A DUAL METHODS APPROACH TO CRUDE PALM OIL PRICE FORECASTING IN MALAYSIA: INSIGHTS FROM ARDL AND LSTM

Mohd Shahrin Bahar
Faculty of Business and Management, Universiti Teknologi MARA Shah Alam Campus
Imbarine Bujang
Faculty of Business and Management Universiti Teknologi MARA Sabah Branch Kota Kinabalu Campus
Abdul Aziz Karia
Faculty of Business and Management Universiti Teknologi MARA Sabah Branch Kota Kinabalu Campus
Nur Zahidah Bahrudin
Faculty of Business and Management, Universiti Teknologi MARA Puncak Alam Campus

Abstract: *Understanding the volatile nature of palm oil prices is crucial due to its significant implications for the economy and the market. Due to its complexity, the central issue of the rise in palm oil price determinants and forecasting depends on various market demand and supply forces. However, many scholars fail to conclude that the factor drives palm oil prices. This study examines the factors affecting Malaysian Crude Palm Oil (CPO) pricing dynamics and uses estimated palm oil prices in forecasting. Using data from the Malaysian Palm Oil Board, spanning January 2004 to December 2021. Methodologically, we employed Autoregressive Distributed Lag (ARDL) and Long Short-Term Memory (LSTM) models to evaluate and forecast CPO prices. Our findings revealed that the LSTM model outperformed the ARDL model in forecasting accuracy. Notably, the LSTM model was more effective with a selection of ten independent variables identified through LASSO and SHAP estimation, compared to using either eleven or four variables based on ARDL regression results. The analysis highlights the significant influence of weather conditions and macroeconomic factors, particularly tax rates, on CPO prices. The findings enhance understanding of market dynamics and assist in accurate forecasting of CPO prices.*

Keywords: *Forecasting, CPO Prices, ARDL, and LSTM*

1. INTRODUCTION

Understanding Crude Palm Oil's (CPO) pricing patterns is essential due to its significant impact on numerous sectors. Analysing these price changes is crucial for stakeholders, including government bodies, farmers, investors, and palm oil production companies, as it aids in maintaining an equilibrium between supply and demand in the broader market where palm oil is a key commodity (Oosterveer, 2015; Isa et al., 2016; Murphy et al., 2021).

Researchers have highlighted the value of understanding core economic principles to grasp price fluctuations. The Cobweb Theorem, described by Pashigian (2008), explains the cyclical nature of supply and demand in markets with production response delays. This theory helps understand market price fluctuations resulting from disparities in production and consumption. Such insights enable the development of effective strategies informed by historical data to modify supply and demand patterns. Additionally, price determination

varies across different markets. In stock markets, prices are set through buyer and seller interactions, reflecting their perceptions and reactions to market conditions (Peterson, 2014; Ma et al., 2021). Conversely, commodity pricing, including CPO, involves diverse methods such as spot markets, futures markets, and direct transactions.

Comprehending CPO pricing is complex, requiring consideration of various factors like currency value fluctuations, trade restrictions, weather conditions, population growth, and the cost of other commodities (Chandrarin et al., 2022; Cespedes & Velasco, 2012; Enghiad, Headey & Fan, 2008; Wilson & Cacho, 2007 Zainalabidin & Rahim, 2012; Zaidi et al., 2022). The demand dynamics are also significantly influenced by the pricing of alternative vegetable oils and broader economic patterns. This complexity hinders accurate forecasting of future trends, affecting the reliability of predictions.

Given this context, accurate models for predicting CPO prices must account for the simultaneous changes in these multifaceted determinants. Moreover, unforeseen events or changes in any single factor can lead to significant deviations from expected prices. This underscores the need for comprehensive models that incorporate a wide range of variables and are adaptable to sudden shifts in market dynamics. The challenge lies in creating predictive models that can navigate the complexity of the factors at play and provide reliable forecasts in the face of uncertainty.

Recent discourse has increasingly focused on comparing traditional statistical methods like the ARDL approach, which explores complex correlations between various parameters and CPO prices, with modern AI techniques like the LSTM model.

The ARDL approach is pivotal for identifying significant variables that influence CPO prices. The analysis investigates the short-term and long-term connections between CPO prices and independent factors. The study used the LSTM model, renowned for managing complex temporal relationships in time-series data, to predict future CPO prices after identifying key factors. The LSTM model's capacity to learn from sequences of data and their underlying patterns makes it an exceptional tool for forecasting in scenarios characterised by intricate interactions of various factors over time. By focusing on the most influential factors identified by the ARDL model, the LSTM model can potentially enhance the precision of forecasts. This synergistic approach between the ARDL and LSTM models allows for a more refined analysis, where the ARDL model's strength in variable identification lays the groundwork for the LSTM model to leverage this information in making accurate future CPO price predictions (Hamid & Shabri, 2017; Ofuoku & Ngniatedema, 2022). Hence, this study aims to contrast the ARDL methodology with the LSTM model, an AI-based technique that detects intricate data patterns over time. Our research focuses on selectively choosing factors crucial for improving forecasting models' accuracy. The procedure is improved by using the Least Absolute Shrinkage and Selection Operator (LASSO) test in the ARDL framework to identify a subset of important variables. Such meticulous variable selection showcases the importance of precise variable identification in navigating the complex dynamics of CPO pricing, aiming to achieve the highest precision in prediction outcomes and offer strategic insights for decision-making in the CPO sector and related industries (Lu et al., 2021).

In summary, this research aims to forge a path to the most accurate CPO price predictions by synergising the ARDL approach's variable selection process with the LSTM model's advanced pattern recognition capabilities. Our research seeks to enhance understanding of the determinants of CPO, forecasting pricing trends, and guiding the industry towards a promising future. By combining established economic concepts with

advanced statistical and AI forecasting methods, this study aims to comprehensively understand the CPO market's intricacies, contributing to a dynamic and informed future in the sector.

2. LITERATURE REVIEW

Palm oil is an agricultural product in Malaysia, substantially contributing to the country's economy. However, the palm oil industry faces challenges, including fluctuating pricing, changing consumer preferences, and environmental concerns. Understanding the factors influencing short- and long-term palm oil pricing in Malaysia is crucial to address these issues. This literature review aims to provide an overview of the available research on the determinants of Malaysian CPO prices.

The discovery of prices plays a fundamental role in the functioning of financial markets, where market players determine the pricing of assets such as stocks, bonds, and commodities based on supply and demand dynamics (Tomek & Robinson, 1981; Netayarak, 2007). This process ensures that prices accurately represent the underlying value of the traded items. The price discovery mechanism varies depending on the market and asset being traded. For example, in the case of stocks, price discovery occurs through transactions between buyers and sellers on the stock market, reflecting their valuations and prevailing market conditions. Similarly, price discovery occurs through spot markets, futures markets, and over-the-counter transactions for commodities like CPO. Buyers and sellers negotiate prices based on current supply and demand, production costs, and broader economic conditions.

CPO prices are influenced by the interplay of supply and demand forces at both the local and international levels. Key factors that influence CPO prices include production and stock levels of CPO and its alternatives (e.g., soybean oil), exchange rates (as a measure of competitiveness and trade), and global economic uncertainties (Putri et al., 2019; Zaidon & Karim, 2019). Economic considerations significantly affect Malaysian CPO prices, as the demand for CPO is influenced by global economic indicators such as exchange rates, GDP, and the Consumer Price Index (CPI) (Putri et al., 2019; Zaidon & Karim, 2019).

CPO export taxation also plays a crucial role in price dynamics. Research has explored the effects of export taxes on CPO prices, with findings suggesting that such taxes may impact the competitiveness of CPOs on the global market, leading to price fluctuations (Hisham et al., 2019). For instance, implementing an import tax on CPOs in Indonesia decreased demand and price declines, which had significant implications for the CPO industry (Hisham et al., 2019). Conversely, reductions in CPO export taxes in Malaysia led to increased demand and price rises, fostering economic benefits (Amin et al., 2019). Moreover, CPO production levels, particularly in Indonesia, have influenced worldwide CPO price changes, reflecting the global impact of production trends. A rise in CPO prices may also adversely affect the price of soybean oil and export and production in Malaysia (Hassan & Balu, 2016).

In light of CPO's economic significance, accurate forecasting of CPO prices is essential in both the agricultural and financial sectors. Recent research has explored the ARDL models and LSTM neural networks to forecast and interpret multivariate CPO price predictions. ARDL The Autoregressive Distributed Lag (ARDL) approach to forecasting has gained prominence for its versatility in econometric analysis. This model, adept at capturing both short-term dynamics and long-term equilibrium relationships, is particularly beneficial in financial and economic forecasting due to its flexibility with variables of different integration orders (Pesaran et al., 2001). A significant advantage of ARDL lies in

its robustness in small sample sizes, making it a preferred choice in studies with limited data (Pesaran & Shin, 1999). However, challenges arise in model complexity and sensitivity to specification, necessitating careful selection of variables and lag lengths (Pesaran et al., 2001). Recent advancements integrate ARDL with machine learning techniques, enhancing forecasting accuracy and model robustness (Du et al., 2020).

Studies have highlighted the potential of LSTM models in analysing complex relationships in multivariate time series data (Sagheer, A., & Kotb, M.; 2019), making it a suitable method for forecasting CPO prices influenced by various economic and environmental factors. Furthermore, multivariate LSTM models have been utilised in financial time series prediction, addressing the challenges posed by global market dynamics and macroeconomic variables that impact CPO prices (Widiputra et al., 2021; Urolagin et al., 2021). The applicability of LSTM models in long-term forecasting aligns well with the dynamic nature of CPO prices, where long-term trends and various influencing factors play significant roles in determining its price (Althelaya et al., 2018). The effectiveness of LSTM-FCNs in multivariate time series classification has also been emphasised, holding promise in classifying different CPO price patterns and identifying potential market trends (Karim et al., 2019).

In conclusion, the ARDL and LSTM models offer the potential for enhanced forecasting of CPO prices, enabling more accurate and interpretable predictions. Understanding the factors influencing CPO pricing is critical for navigating the complexities of the market. Further research to optimise the integration of ARDL and LSTM models for CPO price forecasting is expected to yield even more accurate and insightful results, supporting better decision-making for investors and policymakers in the agricultural and financial sectors. Combining economic factors and advanced forecasting techniques is critical to unlocking a comprehensive understanding of CPO price movements supporting the sustainable growth of the CPO industry in Malaysia and beyond.

3. DATA AND METHODOLOGY

This study utilises monthly frequency data from January 2004 to December 2021, aiming to develop an Autoregressive Distributed Lag (ARDL) model incorporating eleven variables, with the Crude Palm Oil (CPO) price as the dependent variable. The independent variables in the model include the Export of CPO, CPO Production, CPO Export Tax, Stock of CPO, Rainfall (representing weather), Population, Economic Growth, Global Consumption, Price of Soybean, Price of Sunflower, Exchange Rate, and the Consumer Price Index. Data was obtained from reputable institutions like the Malaysian Palm Oil Board, Refinitiv, and Bloomberg. These sources were selected for their comprehensive and precise data on the relevant variables, enhancing our research conclusions' robustness and credibility.

In addition to the traditional ARDL model, this research introduces the Long Short-Term Memory (LSTM) forecasting method. LSTM, a recurrent neural network (RNN), is particularly adept at handling time series data. Its ability to capture temporal dependencies and patterns in the dataset makes it a valuable tool for our analysis. By employing the ARDL model and LSTM, the study seeks to explore complex relationships within the data and potentially improve the accuracy of predictions regarding CPO prices. This dual-model approach allows for a more nuanced understanding of the factors influencing CPO pricing, contributing significantly to the field. ARDL forecasting uses Eviews to analyse and forecast the variables, and LSTM forecasting uses Python.

Table 1: Variables Description

Notation	Variable	Description
Y	Palm Oil Price	Determined by supply-demand factors, production costs, market conditions, and trade dynamics.
X1	Palm Oil Export	The volume of international sales is driven by global demand, trade policies, and market competition.
X2	Palm Oil Production	Influenced by factors such as weather, technology, pests, and government policies.
X3	Palm Oil Stock	Based on production, trade volumes, market demand, and storage capacities.
X4	Tax Rate	Rate affecting palm oil's profitability, price, and market competition.
X5	Weather	Climatic conditions influencing production, yield, and pest incidence.
X6	Population	Determines demand based on size and consumption patterns.
X7	Soybean Price	Affects palm oil pricing as a competing food and industrial ingredient.
X71	Sunflower Price	Influences palm oil pricing as an alternative oilseed.
X8	Economic growth	Impacts price through income, spending, and trade dynamics.
X9	Exchange Rate	Affects global competitiveness and trade pricing.
X10	Consumer Price Index	Indicator of inflation impacting production costs and purchasing power.

3.1. ARLD model

As mentioned by Pasaran, Shin and Smith (2001), the advantage of ARLD is that the variables can be estimated with the combination of I (0) and I (1) series at the same time, with the single equations setup, that makes it simple to implement and interpret. This paper stabilises the variance of the series of all variables transformed into logarithmic form. Below is the ARDL model that used in this study:

$$\Delta Y = \alpha_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \lambda D + \sum(\delta_1 \Delta X_1 + \delta_2 \Delta X_2 + \ldots + \delta_n \Delta X_n) + \varepsilon$$

Where ΔY represents the differenced dependent variable (palm oil price), X_1, X_2, ..., X_n denote the independent variables (Export of CPO, CPO production, CPO Export Tax, Stock of CPO, weather (Rainfall), Population, Economic Growth, World consumption, Price of Soybean, Price of Sunflower, Exchange rate, and Consumer Price Index), α_0 is the short-run intercept term, β_1, β_2, ..., β_n represent the coefficients corresponding to the long-run relationship between the independent variables, λ represents the coefficient of the lagged dependent variable (D), which captures the speed of adjustment towards the long-run equilibrium, δ_1, δ_2, ..., δ_n represent the coefficients corresponding to the short- run relationship between the differenced independent variables, D represents the lagged dependent variable, and ε is the error term.

ARDL models are versatile statistical tools for analysing and forecasting complex interactions among multiple time series variables (Pasaran et al., 2001; Banerjee et al.,1993) concisely offer the following capabilities:

i. Capture Dynamics: They can simultaneously model short-term fluctuations and long-term relationships between variables.
ii. Handle Non-stationarity: These models are adept at dealing with variables that are not stationary, making them highly flexible for various types of data.
iii. Cointegration Analysis: Multivariate ARDL models can test and estimate long-term equilibrium relationships among variables.
iv. Estimate Impacts: They allow for estimating immediate and long-term effects of changes in one variable on others.
v. Complex Interactions: The models can incorporate multiple variables to analyse complex interactions and feedback mechanisms.
vi. Forecasting: They are powerful in forecasting future values based on historical data and variable interactions.
vii. Robust and Flexible: These models provide robust results across different specifications and offer flexibility in including lagged variables.

3.2. Diagnostic tests

The stability of the ARDL model was ensured through a comprehensive evaluation. For initial stability, three tests were applied: Augmented Dickey-Fuller (ADF) identified unit roots, Kwiatkowski-Phillips-Schmidt-Shin (KPSS) assessed stationarity, and Zivot-Andrews (ZA) detected structural breaks. Diagnostic tests were conducted, including Breusch-Godfrey LM for serial correlation, Cameron & Trivedi's IM-test for heteroskedasticity, and Skewness/Kurtosis tests for normality. Distribution was confirmed using techniques like normal probability plots. Meanwhile, structural stability was evaluated using recursive residuals' cumulative sum (CUSUM).

3.3. Long Short-Term Memory (LSTM) Model:

The LSTM model equation for CPO prices (Y) and the independent variables (X1, X2, X3, X4, X5, X6, X7, X71, X8, X9, X10) can be expressed as follows:

$$h_t = LSTM([X_{1t}, X_{2t}, \ldots, X\_q_t], h_{t-1})$$

h_t is the hidden state (output) of the LSTM at time t, and h_{t-1} is the hidden state (output) of the LSTM at the previous time step (t-1). [X_{1t}, X_{2t}, ..., X_q_t] represents the input data, which includes all independent variables (X1, X2, X3, X4, X5, X6, X7, X71, X8, X9, X10) at time t. h_t denotes the input data, which includes all independent variables at time t.

In practice, the LSTM model architecture can be designed with multiple LSTM cells, dropout layers, and possibly other recurrent or dense layers to capture the temporal patterns and dependencies in the multivariate time series data.

Multivariate Long Short-Term Memory (LSTM) models are advanced neural networks designed for forecasting tasks involving multiple interacting time series (Hochreiter, S., & Schmidhuber, J., 1997). The key capabilities include follows:

i. Handling Sequential and Multivariate Data: They excel at processing time series data with multiple input variables, capturing the dynamic interactions among them.
ii. Modelling Long-term Dependencies: LSTMs can remember and utilise long-term historical information, which is crucial for predicting future trends based on past data.

iii. Learning Non-linear Relationships: These models are adept at identifying complex, non-linear patterns in data, surpassing traditional linear models in performance.
iv. Multi-step Forecasting: They can simultaneously predict several future time steps, which are practical for short-term and long-term forecasting needs.
v. Automatic Feature Learning: LSTMs automatically learn relevant features from raw data, minimising the need for manual feature engineering.
vi. Robustness to Missing Data: Their recurrent structure makes them relatively resilient to gaps or missing values in time series data.

3.4. Feature Selection

In our study, not all factor data were significantly associated with CPO prices. To address this, we initially utilised the LASSO method to remove irrelevant factors. This helped us obtain a more focused set of essential variables for our analysis. To further refine our selection, we employed Random Forest to rank the feature importance, enabling us to identify the most influential factors.

LASSO's core principle is to compress irrelevant variables' coefficients to zero in the regression problem (Tibshirani, 1996). The regression problem in this study is expressed as follows:

$$y_i = \omega^T x_i + b$$

$$J(w) = \frac{1}{m}\sum_{i=}^{m}(y_i - \omega^T x_i)^2 + \lambda \sum_{i=1}^{m}|w_i$$

Where y_i, x_i and ω^T represent the monthly CPO prices, factors, and regression coefficients. The cost function J(w) An evaluation is conducted to assess the regression model's accuracy level. Finding the λ ¬ that minimises the value of J(w).

3.5. Division of train set and test set.

For all models, 216 months of data from January 2004 to December 2019 were utilised as the training set. The test set included data from January 2020 to December 2021. It was used to verify predictions made 6, 12, and 24 months ahead, starting from January 2020.

3.6. Model Accuracy Test

The model selection involved using four error measures to evaluate the accuracy of the employed models. In particular, we adopted two scale-dependent metrics, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are used with two scale-independent metrics: Mean Absolute Percentage Error (MAPE) and symmetric Mean Absolute Percentage Error (sMAPE).

4. RESULT

4.1. Variables selection

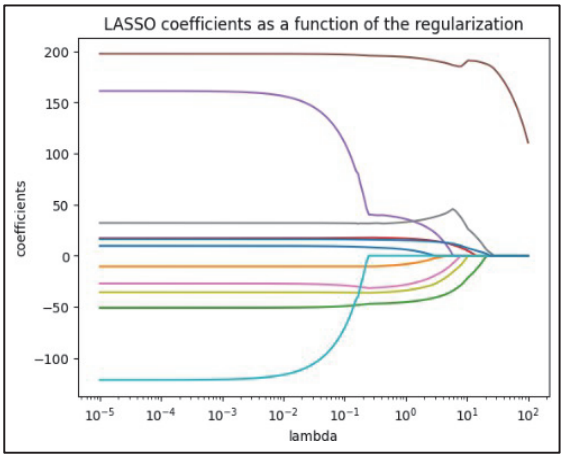
This research study explores the efficacy of Multivariate Autoregressive Distributed Lag (ARDL) and Long Short-Term Memory (LSTM) methods in forecasting Crude Palm Oil (CPO) prices, incorporating a set of 11 macroeconomic variables. Initial analyses involved conducting stationarity tests for each variable to determine their integrated order. Variables Y, X2, and X4 were found to be integrated of order 1 (I(1)), whereas variables X1, X3, X5, X6, X7, X71, X8, X9, and X10 exhibited stationarity at level (I(0)). This

diversity in integration orders indicates the presence of both short-term dynamics and long-term relationships within the model.

Following the stationarity identification, we selected relevant factors based on the ARDL model, specifying the data using the Akaike Information Criterion (AIC) and considering a maximum of four lags for the dependent and independent variables. The lag structure for each variable was determined as ARDL (2, 4, 0, 2, 0, 0, 2, 0, 3, 0, 0, 0) for the 10-variable model and ARDL (2, 0, 2, 0, 3) for the 4-variable model.

Further refinement is involved by employing the Least Absolute Shrinkage and Selection Operator (LASSO) test, aiding in identifying the most influential variables for accurate forecasting. The LASSO test excluded one variable from the initial set, resulting in a refined set of 10 critical factors. The coefficient compression and its impact on model performance are illustrated in Figure 1.

Figure 1: *LASSO Coefficients*



Our analysis encompassed three scenarios presented in Table 3: utilising all eleven independent variables (IV), ten essential variables identified by the LASSO test, and four deemed significant from the ARDL analysis. This approach aimed to enhance forecasting accuracy by eliminating less impactful variables and mitigating overfitting concerns.

Figure 2: *MSE vs Log10(Lambda) with Standard Errors – Lasso*

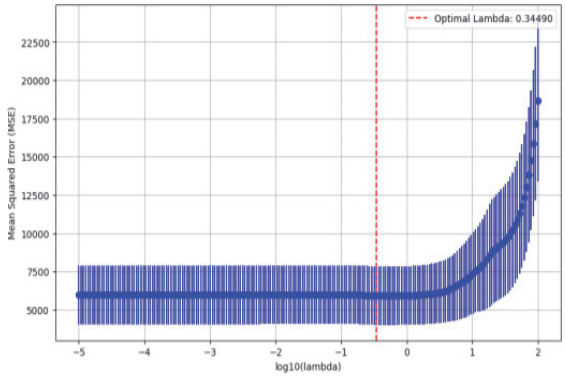


Figure 3: *Top 10 Lasso Selected Coefficients with Error Bars*

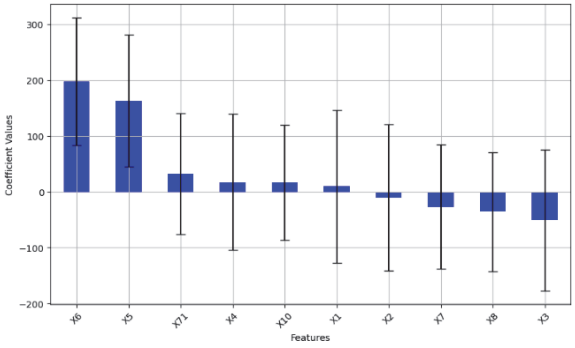


Table 2: Variables Selections

Number of IV	Method	Variables
11	All of the IV	X1, X2, X3, X4, X5, X6, X7, X71, X8, X9, X10
10	LASSO	X1, X2, X3, X4, X5, X6, X7, X71, X8, X10
4	ARDL	X2, X3, X7, X71

4.2. ARDL Multivariate Forecasting

Table 3: Result of ARDL forecasting.

No. of Independent Variables	Test RMSE	Test MAE	Test MAPE	Test SMAPE
11	0.000785	0.000624	2.26	1.05
10	0.000782	0.000623	2.3	1.08
4	0.000833	0.000682	2.5	1.18

Table 3 presents the forecast results for each scenario, indicating the number of independent variables used in the ARDL model. The forecasting accuracy of the ARDL model was assessed under three distinct settings. In the first scenario, utilising all 11 independent variables, the ARDL model achieved a Test RMSE of 0.000785, indicating the root mean squared error between the predicted and actual CPO prices. The Test MAE was 0.000624, indicating the mean absolute error, while the Test MAPE and Test SMAPE were 2.26% and 1.05%, respectively.

In the second scenario, employing ten (10) essential variables selected through the LASSO test, the ARDL model demonstrated improved forecast performance. The Test RMSE decreased to 0.000782, and the Test MAE remained nearly unchanged at 0.000623. The Test MAPE slightly increased to 2.3%, and the Test SMAPE increased to 1.08%.

In the third scenario, utilising four (4) variables based on their significance level from the ARDL analysis, the ARDL model showed comparable forecast accuracy. The Test RMSE was slightly higher at 0.000833, and the Test MAE increased to 0.000682. The Test MAPE was 2.5%, and the Test SMAPE was 1.18%.

Overall, the results highlight the importance of variable selection in multivariate forecasting models. By employing the ten (10) essential variables identified through the LASSO test, the ARDL model achieved the most accurate forecasts. This finding suggests that focusing on relevant macroeconomic factors enhances forecasting accuracy and reduces potential overfitting concerns. Consequently, researchers and practitioners are encouraged to prioritise variable selection techniques, such as the LASSO test, to improve the performance of multivariate forecasting models for CPO prices.

The Multivariate ARDL forecast results demonstrate the impact of variable selection on the model's forecasting performance. Utilising the ten (10) essential variables identified through the LASSO test led to improved forecast accuracy, showcasing the importance of incorporating relevant macroeconomic factors. By emphasising parsimony and focusing on essential variables, the ARDL model offers valuable insights into the dynamic relationships among macroeconomic indicators and their influence on CPO prices.

4.3. LSTM Multivariate Forecasting Result.

Table 4: Forecast performance of the multivariate LSTM model with different size factor sets

No. of Independent Variables	Prediction Horizon	Test RMSE	Test MAE	Test MAPE	Test SMAPE
11	24-steps	0.0322	0.0256	4.9738	4.7850
11	12-steps	0.0331	0.0257	5.0338	4.8320
11	06-steps	0.0361	0.0281	5.6522	5.4071
10	24-steps	0.0282	0.0221	4.2725	4.1468
10	12-steps	0.0299	0.0233	4.536	4.3956
10	06-steps	0.0335	0.0262	5.2212	5.0557
4	24-steps	0.0362	0.0277	5.9268	5.6392
4	12-steps	0.0390	0.0310	6.6612	6.3237
4	06-steps	0.0484	0.0406	9.1453	8.5684
0	24-steps	0.1147	0.0987	1.4238	1.4367
0	12-steps	0.1115	0.0968	1.4083	1.4204
0	06-steps	0.1321	0.1133	1.6664	1.6856

The insights from Table 4 stem from an exhaustive analysis of the LASSO-filtered dataset, pinpointing the most influential factor set. The analysis identifies a set of 10 variables as optimal, surpassing configurations with either 11 or 4 variables. This finding extends beyond the performance of a conventional LSTM model that relies solely on time series data without integrating factors.

A thorough assessment of performance metrics, including RMSE, MAE, MAPE, and SMAPE, across various combinations of independent variables and prediction horizons is crucial. Lower values in these metrics indicate superior model performance, reflecting more minor discrepancies between the model's forecasts and actual observed values.

For the 24-step prediction horizon, the model with 11 independent variables shows an RMSE of 0.0322, MAE of 0.0256, MAPE of 4.9738%, and SMAPE of 4.785%. In contrast, the model with 10 independent variables performs better, with an RMSE of 0.0282, MAE of 0.0221, MAPE of 4.2725%, and SMAPE of 4.1468%. The models with 4 and 0 independent variables yield higher RMSE values of 0.0362 and 0.1147, respectively.

At the 12-steps prediction horizon, the model with 10 independent variables continues to outperform, with RMSE, MAE, MAPE, and SMAPE values of 0.0299, 0.0233, 4.536%, and 4.3956%, respectively. The 11-variable model also shows competitive results.

For the 6-steps prediction horizon, the 10-variable model maintains its lead, with RMSE, MAE, MAPE, and SMAPE values of 0.0335, 0.0262, 5.2212%, and 5.0557%, respectively. The 11-variable model follows closely. Overall, the model with 10 independent variables consistently demonstrates higher accuracy than other configurations. However, the 11-variable model is noteworthy for its SMAPE performance, indicating a balance in percentage errors.

In conclusion, considering the comprehensive assessment metrics, the model with 10 independent variables emerges as the optimal choice for forecasting CPO prices in this context. This decision should consider model complexity, interpretability, and application-specific requirements. Additionally, further insights can be gained through additional statistical tests or cross-validation to refine the model selection process.

The SHAP approach uses the multivariate single-step LSTM model as an example for interpretation. A single sample from January 2020 was selected for the local explanation, and the results are shown in Figure 4.

Figure 4: *Impact of single sample characteristics (January 2020 forecast)*

The 12 test set samples, from January 2020 to December 2021, were selected for the CPO price explanation, and the feature density scatter plot was drawn as shown in Figure 5. Each row corresponds to a feature, while the horizontal axis is the SHAP value. High and low feature values are indicated by red and blue, respectively.

Figure 5: *Scatter plot of feature density*

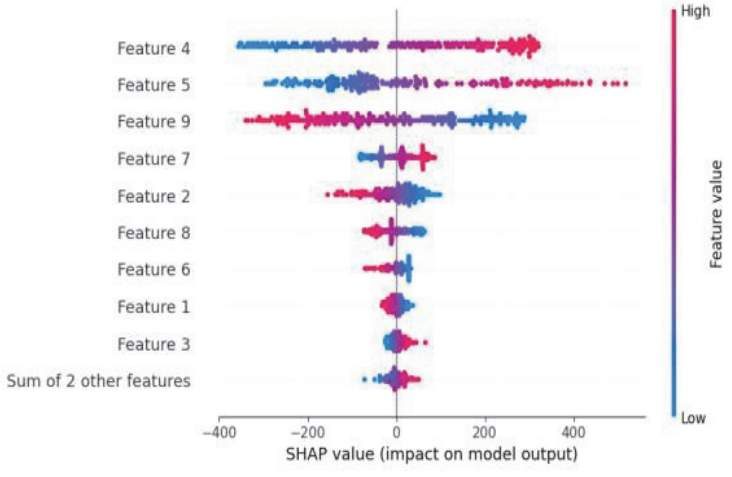
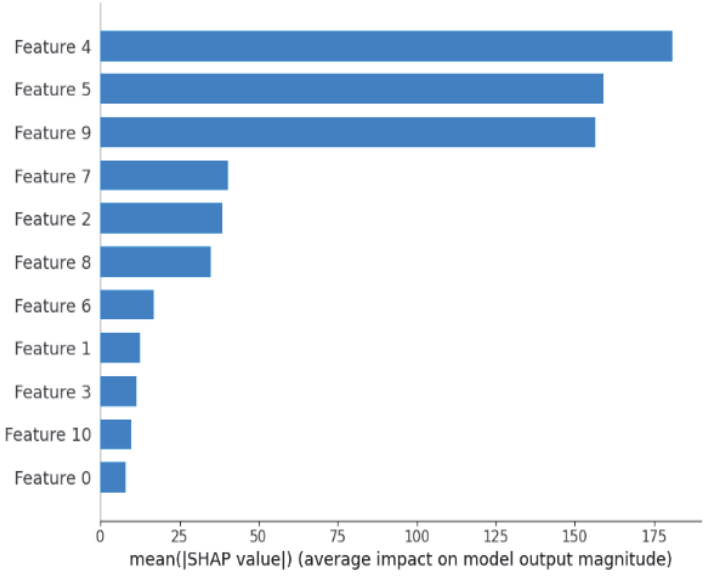


Figure 6: *Feature importance SHAP values*



Specifically, futures 4, the tax rate, shows that its lower value will drive up the predicted value of CPO Prices. The feature 5, weather, indicates that its higher value pushes the predicted value. The scatter of the remaining features oscillates about the SHAP value of 0. There is no spread to either side, which shows that these features have a lower correlation with the projected values. The absolute value of the SHAP value was initially calculated and then averaged to determine the feature significance, as shown in Figure 6.

4.4.Comparative Analysis.

The comparative analysis of the Multivariate ARDL and LSTM models reveals that both approaches exhibit commendable forecasting accuracy. The LSTM models' strengths lie in capturing complex temporal dependencies and non-linear patterns, while the Multivariate ARDL models excel in assessing long-term relationships among the variables.

Figure 7: *Multivariate ARDL model*

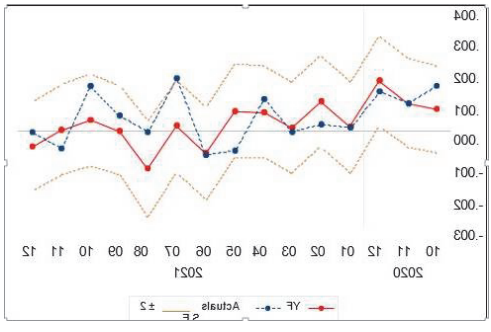
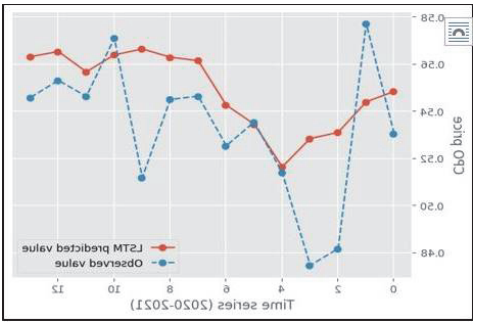


Figure 8: *Multivariate LSTM model*



The ARDL model significantly surpasses the LSTM model in forecasting accuracy, as evidenced by its lower RMSE, MAE, MAPE, and SMAPE values across various data scenarios in this study. This superiority indicates that the ARDL model is more adept at providing precise forecasts for the specific context examined. In terms of consistency, the ARDL model demonstrates a stable performance across different configurations of independent variables, unlike the LSTM model, which shows more significant variability in its performance metrics across various prediction horizons and sets of independent variables. This contrast underscores the value of selecting a forecasting model that aligns with the data's characteristics and the forecasting task's goals. The ARDL model's superior performance in this instance may also highlight the significance of model simplicity and interpretability, especially when the data's underlying relationships can be effectively captured without resorting to the more complex and computationally intensive LSTM models. This finding suggests that, for the data and objectives of this study, the ARDL model's straightforward econometric approach is more beneficial, pointing to the crucial role of model selection in forecasting accuracy.

4.5. Diagnostic Test.

Table 4: Forecast performance of the multivariate LSTM model with different size factor sets

Diagnostic test	X^2 (P-value)	Result
Breusch-Godfrey LM	0.95	No evidence of serial correlations
Breusch-Pagan-Godfrey	0.44	No evidence of heteroscedasticity
Ramsey RESET test	0.33	The model is specified correctly.

The diagnostic tests suggest that the model is stable and well-specified, with no signs of serial correlations or heteroscedasticity, as confirmed by the stability of the error correction coefficients shown in Figures 9 and 10.

Figure 9: *CUSUM test*

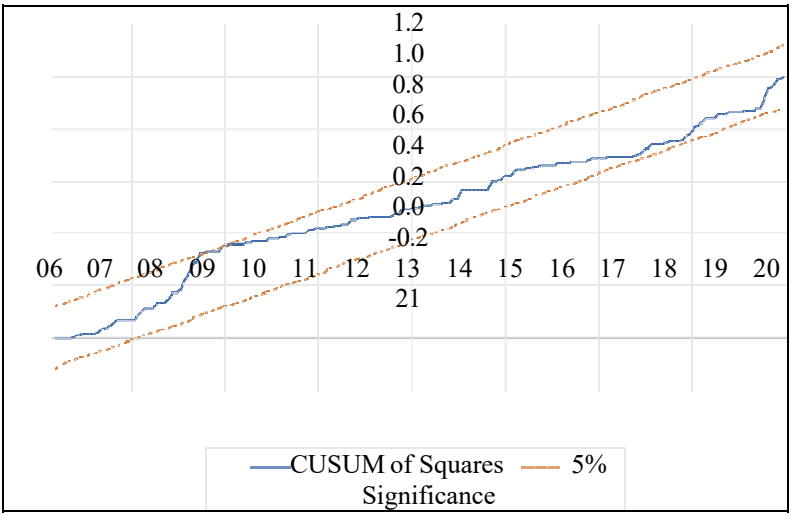
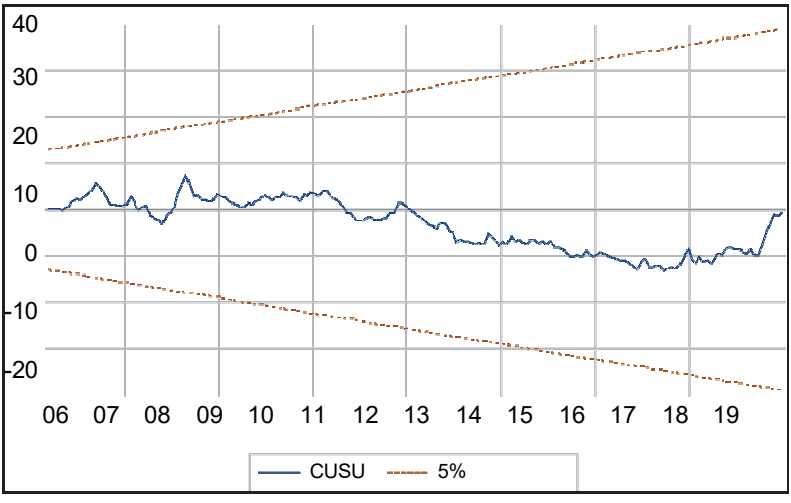


Figure 10: *CUSUM of Squares test*



5. DISCUSSION

The comprehensive analysis and comparison of the ARDL and LSTM models for CPO price forecasting have yielded valuable insights into their respective forecasting capabilities. This discussion section aims to delve deeper into the results' implications and their significance for decision-makers, researchers, and practitioners in the field of CPO price prediction.

One of the key takeaways from this study is the critical importance of variable selection in improving the accuracy of forecasting models. In the case of the ARDL model, the LASSO test played a pivotal role in identifying a subset of essential variables. The results demonstrated that focusing on these relevant factors led to more precise and accurate forecasts. Moreover, the identification of the variables is selected based on the regression result of the ARDL method; 10 variables are selected based on LASSO, and the 4 variables are selected based on the significance level of the variables.

This finding aligns with the idea that incorporating irrelevant variables can introduce noise and potentially lead to overfitting, thus diminishing the model's forecasting prowess (Castle, 2021). Decision-makers and researchers are urged to adopt similar variable selection techniques to enhance the performance of multivariate forecasting models in CPO prices.

On the LSTM side, the analysis revealed that a model incorporating 10 independent variables outperformed other configurations with either more or fewer variables. This outcome underscores the significance of feature engineering and model parsimony. Models that balance incorporating relevant factors and avoiding unnecessary complexity yield superior results. It is worth noting that the LSTM model's ability to capture complex temporal relationships and non-linear patterns was leveraged effectively by selecting these influential variables.

Another intriguing aspect explored in this study was the impact of temporal granularity on forecasting accuracy, particularly in the context of the LSTM model. The analysis considered different prediction horizons (24-steps, 12-steps, and 6-steps), providing

insights into how the model's performance varied with varying time intervals. The results indicated that, for all prediction horizons, the model with 10 independent variables consistently outperformed other scenarios. This suggests that the selected variables, in conjunction with the LSTM architecture, are robust across different timeframes.

Forecasting accurately at different temporal scopes is essential for decision-makers in the CPO market, where short-term and long-term planning is crucial. By understanding the optimal temporal scope for robust forecasting, stakeholders can make more informed decisions regarding inventory management, production planning, and risk mitigation.

The trade-off between the model's complexity and performance is a noteworthy consideration in model selection. While the model with 10 independent variables consistently outperformed others in terms of accuracy, it is essential to acknowledge that increasing model complexity may raise interpretability challenges and computational demands. Therefore, selecting the optimal model should also consider contextual requirements, available computational resources, and the need for transparency in decision-making processes.

Using the SHAP method to interpret the LSTM model's predictions is another valuable aspect of this study. SHAP allowed for identifying individual contributions to various factors, including meteorological variables. This interpretative insight enhances our understanding of the underlying dynamics affecting CPO prices, providing a more holistic view of the forecasting process. Decision-makers and scholars can utilise interpretability techniques to understand better the factors influencing CPO pricing and enhance decision-making.

6. CONCLUSION

The study focuses on providing reliable and accurate models and methods for Multivariate CPO price forecasting. This study highlights the importance of precise variable selection in CPO price forecasting, showcasing the effectiveness of both the ARDL and LSTM models and the ARDL model's superior accuracy and consistency across various scenarios. This study highlights the importance of choosing a simple and interpretable forecasting model that aligns with the data's characteristics and the study's objectives. Moreover, considering different temporal scopes, the study emphasises the need for a balanced trade-off between model complexity and performance. Additionally, using interpretability methods like SHAP enhances our understanding of CPO price dynamics, and selecting variables based on ARDL results is important in identifying the factors that influence CPO prices. This research provides significant information for decision-makers and researchers in the CPO business to make well-informed choices in a changing market despite its limits. Lastly, further research and data refinement are essential for continuous improvement in CPO price forecasting.

ACKNOWLEDGEMENT

The completion of this research was facilitated by the financial assistance provided by the Grant Incentive for Supervision (GIP) from Universiti Teknologi MARA (UiTM). This support, orchestrated through the Research Management Centre (RMC) and the Office of the Deputy Vice-Chancellor (Research & Innovation), was crucial to our work success. We are profoundly grateful to UiTM for offering the support and the necessary resources, which played a pivotal role in achieving our research objectives.

REFERENCES

1. Althaea, K. A., El-Alfy, E. S. M., & Mohammed, S. (2018). Stock market forecast using multivariate analysis with bidirectional and stacked (LSTM, GRU). In 2018 21st Saudi Computer Society National Computer Conference (NCC) (pp. 1-7). IEEE. https://doi.org/10.1109/NCG.2018.8593076
2. Banerjee, A., Dolado, J., Galbraith, J. W., & Hendry, D. (1993). Cointegration, error correction, and the econometric analysis of non-stationary data. Oxford University Press.
3. https://books.google.com.my/books?hl=en&lr=&id=LfVQEAAAQBAJ&oi=fnd&pg=PR7&dq=Banerjee,+A.,+Dolado,+J.,+Galbraith,+J.+W.,+%26+Hendry,+D.+(1993).+Cointegration,+error+correction,+and+the+econometric+analysis+of+non-stationary+data.+Oxford+University+Press&ots=-8SvWMGiUY&sig=twi7s6vceseLKcUE1oByDQ7ipPE
4. Calvin, K., Beach, R., Gurgel, A., Labriet, M., & Loboguerrero Rodriguez, A. M. (2014). Agriculture, forestry, and other land-use emissions in Latin America. Energy Economics, 42(Supplement C), 429-446.
5. https://doi.org/10.1016/j.eneco.2015.03.020
6. Chandrarin, G., Sohag, K., Cahyaningsih, D. S., Yuniawan, D., & Herdhayinta, H. (2022). The response of exchange rate to coal price, palm oil price, and inflation in Indonesia: Tail dependence analysis. Resources Policy, 77, 102750. https://doi.org/10.1016/j.resourpol.2022.102750
7. Castle, J., Doornik, J., & Hendry, D. (2021). Selecting a Model for Forecasting.
8. Econometrics. https://doi.org/10.3390/econometrics9030026.
9. Cespedes, L. F., & Velasco, A. (2012). Assessing the impact of food trade policies on poverty alleviation and welfare. Journal of Policy Modelling, 34(6), 879-894. https://doi.org/10.1057/imfer.2012.22
10. Corley, R. H. V. (2009). How much palm oil do we need? Environmental Science & Policy, 12(2), 134-139. https://doi.org/10.1016/J.ENVSCI.2008.10.011.
11. Enghiad, A., Ufer, D., Countryman, A., & Thilmany, D. (2017). An Overview of Global Wheat Market Fundamentals in an Era of Climate Concerns. International Journal of Agronomy, 2017, 1-15. https://doi.org/10.1155/2017/3931897.
12. Hamid, M. F. A., & Shabri, A. (2017). Palm oil price forecasting model: An autoregressive distributed lag (ARDL) approach. AIP Conference Proceedings, 1842(030026). https://doi.org/10.1063/1.4982864.
13. Hassan, A., & Balu, N. (2016). Examining the long-term relationships between the prices of palm oil and soybean oil, palm oil production and export: Cointegration and causality. Oil Palm Industry Economic Journal, 16(1), 31-37. https://www.researchgate.net/publication/324983472
14. Headey, D., & Fan, S. (2008). Anatomy of a crisis: The causes and consequences of surging food prices. Agricultural Economics, 39(1), 375-391. https://doi.org/10.1111/J.1574-0862.2008.00345.X.
15. Hisham, A. A. M., Karim, Z. A., & Khalid, N. (2019). Determinants of capital expenditure spending in Malaysian palm oil industries: A dynamic panel data analysis. Economic Journal of Emerging Markets, 223-233. https://doi.org/10.20885/ejem.vol11.iss2.art9.
16. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780. https://doi:10.1162/neco.1997.9.8.1735

17. Isa, M., Samsudin, S., & Noh, M. (2016). Cointegration and Causality between the Spot and Futures Markets: Pre and Post Implementation of NKEAs, 2, 98. https://doi.org/10.24191/abrij.v2i1.10068.50-57.
18. Karim, F., Majumdar, S., Darabi, H., & Harford, S. (2019). Multivariate LSTM-FCNs for time series classification. Neural networks, 116, 237-245. https://arxiv.org/pdf/1801.04503
19. Khosla, P. (2011). Nutritional characteristics of palm oil. In Reducing Saturated Fats in Foods(pp.112-127). Woodhead Publishing. https://doi.org/10.1533/9780857092472.1.112
20. Lin, Y.-C., & Huang, M.-H. (2011). The determinants of palm oil price: An empirical analysis. Journal of Agricultural Economics, 62(1), 141-156. https://www.mdpi.com/2071-1050/13/23/13480/pdf
21. Lu, Q., Sun, S., Duan, H., & Wang, S. (2021). Analysis and forecasting of crude oil price based on the variable selection-LSTM integrated model. Energy Informatics, 4(1), 1-20. https://doi.org/10.1186/s42162-021-00166-4
22. Ma, Q., Huang, J., Başar, T., Liu, J., & Chen, X. (2021). Reputation and Pricing Dynamics in Online Markets. IEEE/ACM Transactions on Networking, 29, 1745-1759. https://doi.org/10.1109/TNET.2021.3071506.
23. May, C. Y., & Nesaretnam, K. (2014). Research advancements in palm oil nutrition. European Journal of Lipid Science and Technology, 116(10), 1301-1315. https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejlt.201400076
24. Mohammadi, F., Bilash, A., Abdulla, A. (2020). System dynamics analysis of the determinants of the Malaysian palm oil price. Journal of Oil Palm Research, 32(4), 589-601. http://thescipub.com/PDF/ajassp.2015.355.362.pdf
25. Murphy, D.J., Goggin, K. & Paterson, R.R.M. (2021) Oil palm in the 2020s and
beyond: challenges and solutions. CABI Agric Biosci 2, 39.
26. https://doi.org/10.1186/s43170-021-00058-3
27. Nazlioglu, S., & Soytas, U. (2012). Oil price, agricultural commodity prices, and the dollar: A panel cointegration and causality analysis. Energy Economics, 34(4), 1098-1104.
28. https://www.sciencedirect.com/science/article/pii/S014098831100209X
29. Ofuoku, M., & Ngniatedema, T. (2022). Predicting the Price of Crude Palm Oil: A Deep Learning Approach. International Journal of Strategic Decision Sciences (IJSDS), 13(1), 1-15. https://www.igi-global.com/article/predicting-the-price- of-crude-palm-oil/305830
30. Zainalabidin, S. M., & Rahim, K. A. (2012). Impact of climate change on palm oil production. In Proceedings of USM-AUT International Conference 2012 Sustainable Economic Development: Policies and Strategies (Vol. 167, p. 109). https://www.researchgate.net/profile/Abdelhak_Senadjki/publication/262686349_Poverty_in_Algeria_An_Institutional_Crisis_or_Developmental_Problem/links/0deec53874932d53ff000000/Poverty-in-Algeria-An-Institutional-Crisis- or-Developmental-Problem.pdf#page=119
31. Phitthayaphinant, P., Nissapa, A., Somboonsuke, B., & Eksomtramage, T. (2012). An equation of oil palm plantation areas in Thailand. KKU Research Journal, 11(1), 66-76. https://rtt.kku.ac.th/ejournal/pa_upload_pdf/414679.pdf
32. Putri, P. Y., Achsani, N. A., & Pranowo, K. (2019). The Effects of Macroeconomic Variables and Corporate Financial Performance on Stock Prices of Palm Oil Companies in Indonesia. Jurnal Manajemen & Agribisnis, 16(1), 12-12. /DOI: http://dx.doi.org/10.17358/jma.16.1.12

33. Pashigian, B. P. (2008). „Cobweb Theorem“. The New Palgrave Dictionary of Economics (2nd ed.) https://books.google.com/books?hl=en&lr=&id=EO40DAAAQBAJ&oi=fnd&pg=PR2&dq=+%22Cobweb+Theorem%22.+The+New+Palgrave+Dictionary+of+Economics+(2nd+ed.)&ots=8m56JYBDlX&sig=7-xm8vvetDXl5pqFeoolI2_XzbI
34. Pesaran, M. H., Shin, Y., & Smith, R. J. (2001). Bounds testing approaches to the analysis of level relationships. Journal of Applied Econometrics, 16(3), 289-
35. 326. https://doi.org/10.1002/jae.616
36. Pesaran, M. H., & Shin, Y. (1999). An autoregressive distributed lag modelling approach to cointegration analysis. Econometric Society Monographs, 31, 371-
37. 413.http://request-
38. attachments.storage.googleapis.com/bRv1Dv9b8djCBcOc9hAnvz9gHg1eA4HF1gOGySUCokMEpfXnVvGzMvfj3Hu6YWtrdDaYEeP7BAVQP0FTkZs8JQKRIh6HNaElqtPV/An_Autoregressive_Distributed_Lag_Modeling_Approac. pdf
39. Peterson, P. (2014). Fixing Prices and Fixing Markets. farmdoc daily. https://ageconsearch.umn.edu/record/283094/files/fdd250614.pdf
40. Ricci, E., Peri, M., & Baldi, L. (2019). The Effects of Agricultural Price Instability on Vertical Price Transmission: A Study of the Wheat Chain in Italy. Agriculture. https://doi.org/10.3390/AGRICULTURE9020036.
41. Sagheer, A., & Kotb, M. (2019). Unsupervised pre-training of a deep LSTM-based stacked autoencoder for multivariate time series forecasting problems. Scientific reports, 9(1), 19038. https://www.nature.com/articles/s41598-019-
42. 55320-6
43. Ludin, N., Bakri, M., Kamaruddin, N., Sopian, K., Deraman, M., Hamid, N., Asim, N., & Othman, M. (2014). Malaysian oil palm plantation sector: Exploiting renewable energy toward sustainability production. Journal of Cleaner Production, 65, 9-15. https://doi.org/10.1016/J.JCLEPRO.2013.11.063.
44. Oosterveer, P. (2015). Promoting sustainable palm oil: viewed from a global networks and flows perspective. Journal of Cleaner Production, 107, 146-153. https://doi.org/10.1016/J.JCLEPRO.2014.01.019.
45. Tibshirani R. (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser
46. B. 1996;58(1):267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
47. Tomek, W.G. & Robinson, K.L. 1981. Agricultural Product Prices. Ithaca, New York.: Cornell University Press. https://www.cabidigitallibrary.org/doi/full/10.5555/19811875109
48. Urolagin, S., Sharma, N., & Datta, T. K. (2021). A combined architecture of multivariate LSTM with Mahalanobis and Z-Score transformations for oil price forecasting.Energy,231,120963.
49. https://www.sciencedirect.com/science/article/pii/S0360544221012111
50. Widiputra, H., Mailangkay, A., & Gautama, E. (2021). Multivariate cnn-lstm model for multiple parallel financial time-series prediction. Complexity, 2021, 1-14. https://www.hindawi.com/journals/complexity/2021/9903518/
51. Wilson, N., & Cacho, J. (2007). Linkage between foreign direct investment, trade and trade policy: an economic analysis with application to the food sector in OECD countries and case studies in Ghana, Mozambique, Tunisia and Uganda. https://www.oecd-ilibrary.org/content/paper/152275474424

52. Salami, M. A., & Haron, R. (2018). Long-term relationship of crude palm oil commodity pricing under structural break. Journal of Capital Markets Studies, 2(2), 162-174. https://www.emerald.com/insight/content/doi/10.1108/JCMS- 09-2018-0032/full/pdf
53. Zaidon, N. A., & Karim, A. (2019, September). Z. The Effects of External and Internal Shocks on the Movement of Palm Oil Price: A SVAR Analysis of Malaysia. In Proceedings of the International Conference on Economics, Kota Kinabalu, Sabah, Malaysia (pp. 18-19). https://www.ums.edu.my/fpep/files/paper12019.pdf
54. Zhou, S., Zhou, L., Mao, M., Tai, H., & Wan, Y. (2019). An Optimised Heterogeneous Structure LSTM Network for Electricity Price Forecasting. IEEE Access, 7, 108161-108173. https://doi.org/10.1109/ACCESS.2019.2932999.
55. Zaidi, M. A. S., Karim, Z. A., & Zaidon, N. A. (2022). External and internal shocks and the movement of palm oil price: svar evidence from malaysia. Economies. https://doi.org/10.3390/economies10010007