Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 10, No. 3, June, 2026

Gold Price Forecasting Using Machine Learning Models with Hyperparameter Optimization for Inflation Hedging

Joan Sim Pei Suan [1], Kalaiarasi Sonai Muthu Anbananthen [1*], Raj Kumar Kanan [2]

[1] *Centre for Advanced Analytics, CoE for Artificial Intelligence & Faculty of Information Science and Technology, Multimedia University, Melaka 75450, Malaysia.*

[2] *Research Department of Computer Science, Bishop Heber College (Auto), Affiliated to Bharathidasan University Tiruchirappalli, Tamil Nadu, India.*

Abstract

Gold serves as a hedge against inflation, particularly in emerging markets such as Malaysia, where macroeconomic volatility is pronounced. This study evaluates the predictive performance of six machine learning models; comprising ensemble models (Random Forest, XGBoost, Gradient Boosting Machine, and LightGBM) and deep learning models (Long Short-Term Memory and Gated Recurrent Units) in forecasting Malaysia's gold prices using monthly macroeconomic data from 2009 to 2024. Key indicators include inflation rates, interest rates, exchange rates, oil prices, and stock indices. Hyperparameter tuning is performed using the Optuna framework by comparing three optimization strategies: Tree-structured Parzen Estimator (TPE), Grid Search, and Covariance Matrix Adaptation Evolution Strategy (CMA-ES). Experimental results show that Gradient Boosting, optimized via CMA-ES, achieves the best performance (RMSE = 101.26, R^2 = 0.9972) using the complete feature set. While deep learning models demonstrate improvements following optimization, ensemble models consistently outperform them due to better alignment with the static, cross-sectional nature of the dataset. Feature importance analysis identifies GP_Low, GP_High, and both domestic and international inflation and interest rates as the most significant predictors. This study contributes by benchmarking ensemble and deep learning models, evaluating multiple hyperparameter optimization strategies, and identifying key macroeconomic indicators relevant to gold price forecasting. The findings provide valuable insights for investors, financial analysts, and policymakers in economies sensitive to inflation.

Keywords:

Gold Price Forecasting;
Machine Learning;
Deep Learning;
Hyperparameter Optimization;
Macroeconomic Indicators;
Inflation Hedge.

Article History:

Received:	20	July	2025
Revised:	22	April	2026
Accepted:	02	May	2026
Published:	01	June	2026

1- Introduction

Gold has long been regarded as a strategic and reliable asset, particularly during periods of economic instability and inflationary pressure. A safe-haven asset is defined as one that remains uncorrelated or negatively correlated with other asset classes in times of market stress [1]. Historically, gold has fulfilled this role, acting as both a store of value and a hedge against inflation. In 2023, global gold demand surged to 4,974 tonnes, driven by record central bank acquisitions and growing investor interest, while the average LBMA gold price rose to US$2,386 per ounce—a 23% year-over-year increase [2]. These developments underscore the growing confidence in gold as a protective asset during economic crises [3].

In emerging economies such as Malaysia, gold investment has been recognized as an effective hedge against financial risk, particularly since the 1997/98 Asian Financial Crisis, when policymakers, including then Prime Minister Tun Mahathir Mohamad, emphasized gold's potential role in international trade and monetary stability [4].

* **CONTACT**: kalaiarasi@mmu.edu.my

DOI: https://doi.org/10.28991/ESJ-2026-010-03-016

Given Malaysia's open and trade-dependent economy, it remains highly vulnerable to global shocks such as oil price volatility, U.S. monetary policy shifts, and exchange rate fluctuations [5]. These external pressures often amplify domestic inflationary risks and financial market uncertainty, underscoring the importance of reliable gold price forecasting for investors, financial analysts, and policymakers seeking to manage economic risk and inform inflation-targeting strategies.

Despite its importance, forecasting gold prices remains challenging because of the nonlinear and dynamic relationships between macroeconomic indicators and gold market behavior. Traditional econometric models, such as the AutoRegressive Integrated Moving Average (ARIMA) and Vector AutoRegression (VAR), though widely applied, rely on assumptions of linearity and stationarity that limit their predictive accuracy in volatile environments [6, 7]. Consequently, researchers have increasingly explored machine learning (ML) and deep learning (DL) techniques capable of modeling complex, nonlinear relationships [8]. Ensemble ML methods such as Random Forest, Gradient Boosting Machines (GBM), and XGBoost have demonstrated improved forecasting accuracy, with studies showing Random Forest outperforming ARIMA and Decision Tree models [6], GBM excelling in short-term horizons [9], and XGBoost achieving the highest accuracy compared to five competing models, including Linear Regression and Neural Networks [10]. Similarly, DL models such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) have gained traction for sequential data modeling, with LSTM networks capturing temporal dependencies in COVID-19 and macroeconomic data more effectively than CNN-LSTM and Bidirectional LSTM variants [11]. At the same time, GRU performed better in shorter horizons and LSTM in longer forecasting periods [12]. Hybrid architectures that integrate multiple modeling approaches have further enhanced predictive performance; for example, the LSTM-Attention-CNN model combined memory, attention, and local feature extraction to improve forecasting accuracy [13], while the ICEEMDAN-LSTM-CNN-CBAM hybrid model outperformed 22 benchmark models [14]. More recently, sentiment-augmented approaches incorporating news and social media sentiment indicators have also demonstrated improved one-day-ahead forecasting accuracy [15-18].

Nevertheless, several gaps remain. Most studies focus on global markets or high-frequency trading environments, with limited attention to emerging economies such as Malaysia, where macroeconomic dynamics may differ substantially [5]. Moreover, although ensemble learners and recurrent neural networks have shown promise individually, few studies directly compare their forecasting performance under a unified framework. Finally, hyperparameter optimization—a factor known to influence model accuracy significantly—remains underexplored in gold price forecasting, despite the availability of advanced optimization techniques such as Tree-structured Parzen Estimator (TPE) and Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [19].

To address these gaps, this study benchmarks six ML models "our ensemble learning algorithms (Random Forest, GBM, XGBoost, and LightGBM) and two DL models (LSTM and GRU)" for forecasting Malaysia's gold prices using Malaysia-specific macroeconomic data. The Optuna framework [20] is employed to evaluate three hyperparameter optimization strategies: TPE, Grid Search, and CMA-ES. Specifically, the study seeks to answer the following research questions:

- Which ML model performs best in forecasting Malaysia's gold prices?
- How does hyperparameter optimization influence the predictive accuracy of these models?
- What are the most influential macroeconomic indicators affecting Malaysia's gold price?

The remainder of the paper is structured as follows: Section 2 reviews existing literature across three key areas—macroeconomic indicators, ML models, and optimization strategies. Section 3 outlines the dataset and methodology. Section 4 presents and interprets the results. Section 5 discusses practical implications, limitations, and future research directions.

2- Related Works

This section reviews prior literature relevant to gold price forecasting, structured around three core themes: Macroeconomic Indicators (Section 2-1), ML models (Section 2-2), and Hyperparameter Optimization Techniques (Section 2-3). These themes reflect the multidimensional nature of gold pricing and the methodological evolution from traditional econometrics to modern predictive analytics. The section concludes with a synthesis of the research gaps addressed by this study.

2-1- Macroeconomic Indicators Influencing Gold Prices

Domestic and global macroeconomic variables influence gold prices. Among these, inflation rates are frequently cited as key drivers, as gold is traditionally seen as a hedge against purchasing power erosion [21]. Empirical findings vary: Hashim et al. [21] found a positive relationship across major economies, while Md Hashim [22] demonstrated inflation

to be the most influential factor on Malaysian gold prices between 1980 and 2020. However, others, such as Bin Sukri & Mohd Zain [5] and Mohd et al. [23], reported no significant correlation in the Malaysian context, suggesting that the gold-inflation nexus may be market-specific or temporally unstable.

Interest rates also play a pivotal role in determining the opportunity cost of holding gold. Hashim et al. [21] and Mainal et al. [24] observed a negative correlation between interest rates and gold prices, aligning with the standard economic expectation. In contrast, studies by Md Hashim [22] and Mohd et al. [23] reported statistical insignificance, reflecting variability due to differences in data periods and modelling methods.

Exchange rates and GDP have also emerged as influential. Since gold is globally traded in U.S. dollars, depreciation of local currencies often increases domestic gold prices [5, 24]. GDP reflects investor confidence and purchasing capacity; however, findings have been inconsistent [21, 25].

Additional predictors such as the Volatility Index (VIX) and crude oil prices have been explored as proxies for economic uncertainty and commodity interlinkages, respectively [26, 27]. These studies highlight the multidimensional nature of gold pricing while also revealing contextual and methodological divergences, particularly in emerging markets such as Malaysia.

In this study, identifying the most influential macroeconomic indicators affecting Malaysia's gold price is a central objective. Understanding which variables—such as inflation rates, interest rates, exchange rates, and commodity prices—most strongly impact gold price fluctuations can provide valuable insights for investors, policymakers, and financial analysts. To support this, a feature importance analysis is employed to interpret model predictions and quantify the contribution of each variable to the forecasting output. This interpretability component complements model performance evaluation and enhances transparency in financial decision-making.

2-2- ML Models in Gold Price Prediction

Traditional econometric models such as ARIMA, VAR, and ARDL have long been used in gold price prediction but often fall short in handling non-linearity and time-series volatility [6, 8]. The limitations of these models have motivated the adoption of ML techniques, which offer greater flexibility and predictive power.

Recent studies demonstrate how the gold price forecasting has progressed over the years, from standalone ML and DL models to hybrid and adaptive models. For instance, Taneva-Angelova et al. [15] created a hybrid gold forecasting framework that integrates the classical econometric modelling with ML, DL, and sentiment analysis. Using data from 2014 to 2024, their framework outperformed the traditional econometric approaches. In the same breath, Guo et al. [28] showed that signal decomposition combined with recurrent networks like GRU and LSTM is able to push the forecasting performance by a large margin beyond what is possible using raw sequences. Specifically, their EMD-BiGRU-AM model, built on GRU, performed the best compared to other variants. Other than that, fuzzy logic and optimization approaches have been used in other works. For example, Das et al. [29] designed a combination of Extreme Learning Machine with a Fuzzy Inference System, which was enhanced with a Cluster-based Quasi-Oppositional Crow Search Algorithm (CQCSA). With daily Brent crude oil, WTI crude oil, and gold prices between 2010 and 2022, the proposed model exhibited dramatically lower prediction errors than its benchmarks, testifying to the efficiency of fuzzy and metaheuristic-based optimization in machine learning. Beyond hybrid learning, researchers have also investigated novel input representations.

Liu et al. [30] proved the predictive ability of news sentiment by using textual analysis in predicting short-term gold futures prices. Besides, Salim & Djunaidy [31] developed a hybrid CNN-LSTM architecture that converts daily gold prices (2013-2022) into images via Gramian Angular Fields. The authors who proved the superior performance of CNN-LSTM using image input data revealed that the integration of deep learning, as well as time-series-to-image representations, can be used as an efficient measure to enhance the gold price prediction. Altogether, these developed works clarify the transition towards hybrid and adaptive ML models in the gold price prediction.

While hybrid and adaptive models have shown promising results, they usually require large or multimodal inputs, which are not always available in the localized context. In the case of Malaysia, key macroeconomic indicators such as inflation and interest rates are only reported on a monthly, quarterly, or yearly basis, which results in relatively small datasets. Under such conditions, ensemble models offer an interpretable solution suited for structured tabular data. Ensemble models have demonstrated strong performance [32] across various studies (see Table 1). For example, Manjula and Karthikeyan [9] found that GBM performed best in short-term forecasts, while RF performed better over more extended periods. Jabeur et al. [10] demonstrated the superiority of XGBoost with a 99.4% R^2, outperforming CatBoost, LightGBM, and traditional regression. Similarly, Rady et al. [6] found that GBM outperforms both ARIMA and Decision Trees on monthly gold price data.

Table 1. Summary of ML and DL Models for Gold Price Forecasting

Study	Models Evaluated	Variables Used	Best Performing Model	Evaluation Metrics
Makala & Li [33]	ARIMA, SVM	Daily gold data	SVM	RMSE, MAPE, R^2
Anbananthen et al. [8]	ARIMA, LSTM	12 financial time series indices (e.g., S&P 500, Nikkei 225)	LSTM	RMSE
Manjula & Karthikeyan [9]	RF, GBM, Linear Regression	Crude oil prices, inflation, interest rates, and exchange rates	GBM (short-term), RF (long-term)	MSE, RMSE, MAE
Rady et al. [6]	ARIMA, Decision Tree, RF, GBM	Monthly gold prices	RF	RMSE
Jabeur et al. [10]	XGBoost, CatBoost, RF, LightGBM, Linear Regression, Neural Network	Silver price, oil price, exchange rates, S&P 500, and iron ore	XGBoost	RMSE, MAE, MSE, R^2
Narendran [34]	GPR, ARIMA, CatBoost, LightGBM, LSTM, XGBoost	Daily gold price	LightGBM	MAPE, MAE, Accuracy
Zangana & Obeyd [35]	LSTM, BiLSTM	Oil prices, exchange rates, inflation, and interest rates	BiLSTM	MAE, RMSE, R^2
Mohtasham Khani et al. [11]	CNN-LSTM, Vector LSTM, Encoder-Decoder LSTM, BiLSTM	Gold price, COVID-19 trends, stock indices	Vector LSTM	RMSE, MAE, MSLE
Primananda & Isa [12]	GRU, LSTM	Gold price in multiple currencies, stock indices	GRU (short-term), LSTM (long-term)	RMSE, MAPE
Dewi et al. [36]	GRU, LSTM	Gold prices	GRU	RMSE

In addition to ensemble models, DL models are considered in this study for comparative robustness. DL models, such as LSTM and GRU, have shown promise in capturing sequential patterns [35], and reported improved accuracy using BiLSTM over LSTM. Mohtasham Khani et al. [11] leveraged CNN-LSTM and vector-sequence LSTM to enhance the performance of volatile periods by incorporating COVID-19-related variables. However, several studies, including Jabeur et al. [10], noted that DL models may underperform on small datasets due to their data-hungry nature and complex architecture. Although the use of monthly Malaysian data presents constraints in terms of sample size, including DL approaches allows this study to benchmark their performance against ensemble methods under localized, small-sample conditions.

On the whole, the literature suggests the promising applicability of hybrid and adaptive models. Still, they usually require volume and multi-modal data that are prohibitive in local settings, as shown in studies such as [15, 28, 29] that depended on high-frequency data and decomposed datasets. In comparison, ensemble and deep learning approaches have been consistently validated across diverse studies (see Table 1), with outcomes varying based on model architecture, data volume, and the complexity of economic indicators. However, a lack of comprehensive comparative benchmarking remains between these paradigms, particularly when applied to localized, macroeconomic datasets in emerging markets such as Malaysia. This gap highlights a key motivation for the present study.

2-3- Hyperparameter Optimization in Financial Forecasting

Hyperparameter tuning is a critical yet often underexplored component of financial ML model development. As highlighted by Hutter et al. [19], model performance is susceptible to hyperparameter configurations, particularly in complex, non-linear domains such as economic forecasting. Traditional tuning methods, such as Grid Search and Random Search, are widely used but suffer from inefficiencies. Grid Search is exhaustive and computationally expensive, while Random Search may overlook optimal configurations in high-dimensional search spaces [37].

Recent advancements have introduced more adaptive and efficient optimization frameworks. One such tool is Optuna, a define-by-run hyperparameter optimization library developed by Akiba et al. [20]. Optuna supports multiple optimization strategies, including TPE, CMA-ES, and traditional Grid/Random Search. Its dynamic search space construction, combined with an early stopping (pruning) mechanism, makes it particularly effective for optimizing models in high-dimensional, noisy financial data settings [38]. A recent study from Kausar et al. [39] has justified the applicability of Optuna in financial forecasting. The authors created a CEEMDAN-based hybrid ensemble framework for foreign volatility prediction. They showed that the Optuna-optimized deep learning models consistently performed better than Particle Swarm Optimization (PSO) in terms of accuracy and computation. Beyond financial applications, Optuna has also been adopted in other domains.

For instance, Lai et al. [40] proposed an Optuna-based system (MLOPTA) that optimizes the hyperparameters of SVM, Decision Tree, Random Forest, XGBoost, and LightGBM for disease prediction. While their empirical results were in the healthcare domain, the key insight was that Optuna's adaptive search strategy particularly enhanced recall, making the models less likely to miss rare but important events. Translating this implication to financial forecasting, the

ability to avoid ‘missed signals’ can be critical when detecting directional shifts in noisy and sparse data. Other than the conventional tuning methods, such as Random Search, multiple hyperparameter optimization frameworks have been used in recent studies, such as HyperOpt and Ray Tune. HyperOpt, which is based on Bayesian Optimization, has an edge in computational efficiency [41]. Still, its static search space and absence of a native mechanism for pruning limit its effectiveness in high-dimensional and noisy datasets [42]. While Ray Tune is a distributed experimentation platform that is designed for large-scale training, using it for small datasets could introduce overhead and is often overkill. Other than that, Sequential Model-based Algorithm Configuration (SMAC), a Bayesian Optimization tool that uses Random Forests to handle mixed parameter types and is well–known in Combined Algorithm Selection and Hyperparameter (CASH) optimization. However, it is more complex in terms of computation and implementation, making it more suited for large-scale AutoML frameworks [43].

Considering these trade-offs among the existing frameworks in the context of small and noisy financial datasets, this study employs Optuna to optimize hyperparameters across all selected ML models. Its modular architecture enables fair and consistent comparison across optimization strategies, while its efficiency ensures feasible execution within the constraints of financial forecasting experiments.

2-4- Summary of Research Gaps

While the reviewed studies have contributed valuable insights into gold price forecasting, several persistent gaps remain. First, there are limitations in geographic and data granularity. The majority of existing work relies on high-frequency, globally aggregated datasets, which may not generalize effectively to low-frequency, localized contexts such as Malaysia. In Malaysia, the key macroeconomic indicators, such as inflation rates and interest rates, are only reported on a monthly or annual basis. This reporting cycle makes monthly data the most reliable frequency for capturing the linkages of macroeconomic-gold price relevant to inflation-hedging policy. Using higher-frequency data may introduce noise or misalignment of policy signals, which can decrease interpretability to financial analysts.

Second, comprehensive benchmarking that incorporates ensemble learning and DL models using Malaysia-specific macroeconomic indicators is lacking. To address this, the present study adopts a dual modelling strategy. Six predictive models were selected to represent two distinct ML paradigms: ensemble learning and DL. This approach reflects the hybrid nature of the dataset, which comprises structured tabular economic indicators and exhibits temporal dependencies across a 15-year monthly time series.

Ensemble models—including RF, GBM, XGBoost, and LightGBM- are well-suited for capturing nonlinear interactions and variable importance in structured data. These models are known for their robustness to multicollinearity, scalability, and consistent performance in economic and financial forecasting tasks [9, 10, 44].

In contrast, DL models such as LSTM and GRU can model sequential dependencies and long-term temporal dynamics. These architectures are specifically designed for time series prediction, using gating mechanisms and memory cells to learn from lagged inputs [45, 46].

Third, although hyperparameter tuning significantly impacts model performance, few studies systematically compare multiple optimization strategies within a unified experimental framework. This study addresses this issue by implementing three hyperparameter optimization strategies—TPE, Grid Search, and CMA-ES—using the Optuna framework, which supports flexible and dynamic search space construction and early stopping through pruning.

Finally, while many studies focus on model accuracy, few investigate which macroeconomic variables are most influential in predicting gold prices. Understanding these drivers is critical for enhancing the interpretability and practical relevance of forecasting models. To address this, the study incorporates feature importance analysis to identify and rank key predictors of Malaysia's gold price.

This study addresses these gaps by:

- Benchmarking six ML models using monthly Malaysian macroeconomic data.
- Comparing three hyperparameter optimization strategies (TPE, Grid Search, and CMA-ES) under a unified experimental design using the Optuna framework.
- Examining which macroeconomic variables most strongly influence Malaysia's gold price through feature importance analysis.

3- Methodology

This section outlines the methodological framework adopted to forecast Malaysian gold prices. It covers the dataset and preprocessing procedures, feature selection, ML model design, hyperparameter optimization using Optuna, and performance evaluation metrics.

3-1- Dataset Overview

The dataset used in this study comprises 11 macroeconomic variables collected monthly from January 2009 to September 2024, totaling 189 records, as shown in Table 2. The target variable is GP_Close, representing the closing price of gold in Malaysian Ringgit. The remaining variables serve as predictors and are listed in Table 2 along with their descriptions and sources.

Table 2. Dataset Overview

	Variables	Description	Source
1	GP_Close	Closing price of gold in Ringgit Malaysia	*https://www.investing.com*
2	GP_High	The highest price of gold in Ringgit Malaysia	*https://www.investing.com*
3	GP_Low	Lowest price of gold in Ringgit Malaysia	*https://www.investing.com*
4	MIFR	Malaysia inflation rates (2010=100)	*https://open.dosm.gov.my* *https://tradingeconomics.com*
5	MITR	Malaysia's interest rates	*https://financialmarkets.bnm.gov.my*
6	UIFR	U.S. inflation rates (2010=100)	*https://fred.stlouisfed.org*
7	UITR	U.S interest rates	*https://fred.stlouisfed.org*
8	USD_MYR	Malaysia-US Dollar exchange rates	*https://www.investing.com*
9	VIX	CBOE Volatility Index- closing prices (USD)	*https://www.investing.com*
10	BRENT	Brent crude oil spot prices (USD)	*https://www.eia.gov*
11	KLCI	FTSE Bursa Malaysia KLCI- closing prices (MYR)	*https://www.investing.com*

Table 3 illustrates an excerpt of the raw dataset, showcasing the typical data structure and format. Monthly data were used because key macroeconomic indicators in Malaysia, such as inflation rates and interest rates, are only reported monthly or annually. Using higher-frequency data would either require omitting these essential variables or interpolating missing values, potentially introducing noise and reducing reliability. Monthly data thus provide a consistent and policy-relevant framework for macroeconomic forecasting.

Table 3. Raw Dataset

Date	GP_Closing	GP_High	GP_Low	MIFR	MITR	UIFR	UITR	USD_MYR	VIX	BRENT	KLCI
2009M1	3346.86	3378.65	2775.16	98.1	2.5	96.8	0.15	3.6075	44.84	43.44	884.45
2009M2	3504.14	3731.23	3176.82	98.3	2	97.3	0.22	3.7075	46.35	43.32	890.67
2009M3	3347.12	3618.71	3190.44	98.1	2	97.5	0.18	3.6455	44.14	46.54	872.55
2009M4	3155.94	3413.26	3069.02	97.9	2	97.8	0.15	3.5600	36.50	50.18	990.74
2009M5	3417.75	3508.55	3057.71	98.1	2	98.1	0.18	3.4875	28.92	57.30	1044.11
2009M6	3257.88	3524.28	3170.32	98.2	2	98.9	0.21	3.5150	26.35	68.61	1075.24
2009M7	3363.11	3456.84	3176.89	98.3	2	98.8	0.16	3.5225	25.92	64.44	1174.90
2009M8	3349.47	3450.83	3243.84	98.4	2	99.0	0.16	3.5215	26.01	72.51	1174.27
2009M9	3487.65	3635.14	3275.67	98.7	2	99.0	0.15	3.4610	25.61	67.65	1202.08
2009M10	3565.72	3731.55	3300.51	98.8	2	99.1	0.12	3.4125	30.69	72.77	1243.23
2009M11	4001.96	4106.81	3488.05	99.0	2	99.2	0.12	3.3948	24.51	76.66	1259.11
2009M12	3753.39	4223.18	3621.40	99.2	2	99.0	0.12	3.4240	21.68	74.46	1272.78
2010M1	3693.28	3987.75	3576.07	99.4	2	99.4	0.11	3.4148	24.62	76.17	1259.16
2010M2	3799.23	3906.31	3533.19	99.4	2	99.4	0.13	3.4028	19.50	73.75	1270.78
2010M3	3632.66	3894.47	3536.61	99.4	2.25	99.8	0.16	3.2615	17.59	78.83	1320.57

3-2- Data Preparation and Preprocessing

To ensure model reliability, the following preprocessing steps were applied:

- Data Cleaning: Currency normalization, date formatting, and missing value imputation using forward-fill resampling.
- Data Normalization: Min-Max Scaling was applied to scale all variables between 0 and 1, mitigating the bias toward features with larger magnitudes [47].

- Data Splitting: To ensure robust model evaluation, different data splitting strategies were applied based on model architecture. For tree-based ensemble models, including RF, XGBoost, GBM, and LightGBM, 5-fold cross-validation was employed. This approach partitions the data into five subsets, iteratively training on four folds and validating on the remaining one. It reduces variance and enhances model generalizability [48, 49] and is widely accepted for tabular data structures [50]. For DL models such as LSTM and GRU, a traditional 80/20 chronological train-test split was employed to preserve the temporal sequence of observations. This follows best practices outlined by Brownlee [51], who cautions against using random k-fold cross-validation in time series contexts, as it can introduce data leakage and distort performance estimates due to broken temporal order. While k-fold cross-validation is standard for ensemble models and chronological splits are recommended for time-series DL models to preserve temporal ordering, we ensured fairness by evaluating all models on the same test set using identical performance metrics. This approach reflects best practices for each model type; however, we acknowledge that validation protocols were not fully harmonized across paradigms.

3-3- Feature Selection

Feature selection was conducted using Pearson correlation analysis to assess the linear relationships between each independent variable and the target variable (GP_Close). Variables with a correlation coefficient of 0.6 or greater were included in the Selected Feature Set, representing those with strong linear associations. Conversely, the Full Feature Set retained all available variables, including those with weaker correlations. The strength of these correlations is visualized in Figure 1 and summarized in Table 4, while Table 5 presents the assignment of variables to their respective feature sets.

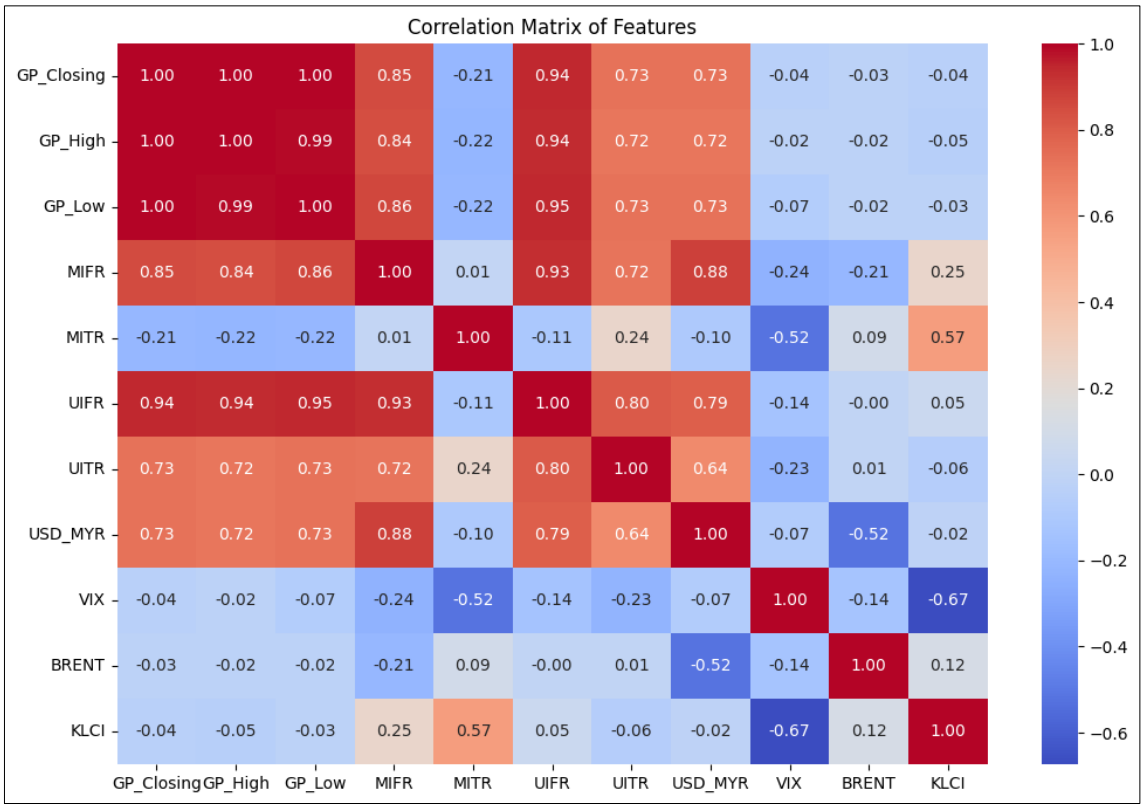

Figure 1. Correlation Matrix

Table 4. Correlation Strength

Correlation with GP_Closing	Variables
Strong correlation relationship	GP_High, GP_Low, MIFR, UIFR, UITR, USD_MYR
Weak correlation relationship	MITR, VIX, BRENT, KLCI

Table 5. Selected and Full Feature Sets

Correlation with GP_Closing	Variables
Strong correlation relationship	GP_High, GP_Low, MIFR, UIFR, UITR, USD_MYR
Weak correlation relationship	MITR, VIX, BRENT, KLCI

Employing both feature configurations enables a dual analysis strategy designed to enhance model robustness. The Selected Feature Set, comprising variables that exhibit strong correlations with the target, helps reduce dimensionality, mitigate noise, and minimize the risk of overfitting. In contrast, the Full Feature Set, which includes both strongly and weakly correlated predictors, may capture complex non-linear interactions and latent relationships that are not detectable through linear correlation analysis alone. This comparative approach supports a more comprehensive evaluation of model performance under varying input complexities.

3-3-1- ML Models

Six models were selected in this study to represent a diversity of learning paradigms, specifically ensemble learning and DL approaches. This variety ensures a comprehensive benchmarking of predictive strategies suited to both structured tabular data and temporal sequences.

Ensemble models:

- RF constructs an ensemble of decision trees during training and outputs the average of their predictions. It leverages bagging (bootstrap aggregation) and random feature selection to reduce variance and improve generalization performance [44].
- GBM builds learners sequentially, where each tree is trained to correct the residuals of the previous model, gradually improving prediction accuracy through iterative refinement.
- XGBoost is an optimized version of GBM that incorporates L1 and L2 regularization to prevent overfitting and improve generalization by penalizing model complexity.
- LightGBM is a fast, histogram-based boosting framework that employs techniques such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to enhance scalability and speed, especially in high-dimensional datasets.

Equation 1 shows the function of the prediction of RF, whereas Equation 2 shows the prediction function of GBM, XGBoost, and LightGBM.

$$\hat{Y}\,\frac{1}{T}\sum_{t=1}^{T} f_t(X) \tag{1}$$

$$\hat{Y} = \sum_{t=1}^{T} f_t(X) \tag{2}$$

L models:

- LSTM networks are a type of recurrent neural network (RNN) designed to capture both short-term and long-term dependencies in sequential data. The architecture incorporates a memory cell and three gates: forget, input, and output, which regulate the flow of information, enabling the model to retain relevant features across time steps.
- GRU is a streamlined variant of LSTM that employs only two gates: update and reset, to control the flow of memory. Its simpler architecture allows for faster training, fewer parameters, and reduced risk of overfitting, particularly in small to medium-sized datasets.

To adapt both LSTM and GRU to the relatively small monthly dataset, we kept the architectures simple with a single recurrent layer containing 32–64 hidden units. We applied dropout during training to reduce overfitting. We used early stopping to halt training when validation performance ceased to improve. These strategies are widely recommended for small datasets as they reduce model complexity, mitigate overfitting, and improve generalization performance without requiring extensive hyperparameter tuning.

Equation 3 shows the prediction function of LSTM and GRU:

$$\hat{Y}_t = \sigma\left(W_y \cdot h_t + b_y\right) \tag{3}$$

In this study, the six models are grouped into two learning paradigms to assess their ability to capture different types of data relationships. The ensemble models (RF, GBM, XGBoost, LightGBM) and the deep learning models (LSTM and GRU) are evaluated under a unified experimental framework to compare their effectiveness in modeling static versus temporal relationships in the context of macroeconomic gold price forecasting. This design enables a structured assessment of how well each paradigm handles different data dependencies within the same forecasting task.

3-4- Hyperparameter Optimization with Optuna

Hyperparameter tuning critically impacts model performance but is often underreported in financial ML studies [19]. This study employs Optuna [20], a define-by-run optimization framework, due to its flexibility, dynamic search space construction, and built-in pruning mechanism. Optuna was selected over alternatives such as Hyperopt and Ray Tune because of its flexible search space design, efficient pruning mechanisms, and ease of use without the overhead of

distributed frameworks. These features allowed for fast experimentation while maintaining accuracy. No major practical limitations were encountered when applying Optuna to our study.

Three optimization strategies are applied in this study to optimize the hyperparameter:

- TPE: A Bayesian method that models good vs. bad parameter regions and selects based on expected improvement [37].
- Grid Search: Exhaustive combinatorial search.
- CMA-ES: Evolutionary strategy using Gaussian sampling and covariance matrix updates.

For each model, a predefined set of hyperparameters and search ranges was established based on prior studies and empirical performance considerations. These are summarized in Table 6. The parameters were sampled using integer ranges, uniform, log-uniform, and categorical options, depending on their nature. Each model was tuned independently using all three optimization strategies to ensure a fair comparison. The best configuration for each model was selected based on the lowest validation RMSE during the tuning process. The final models were then evaluated on the test set using RMSE, MAE, and R^2 to assess performance.

Table 6. Hyperparameter Search Space per Model

Model	Hyperparameter	Search Range / Options
Random Forest	n_estimators	50 – 500 (integer range)
	max_depth	3 – 20 (integer range)
	min_samples_split	2 – 20 (integer range)
	min_samples_leaf	1 – 10 (integer range)
	max_features	[None, "sqrt", "log2"] (categorical options)
	bootstrap	[True, False] (categorical options)
XGBoost	learning_rate	0.01 – 0.3 (log-uniform)
	n_estimators	50 – 100 (integer range)
	max_depth	3 – 15 (integer range)
	subsample	0.5 – 1.0 (uniform)
	gamma	1e-8 – 1.0 (log-uniform)
	colsample_bytree	0.5 – 1.0 (uniform)
	reg_alpha	1e-8 – 10.0 (log-uniform)
	reg_lambda	1e-8 – 10.0 (log-uniform)
LightGBM	learning_rate	0.01 – 0.3 (log-uniform)
	n_estimators	50 – 500 (integer range)
	max_depth	3 – 15 (integer range)
	subsample	0.5 – 1.0 (uniform)
	min_child_samples	5 – 50 (integer range)
	colsample_bytree	0.5 – 1.0 (uniform)
	reg_alpha	1e-8 – 10.0 (log-uniform)
	reg_lambda	1e-8 – 10.0 (log-uniform)
GBM	learning_rate	0.01 – 0 3 (log-uniform)
	n_estimators	50 – 100 (integer range)
	max_depth	3 – 15 (integer range)
	subsample	0.5 – 1.0 (uniform)
	min_samples_split	2 – 20 (integer range)
	min_samples_leaf	1 – 10 (integer range)
	max_features	[None, "sqrt", "log2"]
LSTM	n_units_1	32 – 128 (integer range)
	n_units_2	32 – 128 (integer range)
	batch_size	[16, 32, 64]
	epochs	30 – 100 (integer range)
	dropout_rate	0.1 – 0.5 (uniform)
GRU	n_units_1	32 – 128 (integer range)
	n_units_2	32 – 128 (integer range)
	batch_size	[16, 32, 64]
	epochs	30 – 100 (integer range)
	dropout_rate	0.1 – 0.5 (uniform)

This multi-strategy optimization process enables the study to determine which model performs best overall and evaluate the relative effectiveness of each optimization method across different model architectures.

3-5- Evaluation Metrics

Three evaluation metrics were utilized to evaluate the predictive performance of the ML:

- Root Mean Squared Error (RMSE) measures the standard deviation of residuals. A lower RSME value indicates a better predictive accuracy. Equation 4 presents the formula for RMSE.

$$RMS = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (4)$$

- Mean Absolute Error (MAE) measures the average of the absolute differences between predicted and actual values. A lower MAE suggests a better predictive accuracy. Equation 5 presents the formula for MAE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad (5)$$

- The coefficient of Determination (R^2) measures the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model. It indicates goodness-of-fit. A value of $R^2 = 1$ denotes a perfect fit, where the model describes all variability in the target variable. In contrast, a value of $R^2 = 0$ indicates that the model fails to capture any of the variance. Intermediate values reflect partial explanatory power. Equation 6 presents the formula of R^2.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (6)$$

4- Results and Analysis

This section presents the performance evaluation. It is structured into four parts: Baseline Model Performance (Section 4-1), The Impact of Hyperparameter Optimization Using Optuna (Section 4-2), Visualization of Predicted vs Actual Gold Prices (Section 4-3), and Feature Importance Analysis (Section 4-4).

4-1- Baseline Model Performance

Table 7 summarizes the baseline performance of the six models, including ensemble models (XGBoost, RF, GBM, and LightGBM) and DL models (LSTM and GRU), trained using the default hyperparameter settings. Each model was evaluated on both the Selected Feature Set (SF set) and the Full Feature Set (FF set) using RMSE, MAE, and R^2 as performance metrics.

Table 7. Baseline Model Performance

		RMSE	MAE	R^2
XGBoost	SF set	148.215	117.357	0.9941
	FF set	152.898	118.853	0.9937
RF	SF set	124.711	99.901	0.9958
	FF set	131.752	103.489	0.9953
GBM	SF set	119.056	99.752	0.9962
	FF set	117.808	96.894	0.9963
LighGBM	SF set	177.373	128.472	0.9862
	FF set	207.515	146.834	0.9812
GRU	SF set	5748.321	5414.546	-7.8675
	FF set	5777.822	5445.831	-7.9588
LSTM	SF set	5768.018	5435.548	-7.9284
	FF set	5812.250	5482.368	-8.0659

The baseline results reveal that models trained on the Selected Feature Set generally outperform those using the Full Feature Set across most configurations. This suggests that carefully selected macroeconomic indicators are more informative than a naive feature space expansion and indicates that focusing on strongly correlated variables enhances model generalization by reducing noise and dimensionality. Among all models, GBM consistently achieved the best performance, recording the lowest RMSE (119.056), MAE (99.752), and highest R^2 (0.9962) with the Selected Feature Set. Interestingly, when trained on the Full Feature Set, GBM still led with an even lower RMSE (117.808), MAE (96.894), and slightly higher R^2 (0.9963), indicating its robustness in leveraging both strong and weak predictors through iterative refinement.

All ensemble models demonstrated stable and high accuracy, with R^2 values exceeding 0.98. In contrast, DL models performed poorly, yielding negative R^2 values and RMSE exceeding 5700, indicating severe overfitting and poor generalization. These performance differences can be attributed to the architectural suitability of ensemble models for structured tabular data, particularly when working with relatively small datasets. Similar findings were reported by Bailly et al. [52], who highlighted the ensemble models with the row-column nature of economic data. In contrast, DL models typically require larger training volumes and extensive hyperparameter tuning to perform effectively. The suboptimal performance observed in this study likely reflects issues of overfitting and poor convergence, arising from insufficient data and default parameter configurations. This observation aligns with findings by Dewi et al. [36] and Greff et al. [53] who emphasize the sensitivity of ML and DL to both dataset size and tuning complexity.

In summary, the baseline results highlight the robustness of ensemble models relative to DL models, and suggest that GBM offers a strong foundation for gold price forecasting.

4-2- Optimized Model Performance Using Optuna

To improve model accuracy, hyperparameter optimization was conducted using the Optuna framework, incorporating three strategies: TPE, Grid Search, and CMA-ES. Table 8 results show a substantial improvement across all models after post-optimization, confirming that the default set parameters were suboptimal. GBM again emerged as the top performer, achieving an RMSE of 101.261 and R^2 of 0.9972 using CMA-ES on the Full Feature Set. XGBoost, optimized via TPE, ranked second with RMSE of 107.50 and R^2 of 0.9969. Both RF and LightGBM also showed notable improvements.

Table 8. Optimized Model Performance Metrics

			RMSE	MAE	R^2
XGBoost	TPE	SF set	107.500	88.746	0.9969
		FF set	108.618	90.064	0.9968
	Grid Search	SF set	109.861	89.633	0.9968
		FF set	112.991	94.878	0.9966
	CMA-ES	SF set	110.037	91.506	0.9968
		FF set	109.647	89.754	0.9968
RF	TPE	SF set	117.552	93.627	0.9963
		FF set	115.282	92.988	0.9964
	Grid Search	SF set	120.831	96.019	0.9961
		FF set	125.517	100.710	0.9958
	CMA-ES	SF set	119.684	95.291	0.9962
		FF set	119.852	96.999	0.9961
GBM	TPE	SF set	112.076	92.505	0.9966
		FF set	110.997	84.556	0.9967
	Grid Search	SF set	115.650	99.356	0.9964
		FF set	115.419	91.374	0.9964
	CMA-ES	SF set	113.283	91.123	0.9966
		FF set	101.261	77.826	0.9972
LightGBM	TPE	SF set	145.770	109.672	0.9907
		FF set	151.886	111.999	0.9899
	Grid Search	SF set	212.403	164.235	0.9803
		FF set	184.414	149.990	0.9841
	CMA-ES	SF set	151.445	121.686	0.9900
		FF set	184.215	143.078	0.9852
GRU	TPE	SF set	403.791	271.436	0.9562
		FF set	626.577	332.571	0.8946
	Grid Search	SF set	1066.66	523.206	0.6947
		FF set	980.213	503.716	0.7422
	CMA-ES	SF set	729.662	308.072	0.8571
		FF set	872.707	383.083	0.7956
LSTM	TPE	SF set	589.389	263.189	0.9068
		FF set	1164.43	570.012	0.6361
	Grid Search	SF set	1502.91	1033.05	0.3938
		FF set	1720.16	1294.52	0.2059
	CMA-ES	SF set	705.542	300.977	0.8664
		FF set	836.323	391.820	0.8123

Although GRU and LSTM showed meaningful gains, with R^2 improving from negative values to 0.9562 (GRU) and 0.9068 (LSTM), they still underperformed compared to ensemble models. This reinforces the idea that recurrent neural networks need larger-volume training and more frequent training data to exploit sequential dependencies optimally. Even with improved hyperparameters, these findings reaffirm that DL models are less suited to small-scale tabular financial data without additional feature engineering or architectural modifications. These findings are consistent with previous research by Manjula & Karthikeyan [9], who reported that the strong performance of GBM Machines in predicting gold prices

In terms of optimization strategy, as shown in Table 9, TPE generally provided the best balance of performance and efficiency across models. As a Bayesian optimization approach, TPE models the distribution of good versus and hyperparameter regions and efficiently balances exploration with exploitation. This sample efficiency is particularly useful in macroeconomic forecasting, where computational resources as well as data availability are constrained. CMA-ES, however, yielded the best results specifically for GBM. This can be explained by its suitability for complex, non-convex hyperparameter spaces. GBM's parameters, such as learning rate, estimators, and tree depth, interact in highly non-linear ways that create many local minima.

By exploring multiple regions of the search space simultaneously, the findings suggest that CMA-ES was able to escape poor local optima and converge toward global solutions, which may explain its advantage over the more locally focused TPE method in optimizing GBM on the Full Feature Set. By contrast, Grid Search was the least effective, likely due to its exhaustive nature and inefficiency in high-dimensional parameter spaces.

As observed in Table 9, the Selected Feature Set generally continued to perform better than the Full Feature Set in most of the optimized models. This trend is consistent with the baseline results, reinforcing the need to select features carefully in financial forecasting.

Table 9. Summary of Best Optimized Models

Model	Best Optimization	Feature Set	RMSE	MAE	R^2
GBM	CMA-ES	FF set	101.261	77.826	0.9972
XGBoost	TPE	SF set	107.500	88.746	0.9969
RF	TPE	FF set	115.282	92.988	0.9964
LightGBM	TPE	SF set	145.770	109.672	0.9907
GRU	TPE	SF set	403.791	271.436	0.9562
LSTM	TPE	SF set	589.389	263.189	0.9068

To sum up, these findings highlight the critical importance of hyperparameter optimization in financial forecasting. GBM provides a reliable forecasting tool that remains stable even when new indicators (e.g., geopolitical risk indices and global shocks) are added, making GBM a strong candidate for production forecasting systems. Next, the strong performance of ensemble models when compared to DL models highlights the importance of aligning algorithm choice with data availability. In emerging markets such as Malaysia, where the macroeconomic variables are monthly and relatively less dense, ensemble models are a more robust and interpretable forecasting solution. In contrast, the DL models could be used when they are provided with larger or higher-frequency datasets to maximize their potential. For policymakers or investors in Malaysia, it is more efficient to focus on a smaller number of selected macroeconomic indicators rather than tracking all the possible variables. This implies that even small but well-chosen datasets can yield robust gold price forecasting.

4-3-Actual vs. Predicted Gold Prices

Figure 2 illustrates the predicted gold prices versus actual gold prices, generated using the best-performing model (GBM with CMA-ES on Full Feature Set). The visual alignment confirms the model's ability to follow real-world trends accurately. The inclusion of the Malaysia Inflation Rate (MIFR) curve further supports the role of inflation in driving gold prices, consistent with the literature on gold as an inflation hedge. As inflation rises, both actual and predicted gold prices exhibit an upward trend, reinforcing the economic rationale that gold serves as a hedge against inflation.

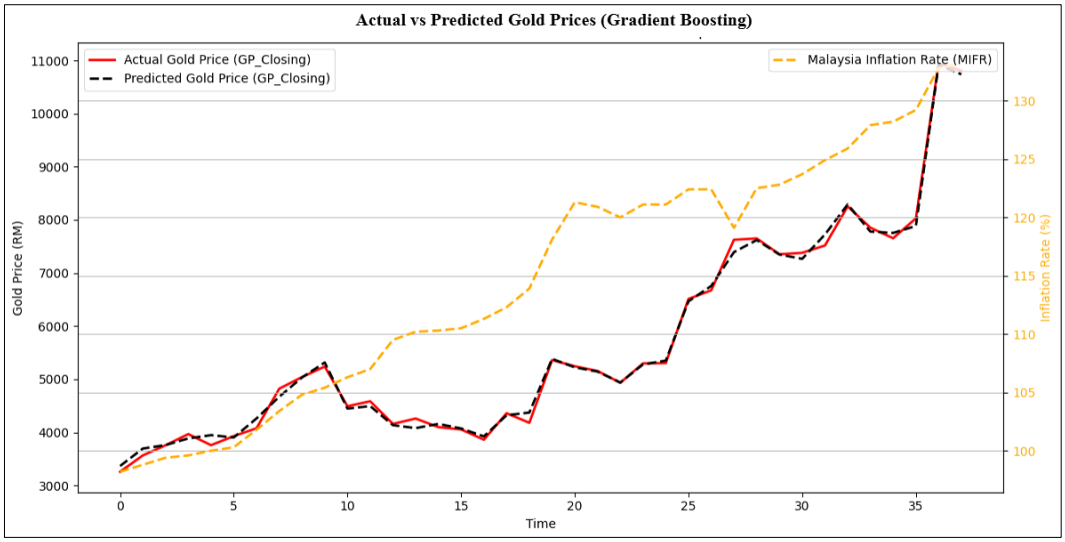

Figure 2. Actual vs. Predicted Gold Prices Using GBM (CMA-ES) and Overlaid MIFR

4-4- Feature Importance Analysis

Identifying the key economic factors that influence Malaysia's gold price is essential for investors and policymakers to make informed decisions about investment and trading strategies. Figure 3 presents the feature importance scores generated by the GBM model optimized with CMA-ES, using the Full Feature Set. The top five most influential features are GP_Low, GP_High, UIFR (U.S. inflation rate), MIFR (Malaysia inflation rate), and UITR (U.S. interest rate). The dominance of GP_Low and GP_High underscores the model's sensitivity to gold's internal pricing behavior. Meanwhile, the significant contributions of both domestic and international inflation and interest rates reflect gold's well-established role as a hedge against inflation and its responsiveness to global monetary policy conditions.

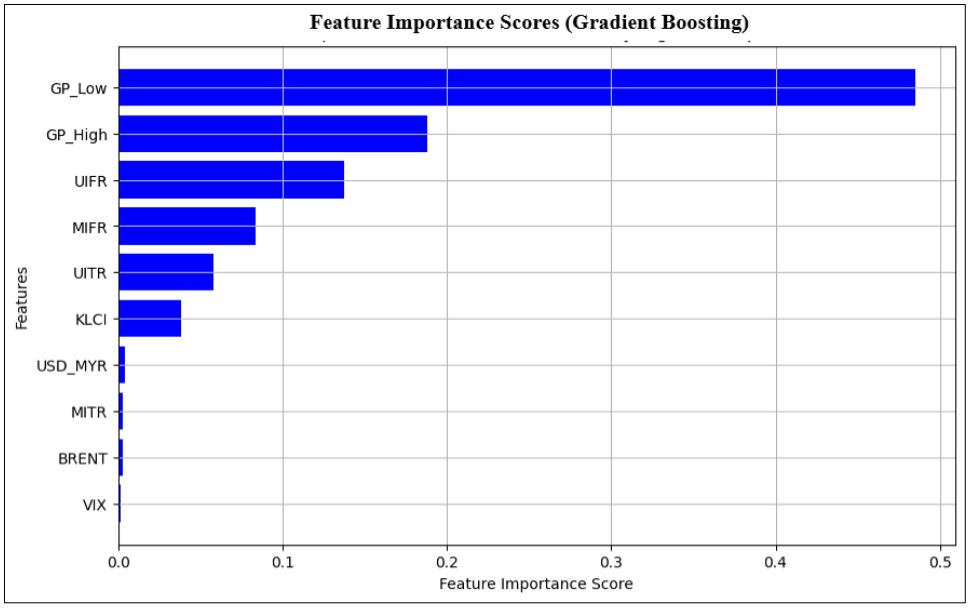

Figure 3. Feature Importance for GBM (Full Feature Set)

5- Discussion

The overall results demonstrate the superiority of ensemble models for predicting the Malaysian gold price using structured macroeconomic data. GBM consistently outperformed all other models in both baseline and optimized scenarios. DL models, although improved with tuning, remained less effective due to the dataset's size and structure. These findings are consistent with Rady et al. [6] and Manjula & Karthikeyan [9], where tree-based ensemble models performed better than traditional models in small-sample contexts. It is also in line with Jabeur et al. [10], who found that tree-based ensemble models outperformed neural networks in forecasting gold prices. However, unlike studies such as Primananda & Isa [12] and Ahmed et al. [54], where DL models are increasingly demonstrating superior performance in financial time series tasks with large-scale and high-frequency datasets, the monthly-frequency dataset and limited size in this study constrained DL effectiveness in capturing long-term temporal dependencies.

The improvement from Optuna optimization in our study mirrors results from Kausar et al. [39] and Hanifi et al. [55], which showed that Optuna's adaptive strategies generally outperform traditional optimization methods. Among optimization strategies, TPE offered consistent gains across models, while CMA-ES delivered the best single result. Grid Search, despite being exhaustive, was the least effective due to its inefficiency in high-dimensional search spaces.

Performance across feature sets showed that the Selected Feature Set generally yielded better results, likely due to noise reduction and improved generalization. An exception was GBM, which benefited from the richer Full Feature Set, highlighting its robustness in handling weakly correlated predictors.

The forecasted and actual gold prices (Figure 2) further illustrate that the GBM model optimized with CMA-ES is able to closely follow the observed price movements and aligns with inflationary trends. This consistency with gold's role as an inflation hedge indicates that the model is not only statistically accurate but also captures the underlying economic behavior.

In feature importance analysis (Figure 3), the prevalence of GP_Close and GP_High suggests the model's ability to incorporate gold's internal trading signals. At the same time, the strong contributions of inflation and interest rates validate gold's role as a hedge against purchasing power erosion and monetary uncertainty. These findings are in line with Md Hashim [22], who reported that inflation is the most influential factor in determining the Malaysian gold price volatility, and Ahmed et al. [56], who demonstrated that U.S. money policy shocks have significant spillover effects on emerging markets, which produces a stronger impact of U.S. interest rates and inflation on the asset prices. Nevertheless, the findings do not coincide with the results obtained by Leng et al. [57] and Shiva & Sethi [58], who reported the significant influences of equity indices on gold prices in Malaysia and India contexts, respectively. This divergence may stem from the use of monthly macroeconomic data, which emphasizes gold's macro hedge function rather than capturing short-term sentiment.

These findings, as summarized in Table 10, directly address the first, which concerns the ability of ensemble and deep learning models to capture static versus temporal relationships in gold price forecasting. The superior performance of ensemble models, such as GBM and XGBoost, suggests that static, cross-sectional relationships—where concurrent macroeconomic indicators explain gold prices—are the dominant patterns in this dataset. In contrast, the underperformance of LSTM and GRU, even after optimization, indicates that temporal dependencies (i.e., how past economic conditions influence future gold prices) are either weak or not effectively learnable given the monthly frequency and limited sample size. This reinforces the conclusion that, in small-scale, low-frequency macroeconomic datasets, ensemble models are more suitable than recurrent DL architectures for predictive forecasting tasks.

Table 10. Comparison between previous studies and the present study

Study	Data / Context	Models Used	Key Findings	Relation to Present Study
Rady et al. [6]	Monthly gold price data (362 data points)	RF, ARIMA, GBT, DT	Ensembles outperformed traditional models.	Consistent- our GBM is also superior in short/ medium horizons
Manjula & Karthikeyan [9]	Monthly gold price data (228 data points)	Linear Regression, RF, GB	Ensembles outperformed traditional models.	Consistent- our GBM is also superior in short/ medium horizons
Jabeur et al. [10]	Monthly gold price data (408 data points)	XGBoost, CatBoost, RF, LightGBM, NN, Linear Regression	Ensembles outperformed NN.	Matches our finding that ensembles outperform DL in small/medium samples
Primananda & Isa [12]	Daily gold price data (approximately 7305 data points)	LSTM, GRU	DL models performed well in large samples.	Differs- our monthly dataset limited the effectiveness of DL
Ahmed et al. [54]	Daily gold price data (10850 data points)	LSTM, ARIMA, Covariance Matrix Estimation, Deep Regression, SVR, CNN	DL models, particularly LSTM, outperformed the compared models.	Differs- our monthly dataset limited the effectiveness of DL
Kausar et al. [39], Hanifi et al. [55]	Financial time series forecasting	ML/ DL models with Optuna optimization	Optuna tuning enhanced forecasting performance.	Consistent- our study also found that Optuna improved model accuracy.
Md Hashim [22]	Malaysia, yearly data from 1980 to 2020	Regression (POLS)	Inflation significantly influences Malaysian gold prices.	Consistent- our feature importance also showed inflation as a dominant driver of Malaysian gold prices.
Ahmed et al. [56]	Emerging markets, macro-financial linkages	Econometric/ VAR analysis	U.S. monetary policy shocks have significant spillover effects on EMs, amplifying the effect of U.S. interest rates and inflation on asset prices.	Consistent- our feature importance also showed the strong impact of UIFR and UITR on Malaysian gold prices.
Leng et al. [57]	Malaysia, financial market context	Econometric analysis	Equity indices significantly influence Malaysian gold prices.	Differs- our study found weak equity effects.
Shiva & Sethi [58]	India, commodity and financial market context	Econometric/ Time-series analysis	Equity indices significantly influence gold prices in the Indian context	Differs- our study found weak equity effects

Overall, the discussion underscores that the models not only provide empirical answers to the research questions but also demonstrate their practical limitations. Results show the areas where ensemble approaches excel and where deep learning is limited in such a context. Based on these insights, the following sections summarize the contributions to the study and present broader implications, limitations, and future directions.

6- Conclusion

This study benchmarks the predictive performance of six machine learning models "four ensemble learners (Random Forest, Gradient Boosting, XGBoost, and LightGBM) and two deep learning models (LSTM and GRU)" for forecasting Malaysia's gold prices using monthly macroeconomic data from 2009 to 2024. The models were optimized using three strategies (TPE, Grid Search, and CMA-ES) in a unified Optuna framework. Hyperparameter tuning significantly improved predictive accuracy, with GBM optimized by CMA-ES achieving the highest performance, underscoring the strength of ensemble methods in capturing static, cross-sectional relationships. Although deep learning models also benefited from optimization, their performance remained constrained by the low-frequency, small-sample dataset. Feature importance analysis confirmed that internal gold price movements and macroeconomic indicators "particularly domestic and U.S. inflation and interest rates" were the most influential predictors, reaffirming gold's role as an inflation hedge.

Despite these contributions, several limitations remain. Validation protocols differed across model types, were appropriate within each paradigm, but not fully harmonized. Future research should incorporate stress-testing for structural breaks such as COVID-19 and post-2022 interest rate hikes, employ higher-frequency data where available, and extend forecasts to other safe-haven assets such as silver or sovereign bonds to assess generalizability. Hybrid ensemble–DL approaches could also be explored to leverage the strengths of both paradigms, potentially improving robustness across economic regimes. Furthermore, integrating sentiment indicators from news or social media at monthly horizons could capture behavioral drivers of gold prices often missed by macroeconomic variables. Expanding feature importance analysis across multiple models would further enhance interpretability, consistency, and transparency, providing deeper insights for policymakers, investors, and researchers alike.

7- Declarations

7-1- Author Contributions

Conceptualization, J.S.P.S. and K.S.M.A.; methodology, J.S.P.S. and R.K.K.; validation, R.K.K. and J.S.P.S.; formal analysis, R.K.K. and K.S.M.A.; investigation, J.S.P.S.; resources, K.S.M.A.; data curation, J.S.P.S.; writing—original draft preparation, J.S.P.S.; writing—review and editing, K.S.M.A. and R.K.K.; supervision, K.S.M.A. and R.K.K. All authors have read and agreed to the published version of the manuscript.

7-2- Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7-3- Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

7-4- Institutional Review Board Statement

Not applicable.

7-5- Informed Consent Statement

Not applicable.

7-6- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

8- References

[1] Baur, D. G., & Lucey, B. M. (2010). Is gold a hedge or a safe haven? An analysis of stocks, bonds and gold. Financial Review, 45(2), 217–229. doi:10.1111/j.1540-6288.2010.00244.x.

[2] World Gold Council. (2024). Gold Demand Trends: Full Year 2024. World Gold Council, London, United Kingdom.

[3] Bildirici, M. E., & Sonustun, F. O. (2018). The effects of oil and gold prices on oil-exporting countries. Energy Strategy Reviews, 22, 290–302. doi:10.1016/j.esr.2018.10.004.

[4] Ibrahim, M. H. (2012). Financial market risk and gold investment in an emerging market: the case of Malaysia. International Journal of Islamic and Middle Eastern Finance and Management, 5(1), 25–34. doi:10.1108/17538391211216802.

[5] Bin Sukri, M. K. A., & Mohd Zain, N. H. (2015). The relationship between selected macroeconomic factors and gold price in Malaysia. International Journal of Business, Economics and Law, 8(1), 182-193.

[6] Rady, E. H. A., Fawzy, H., & Fattah, A. M. A. (2021). Time series forecasting using tree based methods. Journal of Statistics Applications & Probability, 10(1), 229–244. doi:10.18576/JSAP/100121.

[7] Siami-Namini, S., Tavakoli, N., & Siami Namin, A. (2018). A Comparison of ARIMA and LSTM in Forecasting Time Series. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 1394–1401. doi:10.1109/ICMLA.2018.00227.

[8] Anbananthen, S. K., Sainarayanan, G., Chekima, A., & Teo, J. (2006). Data Mining using Pruned Artificial Neural Network Tree (ANNT). 2nd International Conference on Information and Communication Technologies ICTTA 2006, 1350–1356. doi:10.1109/ictta.2006.1684577.

[9] Manjula, K. A., & Karthikeyan, P. (2019). Gold Price Prediction using Ensemble based Machine Learning Techniques. 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 1360–1364. doi:10.1109/icoei.2019.8862557.

[10] Jabeur, S. Ben, Mefteh-Wali, S., & Viviani, J. L. (2024). Forecasting gold price with the XGBoost algorithm and SHAP interaction values. Annals of Operations Research, 334(1–3), 679–699. doi:10.1007/s10479-021-04187-w.

[11] Mohtasham Khani, M., Vahidnia, S., & Abbasi, A. (2021). A Deep Learning-Based Method for Forecasting Gold Price with Respect to Pandemics. SN Computer Science, 2(4), 335. doi:10.1007/s42979-021-00724-3.

[12] Primananda, S. B., & Isa, S. M. (2021). Forecasting Gold Price in Rupiah using Multivariate Analysis with LSTM and GRU Neural Networks. Advances in Science, Technology and Engineering Systems Journal, 6(2), 245–253. doi:10.25046/aj060227.

[13] He, Z., Zhou, J., Dai, H.-N., & Wang, H. (2019). Gold Price Forecast Based on LSTM-CNN Model. 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), 1046–1053. doi:10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00188.

[14] Liang, Y., Lin, Y., & Lu, Q. (2022). Forecasting gold price using a novel hybrid model with ICEEMDAN and LSTM-CNN-CBAM. Expert Systems with Applications, 206, 117847. doi:10.1016/j.eswa.2022.117847.

[15] Taneva-Angelova, G., Raychev, S., & Ilieva, G. (2025). A Framework for Gold Price Prediction Combining Classical and Intelligent Methods with Financial, Economic, and Sentiment Data Fusion. International Journal of Financial Studies, 13(2), 102. doi:10.3390/ijfs13020102.

[16] Anbananthen, K. S. M., & Elyasir, A. M. H. (2013). Evolution of opinion mining. Australian Journal of Basic and Applied Sciences, 7(6), 359-370.

[17] Hajek, P., & Novotny, J. (2022). Fuzzy Rule-Based Prediction of Gold Prices using News Affect. Expert Systems with Applications, 193, 116487. doi:10.1016/j.eswa.2021.116487.

[18] Kostyrin, E., & Drynkin, S. (2025). Innovative Approach to the Optimal Distribution of Citizens' Pension Savings to Non-State Pension Funds. Emerging Science Journal, 9(1), 504–523. doi:10.28991/ESJ-2025-09-01-028.

[19] Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). Automated machine learning: methods, systems, challenges. Springer, Cham, Switzerland. doi:10.1007/978-3-030-05318-5.

[20] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2623–2631. doi:10.1145/3292500.3330701.

[21] Hashim, S. L., Ramlan, H., Razali, N. H., & Nordin, N. Z. (2017). Macroeconomic variables affecting the volatility of gold price. Journal of Global Business and Social Entrepreneurship (GBSE), 3(5), 97-106.

[22] Md Hashim, S. L. (2022). Analysis on Factors Influence the Price of Gold in Malaysia. Advanced International Journal of Business, Entrepreneurship and SMEs, 4(11), 16–22. doi:10.35631/aijbes.411002.

[23] Mohd Nasir, F. Z., Wan Zakaria, W. M. F., Musa, M. H., Burhanuddin, M. A., & Azman Ong, M. H. (2018). Economic forces on gold price in Malaysia. e-Academia Journal, 7(2), 54-65.

[24] Mainal, S. A., Mohd Selamat, A. H., Abd Majid, N. D. S., & Noorzee, K. N. I. (2023). Factors Influencing the Price of Gold in Malaysia. Information Management and Business Review, 15(3(I)), 195–205. doi:10.22610/imbr.v15i3(i).3529.

[25] Sailaja, V. N., Kumar, A., & VS, P. K. (2022). A study on macro-economic variables and their impact on gold price in India. Academy of Marketing Studies Journal, 26(5), 1-16.

[26] Cohen, G., & Aiche, A. (2023). Forecasting gold price using machine learning methodologies. Chaos, Solitons & Fractals, 175, 114079. doi:10.1016/j.chaos.2023.114079.

[27] Cologni, A., & Manera, M. (2008). Oil prices, inflation and interest rates in a structural cointegrated VAR model for the G-7 countries. Energy Economics, 30(3), 856–888. doi:10.1016/j.eneco.2006.11.001.

[28] Guo, C., Luo, Y., Qin, B., & Ge, X. (2025). Enhancing Gold Price Forecasting through Decomposition Algorithms and Deep Learning: A Comprehensive Comparative Analysis and Hybrid Model Matching Approach. Proceedings of 2025 4th International Conference on Cyber Security, Artificial Intelligence and the Digital Economy, CSAIDE 2025, 355–361. doi:10.1145/3729706.3729762.

[29] Das, S., Sahu, T. P., & Janghel, R. R. (2022). Oil and gold price prediction using optimized fuzzy inference system based extreme learning machine. Resources Policy, 79, 103109. doi:10.1016/j.resourpol.2022.103109.

[30] Liu, Y., Zhang, Y., & Peng, X. (2024). Textual analysis and gold futures price forecasting: Evidence from the Chinese market. Finance Research Letters, 69, 106116. doi:10.1016/j.frl.2024.106116.

[31] Salim, M., & Djunaidy, A. (2024). Development of a CNN-LSTM Approach with Images as Time-Series Data Representation for Predicting Gold Prices. Procedia Computer Science, 234, 333–340. doi:10.1016/j.procs.2024.03.007.

[32] Anbananthen, K. S. M., Busst, M. B. M. A., Kannan, R., & Kannan, S. (2023). A Comparative Performance Analysis of Hybrid and Classical Machine Learning Method in Predicting Diabetes. Emerging Science Journal, 7(1), 102–115. doi:10.28991/ESJ-2023-07-01-08.

[33] Makala, D., & Li, Z. (2021). Prediction of gold price with ARIMA and SVM. Journal of Physics: Conference Series, 1767(1), 012022. doi:10.1088/1742-6596/1767/1/012022.

[34] Narendran, S. (2024). Exploring the Efficacy of Various Machine Learning Approaches for Gold Price Forecasting: Insights and Analysis. 2024 7th International Conference on Circuit Power and Computing Technologies (ICCPCT), 1319–1324. doi:10.1109/ICCPCT61902.2024.10673291.

[35] Zangana, H. M., & Obeyd, S. R. (2024). Deep Learning-based Gold Price Prediction: A Novel Approach using Time Series Analysis. Sistemasi, 13(6), 2581. doi:10.32520/stmsi.v13i6.4651.

[36] Dewi, N. P. J. R., Ginantra, N. L. W. S. R., Darma, I. W. A. S., & Indrawan, I. G. A. (2023). Application of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) Algorithm in Gold Price Prediction. 2023 11th International Conference on Cyber and IT Service Management, CITSM 2023, 1–6. doi:10.1109/CITSM60085.2023.10455749.

[37] Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. Advances in Neural Information Processing Systems (NIPS 2011), 24.

[38] Kee, T., & Ho, W. K. O. (2025). Optimizing Machine Learning Models for Urban Sciences: A Comparative Analysis of Hyperparameter Tuning Methods. Urban Science, 9(9), 348. doi:10.3390/urbansci9090348.

[39] Kausar, R., Iqbal, F., Raziq, A., Sheikh, N., & Rehman, A. (2024). Enhanced Foreign Exchange Volatility Forecasting using CEEMDAN with Optuna-Optimized Ensemble Deep Learning Model. Sains Malaysiana, 53(9), 3229–3239. doi:10.17576/jsm-2024-5309-25.

[40] Lai, L. H., Lin, Y. L., Liu, Y. H., Lai, J. P., Yang, W. C., Hou, H. P., & Pai, P. F. (2024). The Use of Machine Learning Models with Optuna in Disease Prediction. Electronics (Switzerland), 13(23), 4775. doi:10.3390/electronics13234775.

[41] Pokhrel, P., & Lazar, A. (2023). A Comparison of AutoML Hyperparameter Optimization Tools for Tabular Data. The International FLAIRS Conference Proceedings, 36. doi:10.32473/flairs.36.133357.

[42] Shekhar, S., Bansode, A., & Salim, A. (2021). A Comparative study of Hyper-Parameter Optimization Tools. 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 1–6. doi:10.1109/CSDE53843.2021.9718485.

[43] Lindauer, M., Eggensperger, K., Feurer, M., Biedenkapp, A., Deng, D., Benjamins, C., Ruhkopf, T., Sass, R., & Hutter, F. (2022). SMAC3: A Versatile Bayesian Optimization Package for Hyperparameter Optimization. Journal of Machine Learning Research, 23(54), 1–9.

[44] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. doi:10.1023/A:1010933404324.

[45] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780. doi:10.1162/neco.1997.8.8.1735.

[46] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv Preprint, arXiv:1412.3555. doi:10.48550/arXiv:1412.3555.

[47] Mahmud Sujon, K., Binti Hassan, R., Tusnia Towshi, Z., Othman, M. A., Abdus Samad, M., & Choi, K. (2024). When to Use Standardization and Normalization: Empirical Evidence from Machine Learning Models and XAI. IEEE Access, 12, 135300–135314. doi:10.1109/ACCESS.2024.3462434.

[48] Potoski, M. (2013). Predicting gold prices. CS 229: Machine Learning Final Projects, Stanford University, Stanford, United States.

[49] Kumar, R., Moolchandani, J., Shukla, A., Sahu, S., Thada, V., & Chole, V. (2024). Machine Learning-Based Prediction of Gold Prices Using Economic Indicators. 2024 13th International Conference on System Modeling & Advancement in Research Trends (SMART), 520–524. doi:10.1109/SMART63812.2024.10882507.

[50] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence (IJCAI), 14(2), 1137–1145.

[51] Brownlee, J. (2017). Long short-term memory networks with python: develop sequence prediction models with deep learning. Machine Learning Mastery, Vermont, Australia.

[52] Bailly, A., Blanc, C., Francis, É., Guillotin, T., Jamal, F., Wakim, B., & Roy, P. (2022). Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. Computer Methods and Programs in Biomedicine, 213, 106504. doi:10.1016/j.cmpb.2021.106504.

[53] Greff, K., Srivastava, R. K., Koutnik, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A Search Space Odyssey. IEEE Transactions on Neural Networks and Learning Systems, 28(10), 2222–2232. doi:10.1109/TNNLS.2016.2582924.

[54] Ahmed, S., Akinci, O., & Queralto, A. (2024). U.S. Monetary Spillovers to Emerging Markets: Both Policy Drivers and Vulnerabilities Matter. doi:10.2139/ssrn.4973935.

[55] Hanifi, S., Cammarono, A., & Zare-Behtash, H. (2024). Advanced hyperparameter optimization of deep learning models for wind power prediction. Renewable Energy, 221, 119700. doi:10.1016/j.renene.2023.119700.

[56] Ahmed, S., Akinci, O., & Queralto, A. (2024). US Monetary Policy Spillovers to Emerging Markets: Both Shocks and Vulnerabilities Matter. FRB of New York Staff Report, 1321. doi:10.17016/IFDP.2021.1321r1.

[57] Leng, C. T., Yi, L. J., Woh, L. K., & Cheong, S. (2020). Asymmetric volatility spillover between oil market, gold market and Malaysian stock market. Bachelor of Finance, Universiti Tunku Abdul Rahman, Petaling Jaya, Malaysia.

[58] Shiva, A., & Sethi, M. (2015). Understanding Dynamic Relationship among Gold Price, Exchange Rate and Stock Markets: Evidence in Indian Context. Global Business Review, 16(5_SUPPL), 93–111. doi:10.1177/0972150915601257.