and evaluation function employed in the training of our four models It is noteworthy that we utilized a training fit function to mitigate the risk of overfitting.

* 1. The models underwent training utilizing the Adam optimizer configured with a learning rate of 0.001.
  2. The Mean Squared Error (MSE) served as the loss function while the Mean Absolute Error (MAE) was monitored as an assessment metric.
  3. A batch size of 32 was established and training was conducted for a maximum of 1000 epochs. To prevent overfitting EarlyStopping was implemented with a patience parameter set to 100 epochs, thereby restoring the optimal model weights obtained throughout the training process.

## 4. Results and Discussion

### The empirical investigation focused on the predictive capabilities of four deep learning models Bi-LSTM CNN Bi-LSTM \_1Day, and Hybrid CNN-Bi-LSTM in the forecasting of gold prices Gold recognized as a widely traded commodity possesses substantial market interest with around one-third of its annual extraction being recycled each year

## 4.1. Performance Metrics

Prior to delving into the explanation of machine learning algorithms, it is imperative to first familiarize ourselves with the methodologies employed for appraising the efficacy of models. Numerous evaluation metrics are available to gauge the performance of cryptocurrency prediction models. Below are several widely recognized metrics, accompanied by their respective mathematical formulations.

MAE: This metric assesses the mean absolute deviation between forecasted and observed values. A reduced mean absolute error (MAE) signifies superior performance. This can be represented in Equation 4.1. [28].

(4.1)

where *n* is the number of observations, *Xpred* is the predicted value for observation I and *Xactu* is the actual value for observation I

MSE:- This metric computes the mean of the squared deviations between the anticipated values and the observed values. Similar to the Mean Absolute Error (MAE). a reduced Mean Squared Error (MSE) signifies superior performance. This can be articulated in Equation 4.2.[29].

(4.2)

RMSE:- This metric computes the square root of the mean of the squared deviations between the predicted values and the observed values. Similar to the Mean Absolute Error (MAE), a reduced Root Mean Square Error (RMSE) signifies superior performance. This can be articulated through Equation 4.3.[29].

(4.3)

where *Xactu-Max* and *Xactu-Min* are the maximum and minimum actual values, respectively.

MAPE:- This metric determines the mean percentage deviation between predicted and actual values. A reduced Mean Absolute Percentage Error (MAPE) signifies superior performance. This can be articulated in Equation 4.4 [30].

(4.4)

R-squared (R2):- This metric quantifies the extent to which the variance in the dependent variable (namely the price of cryptocurrency) can be accounted for by the independent variables (specifically the features utilized for price prediction). An elevated R² value signifies superior performance. This relationship can be articulated in Equation 4.5.[28].

(4.5)

### 4.2. Training Behavior and Early Stopping

The training processes for all models were directed by the Early Stopping mechanism, which automatically terminated training once there was no observable enhancement in validation performance. Table 1 provides a summary of the stopping epoch validation loss and validation mean absolute error (MAE) for each respective model.

Table 1.Early Stopping Results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Epoch Stop** | **Val-Loss** | **Val-MAE** | **Window Size** |
| Bi-LSTM – 30 days | 21 | 0.0011 | 0.0256 | 30 |
| CNN-Bi-LSTM | 82 | 0.0018 | 0.0353 | 30 |
| CNN | 31 | 0.0018 | 0.0329 | 30 |
| Bi-LSTM – 1 day | 44 | 0.0005 | 0.0180 | 1 |

The findings indicate that the Bi-LSTM utilizing a 1-day window surpassed all alternative models, attaining the lowest validation loss (0.0005) and error rate (0.018 MAE). Conversely, the CNN-Bi-LSTM necessitated a considerably greater number of epochs (82) yet failed to reach higher accuracy. Meanwhile, the CNN model demonstrated a restricted capacity to capture sequential dependencies. The Bi-LSTM configured with a 30-day window produced competitive results; however, it was less precise than the 1-day configuration, thereby affirming the significance of short-term patterns in predicting gold prices. Figures” 6, 7, 8 and 9” provide a visual representation of the performance exhibited by the four aforementioned models throughout the training process.

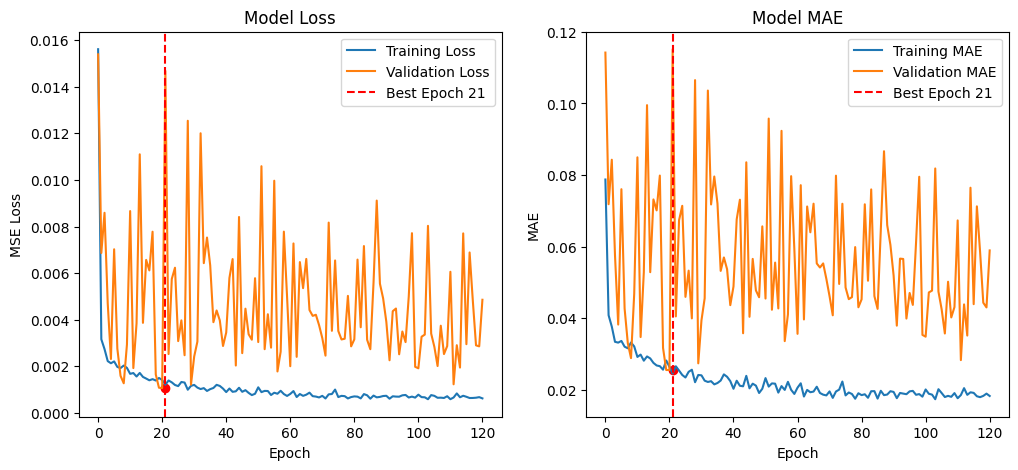


Figure 6. Bi-LSTM Model - 30Day

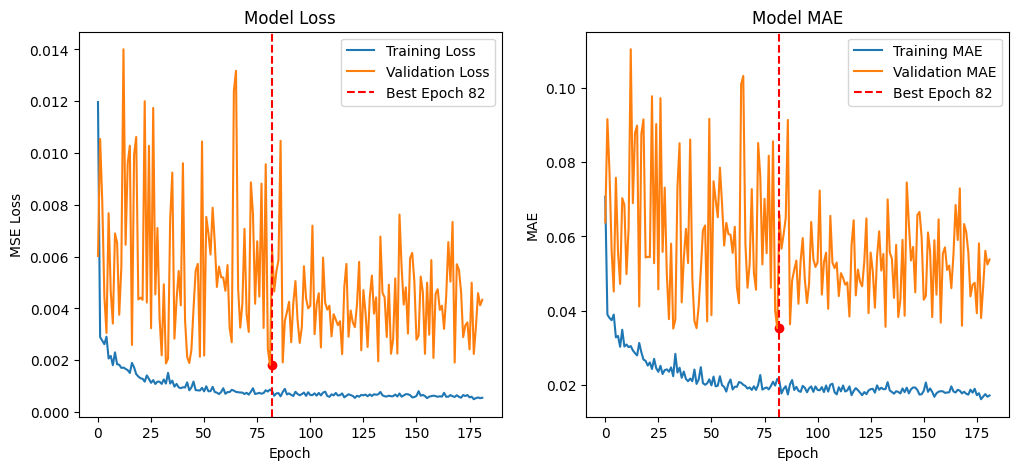


Figure 7. CNN-Bi-LSTM Model

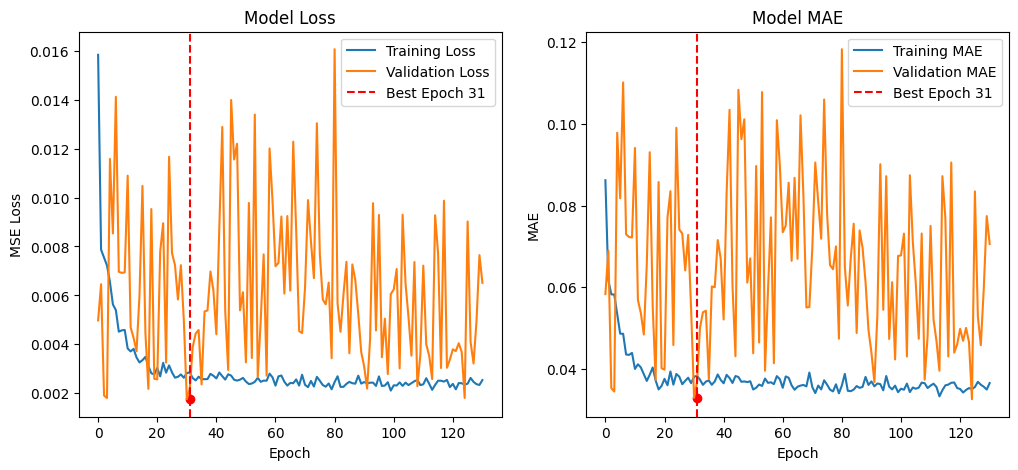


Figure 8. CNN Model

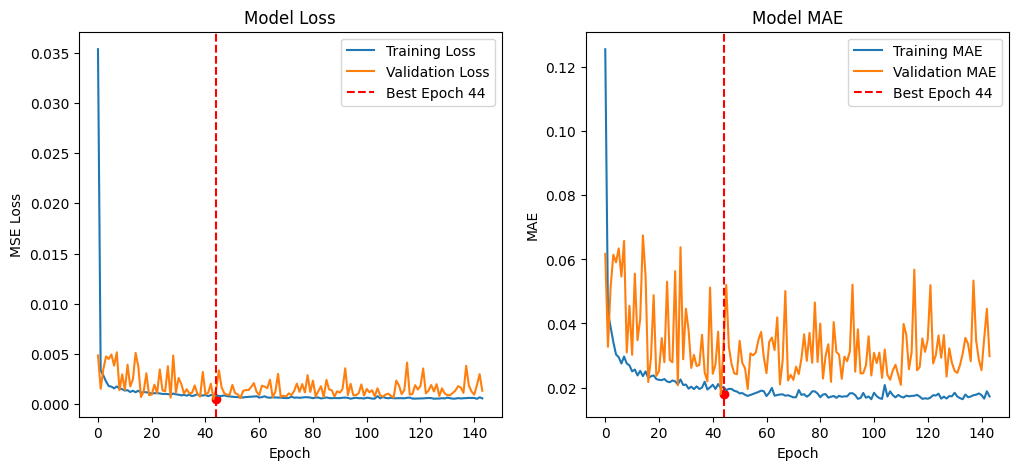


Figure 9. Bi-LSTM Model - 1Day

### 4.3. Performance of deep learning models (Bi-LSTM, CNN, Bi-LSTM -1Day, and Hybrid CNN-Bi-LSTM)

### The empirical assessment underscores the relative capabilities of four deep learning architectures which were trained utilizing the gold price dataset as depicted in Table 2, the Bi-LSTM (1-day) model surpassed all other models in nearly every evaluation metric. Notably it recorded the lowest RMSE (0.019) MAE (0.0133) and MAPE (0.80%) alongside the highest R² score (0.96) signifying a robust predictive ability and a close correspondence with actual fluctuations in gold prices as shown in Figure 10. In contrast the 30-day Bi-LSTM showed suboptimal performance, with an RMSE of 0.0337 and an R² value of 0.93 indicating that extending the temporal window may have introduced noise rather than enhancing predictive precision. Furthermore, the CNN and CNN-Bi-LSTM hybrid models exhibited moderate accuracy (RMSE ≈ 0.035, R² ranging from 0.89 to 0.93) suggesting that convolutional layers in isolation could not adequately capture the temporal dependencies that characterize the dynamics of gold prices and as shown in Figure 11.

### 

Table 2. Performance Comparison Across Models and Sequence Windows.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Sequence Window** | **RMSE** | **MSE** | **MAE** | **MAPE** | **R²** |
| **Bi-LSTM (30-day)** | 30 days | 0.0337 | 0.0011 | 0.0237 | 1.45% | 0.93 |
| **CNN-Bi-LSTM** | 30 days | 0.0349 | 0.0012 | 0.0279 | 1.77% | 0.89 |
| **CNN** | 30 days | 0.0350 | 0.0012 | 0.0277 | 1.82% | 0.93 |
| **Bi-LSTM (1-day)** | 1 day | 0.0190 | 0.0004 | 0.0133 | 0.80% | 0.96 |

### 

Figure 10. Actual an prediction gold price using Bi-LSTM (1-Day)

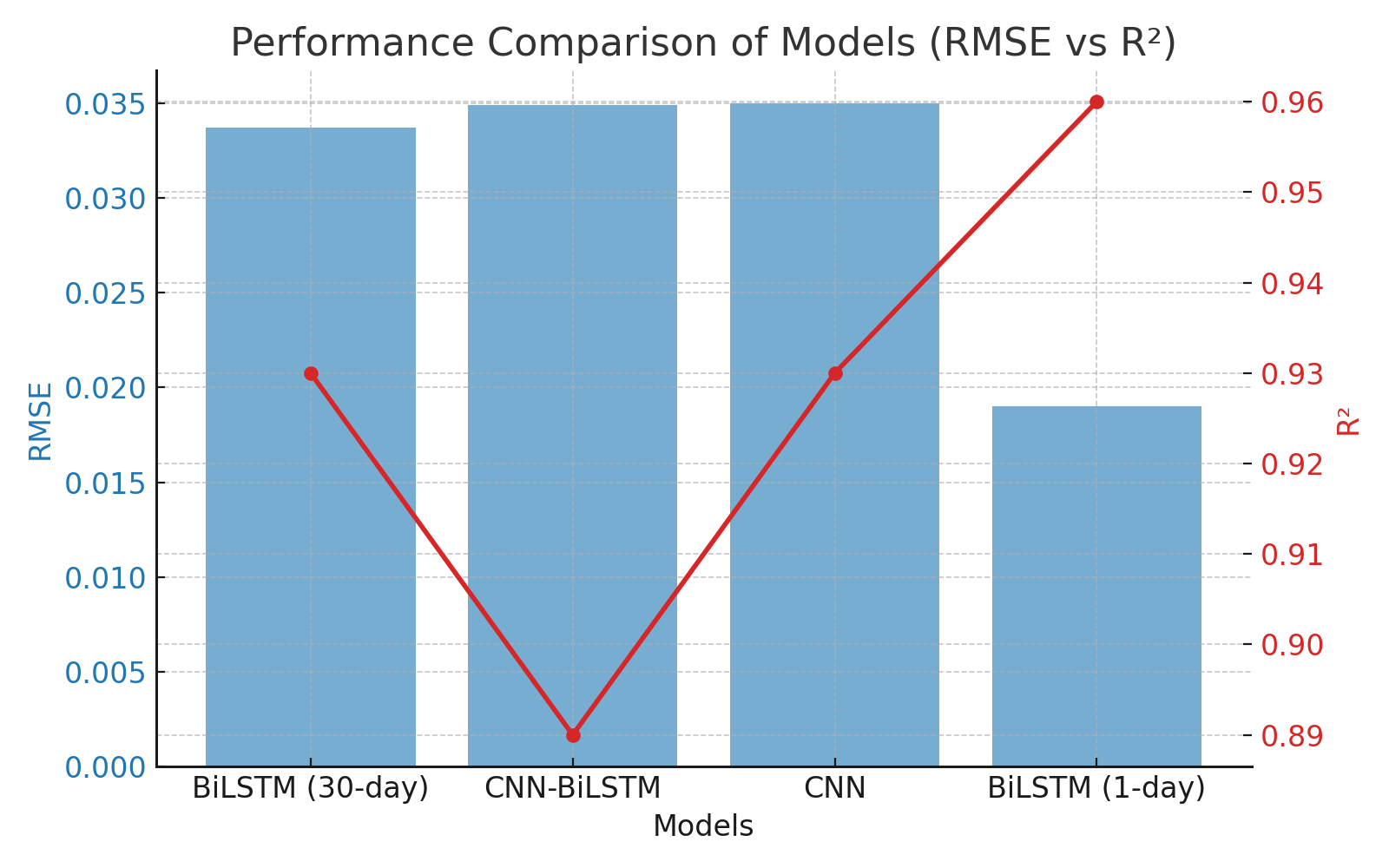


Figure 11. Performance Comparison of Models (RMSE vs R2) the red line for R2.

### 4.4. Feature Importance Using SHAP and permutation methods

In addition to the evaluation of the model the analysis of feature importance using SHAP and permutation methods yielded significant insights into the factors influencing fluctuations in gold prices. The features that ranked highest consistently included gold\_high, gold\_low, and gold\_open, thereby validating that intraday volatility and the conditions present at the market's opening are the most salient determinants in predicting gold price movements. Additionally, secondary factors such as oil\_high, silver\_close and the Consumer Price Index (CPI) imply that the interrelationships between commodities and macroeconomic indicators are also influential, though to a lesser degree. Conversely, in the context of this model, the S&P 500, EUR/USD, and Geopolitical Risk (GPR) appear to have negligible impacts. The coefficient estimates are statistically insignificant and their levels suggest that gold would be less responsive to these drivers in the out-of-sample period as shown in Figure 12.

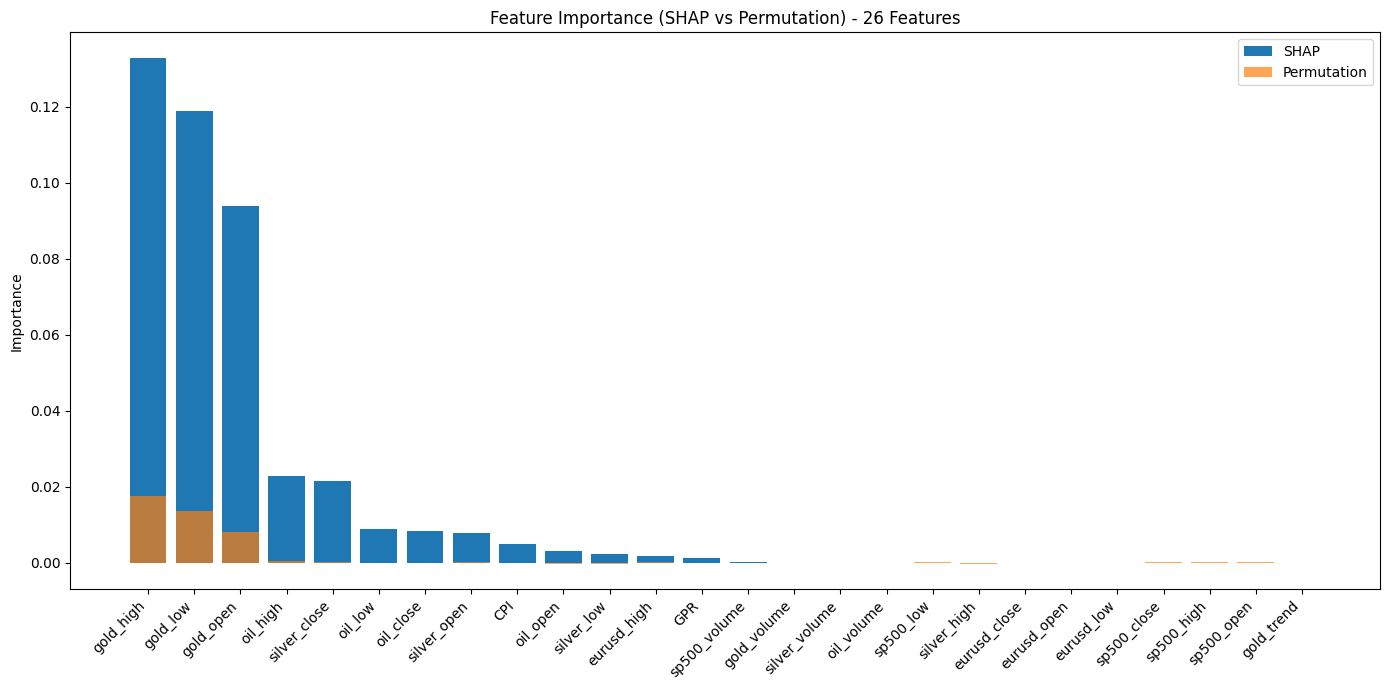


Figure 12. Feature Impotence (SHAP vs Permutation) - 26 Features.

## 5. Discussion and Analysis

In conclusion, this paper has provided insights into the data-driven forecasting capabilities of a range of deep learning architectures trained with limited features across a variety of time horizons, and carefully prioritized. Notable is the validation across three understandings of forecasting complexity. Certainly, a unifying yet novel conclusion is arriving at with effort- and computation-intensive neural architectures that the time horizon considered is the crucial factor to take into account at all stages of the forecasting pipeline. The outstanding performance of the one-day Bi-LSTM model clearly demonstrates that sequential short-term memory-based architectures are extremely effective for gold price forecasting in the presence of complex co-movements of driving factors. Furthermore, the results of the extensive permutation feature importance analysis affirm that for timely, gold price forecasting based on fundamental short-term commodity drivers is surprisingly superior to macroeconomic explanations. This consequently carries crucial implications for market participants and policy makers of our era. Simply put, our results suggest that the feat of performing complex gold price forecasting with a premium on predicting power and computational efficiency is indeed attainable by concentrating primarily on recent price action and fundamental commodity relationships. The further exploration of these relationships could be leveraged in formulating portfolio selection strategies and shaping effective regulatory framework decisions.

## 6. Conclusion

This study proposed a hybrid deep learning framework to predict gold price based on high-frequency financial market signals, macroeconomic indicators, and geopolitical risk. A wide range of models were trained and compared, including Bi-LSTM with multiple input sequence lengths, CNN-Bi-LSTM, and CNN approaches. The models were compared using rigorous evaluation metrics, including RMS, MSE, MAE, MAPE, R², and early stopping rules based on these metrics to avoid overfitting. Results show that Bi-LSTM with a 1-day sequence performed best among all models, attaining the lowest RMSE of 0.0190, MAE of 0.0133, and highest R² of 0.96.

The effectiveness of the Bi-LSTM model indicates that short-term temporal associations are more suitable for modeling the dynamics of the gold price than longer temporal associations. The CNN-Bi-LSTM and CNN models exhibited higher error rates, reaffirming the advantages of recurrent architectures for temporal and sequential financial time series forecasting.

Utilizing SHAP and permutation-based feature importance methodologies only revealed that the features pertaining to gold (namely gold\_high, gold\_low, and gold\_open) were of utmost significance; oil and silver features followed, whereas macroeconomic features as well as geopolitical events ranked last. This was expected, considering the principal factors that determine the price of gold as well as the above-mentioned co-dependence between commodities. Overall, the Bi-LSTM architecture seems to work well for time-shifted financial time series based on these results, and the importance of the features has been demonstrated. Possible future work could include the use of ensemble models incorporating real-time sentiment analysis of geopolitical occurrences, or the application of reinforcement learning to optimize the decision-making process within an algorithmic trading strategy.

The proposed Bi-LSTM architecture remains robust across diverse data settings, including price, fundamental, and technical datasets, suggesting its resilience and adaptability. Sensitivity analyses reveal that essential features influencing gold price trends include fluctuations in the US dollar's value, global inflation rates, changes in US Treasury yields, the price and volatility of gold futures, and macroeconomic indicators such as GDP growth. Also, the model's profit-maximizing trading strategy yields significant empirical returns, outperforming many established benchmarks in the literature. Beyond merely confirming the predictive power of factors outlined in previous literature, these findings highlight the intricate dynamics between gold and other indicators. Empirical results lend credence to refined versions of existing explanations in the economic literature, but their statistical significance does not definitively validate these theoretical frameworks. This paper could be of value to practitioners considering algorithmic trading strategies, especially regarding model selection or feature choices. Future investigations could center on applying ensemble models or classic ML algorithms to traditional low-frequency datasets, integrating up-to-the-minute news-based sentiment analysis features and other high-frequency variables, or deploying RL frameworks leveraging the findings of this paper in dynamic, decision-centric trading models.