Techno-Science Research Journal V14 (Special Issue) (2026) P 10-18

Content list available at ITC

Techno-Science Research Journal

Journal Homepage: http://techno-srj.itc.edu.kh/

Techno-Science Research Journal

Comparing Time Series, Machine Learning, and Deep Learning Models for Paddy Rice Price Forecasting in Battambang Province, Cambodia

Phirum Leng*, Sokkhey Phauk, Malai Nhim, Chantara Tann

Department of Applied Mathematics and Statistics, Institute of Technology of Cambodia, Russian Federation Blvd. P.O. Box 86, Phnom Penh, Cambodia

Received: 16 June 2025; Revised: 18 July 2025 Accepted: 10 September 2025; Available online: 31 July 2026

Abstract: *In Cambodia, paddy rice is a staple food and a key income source, with Battambang Province being one of the leading rice-producing regions. Accurate paddy rice price forecasts are crucial for farmers, traders, policymakers, and financial institutions to make informed decisions, mitigate risks, and stabilize the market. Several factors influence price fluctuations, including domestic production, global market trends, input costs, and macroeconomic variables like crude oil prices and exchange rates. Developing forecasting models that capture these dynamics is crucial for agricultural sustainability and economic resilience. This study compares traditional time series models (ARIMAX) with machine learning models (XGBoost) and deep learning models (LSTM) to evaluate forecasting accuracy. The results show that XGBoost outperforms both ARIMAX and LSTM, achieving the lowest RMSE (29.90 KHR/Kg), MAE (25.47 KHR/Kg), and MAPE (1.70%). In contrast, ARIMAX performs moderately, while LSTM exhibits higher error rates. This study provides stakeholders with a data-driven approach to agricultural decision-making, supports sustainable farming practices, and enhances market predictability, thereby contributing to economic stability in Cambodia's rice industry. The findings also inform policy recommendations to improve trade efficiency and food security.*

Keywords: Rice Price Forecasting, Time Series Models, Machine Learning, Deep Learning, ARIMAX, XGBoost, LSTM, Market Stability, Forecasting Accuracy

1. INTRODUCTION

As a staple crop and a significant export, rice is essential to Cambodia's economy, and Battambang is a major center for production [1]. However, changes in rice prices pose difficulties for farmers, traders, and legislators, affecting economic stability and food security [3]. While machine learning and deep learning models, such as XGBoost and LSTM, have shown promise in increasing accuracy, conventional forecasting techniques like ARIMA struggle to capture intricate pricing patterns. However, few studies have combined the two methods to improve the forecasts [2]. This study evaluates XGBoost, LSTM, and time series models (ARIMAX) for predicting rice prices in Battambang. It aims to determine the most suitable strategy by examining key variables, including crude oil prices and the USD exchange rate. The results will help with economic planning, market stability, and decision-making in the rice industry in Cambodia. Since the agriculture industry faces both market volatility and external shocks, accurate price forecasting becomes a crucial tool for ensuring resilience and informed policymaking. By contrasting machine learning and deep learning techniques with traditional models, this study seeks to improve forecast accuracy while illuminating the seasonal and nonlinear patterns of rice prices [4]. Finally, the study applied a comparative analysis of the models to support supply chain and rice stock management, focusing on their advantages and disadvantages.

Additionally, this study supports the Global Sustainable Development Agenda, particularly the Sustainable Development Goals (SDGs) related to hunger, poverty, and economic growth. By providing accurate price forecasts, the study aims to contribute to SDG 1 (No Poverty), SDG 2 (Zero Hunger), and SDG 8 (Decent Work and Economic Growth [5], helping policymakers, stakeholders, and farmers make

* Corresponding author: *Leng Phirum*
E-mail: lphirum4@gmail.com

informed decisions that promote long-term agricultural sustainability and economic resilience.

Our research applied time series (TS), machine learning (ML), and deep learning (DL) to analyze the performance of models and predict paddy rice prices. This will help examine the basic significance of accurately predicting the paddy rice price, highlighting the main obstacles such as supply chain, market volatility, by describing policy, social, and economic impacts for different stakeholders in the nation that rely on rice or paddy rice.

While recent studies have applied statistical, machine learning (ML), and deep learning (DL) models to forecast rice prices, showing that ML and DL, especially LSTM-based models, often outperform traditional methods like ARIMA, some, like Ray et al. (2023), also highlight the added value of using external variables. This review underscores the growing effectiveness of advanced models and supports their potential for improving rice price forecasting, particularly in regions like Battambang Province, Cambodia.

2. LITERATURE REVIEW

A few recent studies have explored the various approaches to forecasting agricultural commodity prices, particularly rice, using both traditional and modern predictive models. The Autoregressive Integrated Moving Average (ARIMA) and its extension (ARIMAX) models have been widely adopted due to their simplicity and interpretability, but they often fall short when dealing with nonlinear patterns and external shocks. To address these limitations, researchers have increasingly turned to Machine Learning (ML) and Deep Learning (DL) techniques, including XGBoost, Random Forest, and Long Short-Term Memory (LSTM) networks, which are capable of capturing complex temporal and nonlinear relationships.

For instance, Chanchala et al. (2019) demonstrated that time series and regression models effectively capture seasonality and trends in rice prices, particularly when analyzing macroeconomic influences such as crude oil and exchange rates. This directly supports our incorporation of macroeconomic variables as exogenous inputs in time series models like ARIMAX, helping to contextualize price movements in Cambodia's rice market.

Another comparative study by Chen et al. (2021) has confirmed that comparing traditional models with machine learning and deep learning aligns closely with our methodology. Their finding that LSTM outperformed other models reinforces the value of including deep learning techniques in our comparative analysis. Moreover, their use of XGBoost and Random Forest further validates our choice of machine learning methods.

Despite the geographic context and forecasting target, Noorunnahar et al. (2023) compare performance between ARIMA and XGBoost for rice production forecasting in Bangladesh, showing that machine learning models can outperform traditional models in agricultural contexts. This comparison directly informed our hypothesis that advanced machine learning models may improve forecasting accuracy for rice prices in Battambang.

Therefore, these studies contribute to the existing body of knowledge by comparing traditional time series models (ARIMAX) with ML and DL models (XGBoost and LSTM), incorporating key exogenous variables such as crude oil prices and exchange rates. This approach not only improves forecasting accuracy but also enhances understanding of the underlying drivers of rice price fluctuations in Battambang Province, Cambodia's leading rice-producing region.

3. METHODOLOGY

3.1 Dataset

The dataset used in this study comprised historical paddy rice prices from January 2007 to June 2022, accumulating of 183 monthly records obtained from the World Food Program's website. Additional factors, such as crude oil prices and USD exchange rates from World Bank Commodity Prices and the National Bank of Cambodia (NBC), were incorporated to capture macroeconomic influences on paddy rice prices.

This study uses a monthly time series dataset to forecast paddy rice prices in Battambang Province, Cambodia. The dataset includes wholesale rice prices (Riels/Kg) as the target variable, along with crude oil prices (USD/barrel) and the USD-Khmer Riel exchange rate as external economic factors.

To prepare the dataset for analysis, missing values were handled, continuous variables were scaled, and time indices were aligned. Exploratory data analysis (EDA) was conducted to identify trends, correlations, and stationarity within the data. The inclusion of economic indicators aims to improve forecasting accuracy and offer deeper insight into the factors influencing rice price volatility over time. An overview of key features and their descriptions is provided in Table 1 below.

Table 1. Sample of the Dataset Used

date	Crude oil avg_$/bbl	KHR=X	price_kh
1/15/2007	53.52	3878.56	956.67
2/15/2007	57.56	3863.49	970.00

date	Crude oil avg_$/bbl	KHR=X	price_kh
3/15/2007	60.60	3845.16	967.50
…	…	…	…
2/15/2022	93.54	4029.42	1500.00
3/15/2022	112.40	4017.29	1560.00

3.2 Data Preparation and Preprocessing

The work begins with data preparation and preprocessing to ensure its suitability for analysis. Data profiling is performed to assess quality and detect anomalies, followed by data cleaning to address errors, inconsistencies, and missing or duplicate values. The Date column is transformed into a datetime format for more informative analysis. Visualizations are used to understand the data distribution and characteristics.

The significance of this processing is that stable values are replaced for periods longer than 6 months (Apr 2017 to Oct 2017) and (Aug 2020 to Mar 2021). Figure 1 below shows the data after the stable values were replaced.

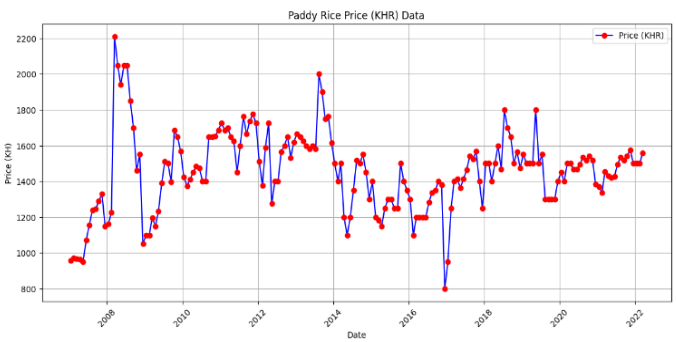

Fig. 1. Time plot of Paddy Rice Price (KHR)

3.3 Correlation Matrix

A statistical metric of correlation measures how strongly and in which direction two variables have a linear relationship. The plot below indicates the relationship of variables in this research.

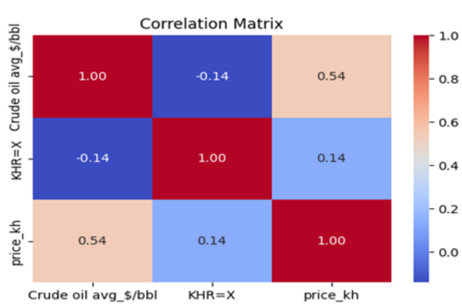

Fig. 2. Correlation Matrix of variables

The correlation matrix above illustrates the relationships between crude oil price, exchange rate, and rice prices in Battambang, Cambodia. Crude oil shows a moderate positive correlation (0.54) with rice prices, suggesting that rising oil costs may increase rice prices. The exchange rate weakly correlates with crude oil (-0.14) and rice prices (0.14), indicating limited influence. The color gradient reflects the strength and direction of each relationship. Overall, crude oil prices appear to have the most significant impact on rice prices.

3.4 Feature Scaling

To ensure a proper contribution during feature selection, feature scaling standardizes the range of independent variables, and to stop features with higher values from dominating, all features are scaled to the 0-1 range using MinMaxScaler().

3.5 Data Splitting

To assess model performance and prevent overfitting, the data is divided into training and testing sets. The model is trained on the first 146 instances (index 0-145) and evaluated on the remaining 38 unseen instances (index 146-183) to measure its accuracy.

3.6 Model Selection

Model selection is crucial in this research. We applied ARIMAX (with exogenous variables), machine learning, and deep learning models to assess their performance. This approach captures both linear and nonlinear relationships, including the impact of external variables. Time series models are considered first because rice prices show temporal patterns, trends, and seasonality. ARIMA is suitable for univariate analysis, focusing on a single variable, leading to ARIMAX, which supports multivariate analysis by incorporating external factors such as crude oil prices and USD exchange rates. This provides a solid foundation for comparison and improves forecasting accuracy.

3.6.1 Autoregressive Integrated Moving Average (ARIMA)

ARIMA is a classical time series forecasting method designed for univariate data. It captures patterns such as trend and seasonality by combining three components: AR (p) is autocorrelation from previous time steps, I (d) is the number of differencing to ensure stationarity, and MA (q) incorporates the influence of past forecast errors. The formula of ARIMA is demonstrated below:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \qquad \text{(Eq. 1)}$$

Where:

X_t : Value of the time series at time t
c : Constant term (intercept)
φ_i : Autoregressive (AR) coefficients, where i ranges from 1 to p
θ_j : moving average (MA) coefficients with j ranging from 1 to q
ε_t : Error terms (white noise) at time t

ARIMA is applied in the initial step to analyze the univariate time series, allowing us to evaluate how much of the rice price variation can be explained and predicted using only historical rice price data. Figure 3 below shows the process for ARIMA:

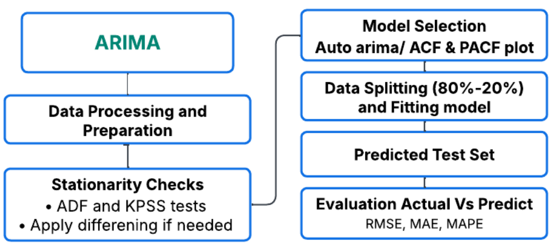

Fig. 3. ARIMA Process Flow

3.6.2 Autoregressive Integrated Moving Average with Exogenous Regressors (ARIMAX)

The ARIMAX model extends the classical ARIMA model by incorporating external (exogenous) variables that may influence the target variable [6]. The ARIMAX model is useful when external factors, such as oil prices and exchange rates, are expected to impact the variable of interest, which is the rice price in this study. The general form of the ARIMAX model is

$$y_t = \alpha + \sum_{i=1}^{p} \emptyset_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \sum_{k=1}^{K} \beta_k X_{k,t} + \varepsilon_t \qquad \text{(Eq. 2)}$$

Where:

y_t : Target variable at time (t)
α : Intercept (constant term)
$\emptyset_i$: Autoregressive coefficients (AR part)
θ_j : Moving average coefficients (MA part)
ε_t : Error terms (white noise)
β_k : Coefficients for exogenous variables
$X_{k,t}$: Exogenous variable(s) at time t
ρ, d, q : Orders of the AR, differencing, and MA, respectively
K : Number of exogenous variables

Figure 4 below shows the whole process for ARIMAX:

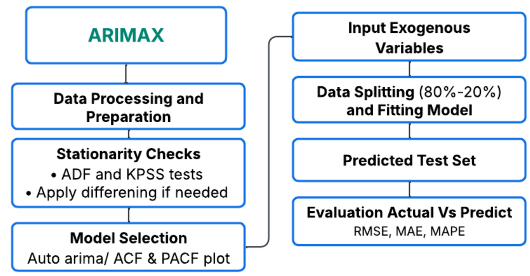

Fig. 4. ARIMAX Process Flow

3.6.3 Extreme Gradient Boosting (XGBoost)

As introduced by Chen and Guestrin (2016), XGBoost is a type of machine learning algorithm that can capture complex feature data and non-linear correlations. To decrease overfitting and increase prediction accuracy, it minimizes a regularized objective function [7]. The objective function in XGBoost is defined as:

$$L(\emptyset) = \sum_{i=1}^{n} l(y_i, \hat{y}_i(t)) + \sum_{k=1}^{t} \sigma(f_k) \qquad \text{(Eq. 3)}$$

Where: $L(\emptyset)$ is the loss function that needs to be minimized.

y_i : The target variable's actual value at time i
$\hat{y}_i(t)$: Target variable's predicted value at time i
$l(y_i, \hat{y}_i(t))$: loss function
n : Number of data points in the time series
T : Number of regularization terms
(f_k) : The regularization term's set of features or model parameters
$\sigma\ (f_k)$: regularization term (e.g., L1/L2 norm)

Below is the flowchart of XGBoost implementation:

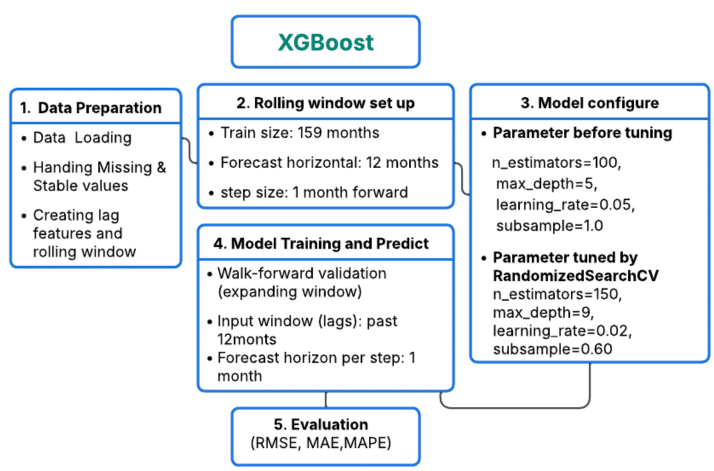

Fig. 4. XGBoost Process Flow

3.6.4 Long-Short Term Memory (LSTM)

LSTM is a specialized type of Recurrent Neural Network (RNN) designed to learn long-term dependencies in sequential data. Unlike traditional RNNs, which suffer from vanishing or exploding gradients when modeling long sequences, LSTM addresses this limitation by introducing gated mechanisms that regulate the flow of information [8], which is expressed by the formula below:

$$f_t = \sigma(W_{fh}h_{t-1} + W_{fx}x_t + b_f) \quad \text{(Eq. 4)}$$

$$i_t = \sigma(W_{ih}h_{t-1} + W_{ix}x_t + b_i) \quad \text{(Eq. 5)}$$

$$\tilde{c}_t = tanh(W_{\tilde{c}_t h}h_{t-1} + W_{\tilde{c}_t x}x_t + b_{\tilde{c}}) \quad \text{(Eq. 6)}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \quad \text{(Eq. 7)}$$

$$o_t = \sigma\left(W_{oh}h_{t-1} + W_{ox}x_t + b_o\right) \quad \text{(Eq. 8)}$$

$$h_t = o_t \cdot tanh(c_t) \quad \text{(Eq. 9)}$$

Where:

f_t : Forget gate- eliminates unnecessary historical data
i_t : Input gate- controls the fresh data is added to memory
$\tilde{c}_t$: Candidate state-recommended values for updating the memory
c_t : Cell state-updated memory that combines fresh and old data
o_t : Output gate- decide which memory is chosen to output
h_t : Hidden state- the LSTM's final output at time t

These equations represent two components of the loss function: (1) the sum of the losses between the actual and predicted values, (2) the sum of regularization terms, which help prevent overfitting or enforce specific constraints.

LSTM workflow in this research is indicated in Figure 5 below:

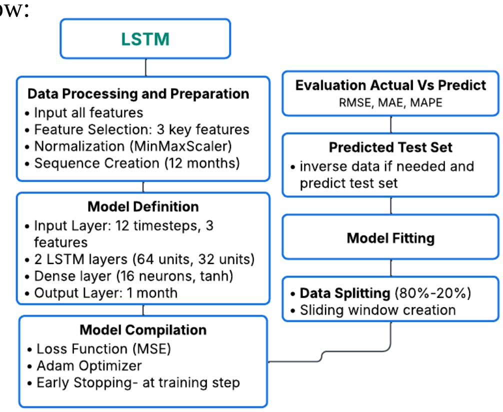

Fig. 5. LSTM Process Flow

3.7 Parameter optimization and Evaluation metrics

To fine-tune the ARIMAX model, the values of p, d, and q were selected based on the Akaike Information Criterion (AIC), which measures the trade-off between model fit and complexity. The data was first made stationary using differencing, and the ideal value of d was then found using the Augmented Dickey-Fuller (ADF) test. The model with the lowest AIC value was then picked to determine the values of p and q. The AIC is calculated as:

$$AIC = 2k - \ln(L) \quad \text{(Eq. 10)}$$

Where:

k : Number of parameters estimated
L : Maximum likelihood of the model

After fitting the ARIMAX models, their predictive performance was evaluated using a validation dataset (index 146–183). Three widely used error metrics were applied:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \text{(Eq. 11)}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \text{(Eq. 12)}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \times 100 \quad \text{(Eq. 13)}$$

Where:

n : Total observations count
A_t, y_i : Actual values of data
$F_t, \hat{y}_i$: Forecasted/Predicted Values

These metrics quantify the average deviation between predicted and actual values, with lower scores indicating better model performance.

4. RESULTS AND DISCUSSION

4.1 ARIMA model

ARIMA starts with converting the date column into a datetime index after data cleaning and processing, and exploratory data analysis. The stationary checks used with ADF (p-value: 0.0096) and KPSS tests with p-value: 0.10, respectively, confirm stationarity in the paddy rice price.

To identify the most suitable order, both ARIMA and SARIMA models are fitted using the auto_arima () function. This function selects the optimal configuration based on AIC and BIC. The ARIMA model is configured with a seasonal component of 12 months, but not significant due to the high p-value in the results ar.S.L12: p = 0.367 and ar.S.L24: p = 0.205, which means the seasonal autoregressive component at these lags is not really strong predictor in this case. Thus, the result from auto_arima is ARIMA (1,0,0).

To better select the best models, we also plot the ACF and PACF in Figures 7 and 8 as below:

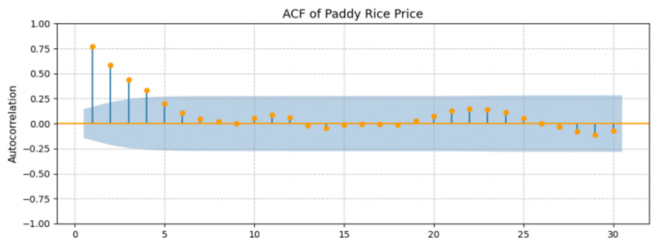

Fig. 7. ACF plot of paddy rice price (KHR) in ARIMA

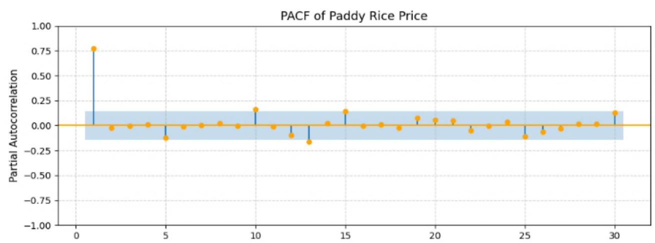

Fig. 8. PACF plot of paddy rice price (KHR) in ARIMA

The ACF and PACF plots in Figures 7 and 8 show strong autocorrelation at lag 1 with a sharp PACF cutoff, supporting an AR(1) process. Minor spikes at seasonal lags (12, 24) are not significant, aligning with the SARIMA output, which supports a non-seasonal ARIMA(1,0,0) model.

After model selection, the dataset is split into training (80%) and testing (20%) sets, and the models are trained solely on the training data. Predictions for the test period and evaluated using the RMSE, MAE, and MAPE. Figure 9 below indicates the actual and predicted test set of paddy rice price.

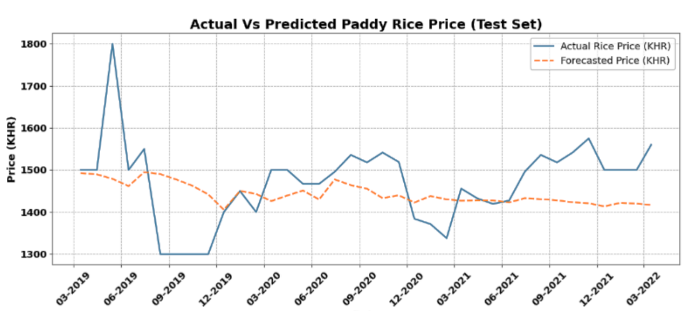

Fig. 9. Actual Vs Predicted Test Set by ARIMA (1,0,0)

The result in Figure 9 shows that while the ARIMA model captures the general trend of paddy rice prices, it underestimates the magnitude of short-term fluctuations. The forecasted values remain relatively stable compared to the more volatile actual prices, indicating limited responsiveness to sudden market changes.

4.2 ARIMAX model

In this study, the ARIMAX model was applied using crude oil price and exchange rate as covariates to capture external effects on rice price as well as to predict its price.

The first step is to test for stationarity using the ADF and KPSS tests. Among the three variables, only Crude Oil Price is non-stationary, with p-values of 0.080 (ADF) and 0.022 (KPSS), indicating the need for transformation or differencing. Then, after applying the 1st differencing, the crude oil appears to be stable. The next step was to identify the p and q using the autocorrelation and partial autocorrelation functions (ACF and PACF) [9] as shown in Figures 10 and 11. The order of p and q with the minimum AIC and BIC selected, and the models were ARIMAX (1,0,0), ARIMAX (1, 1, 1), and ARIMAX (1, 1, 2). Among the above models, ARIMAX (1, 1, 2) is the most suitable model based lowest AIC and BIC.

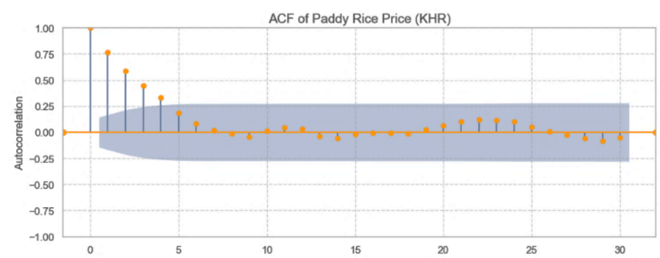

Fig. 10. ACF plot of paddy rice price (KHR) in ARIMAX

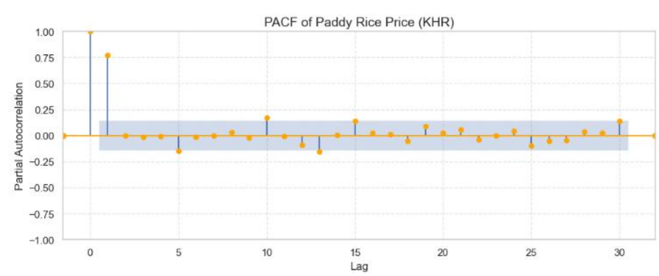

Fig. 11. PACF plot of paddy rice price (KHR) in ARIMAX

Refer to Figures 10 and 11, the PACF cuts off after lag 1, indicating an AR (1) component, and the ACF plot displayed a progressive decrease. The choice of ARIMAX (1,1,2) was influenced by this observation as well as the minimum AIC/BIC.

Figure 12 below shows the actual and predicted test set of paddy rice price in ARIMAX (1,1,2).

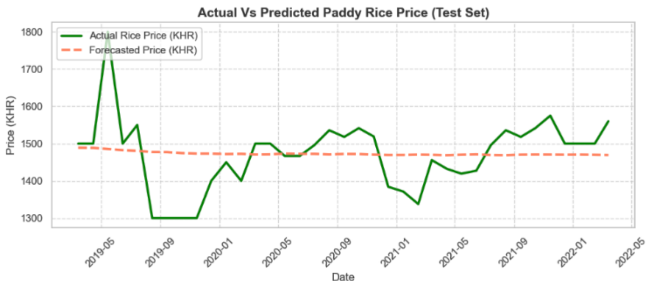

Fig. 12. Actual Vs Predicted Test Set by ARIMAX (1,1,2)

Figure 12 compares the actual paddy rice prices with the predicted prices from the ARIMAX(1,1,2) model over the test period. While the actual prices show significant fluctuations, the predicted values remain relatively stable, indicating that the model struggles to fully capture the short-term volatility in the rice price.

4.3 XGBoost model

The XGBoost model was trained using a rolling forecast (walk-forward) approach with a one-month step size, applied to a dataset containing 183 monthly observations. Lag features from the previous 12 months were used to capture temporal dependencies in the target variable. Two model configurations were evaluated: a baseline model with parameters n_estimators = 100, learning_rate = 0.05, max_depth = 5, and subsample = 1.0, and a tuned model with parameters n_estimators = 150, learning_rate = 0.02, max_depth = 9, and subsample = 0.60. The tuned parameters were obtained using RandomizedSearchCV with TimeSeriesSplit on the initial training window, and then applied across all walk-forward steps. At each forecast step, the model was retrained on all available data up to that point before predicting the next month. The tuned model achieved better accuracy, as illustrated in Figure 13 and discussed in Section 4.5.

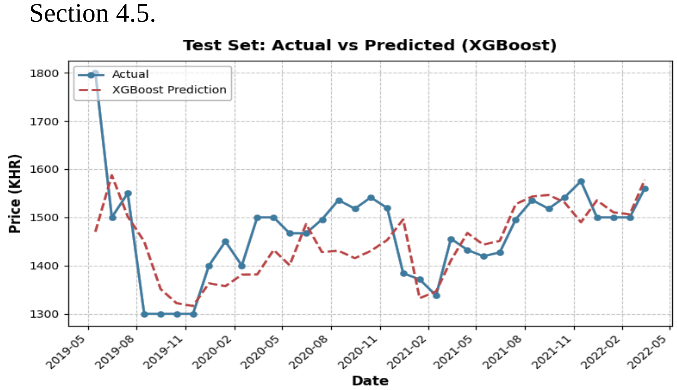

Fig. 13. Actual vs Predicted Test Set by XGBoost

4.4 LSTM model

The LSTM model was implemented to capture long-term dependencies and sequential patterns in rice price movements. All features were normalized using MinMaxScaler. A sequence length of 12 months was selected, allowing the model to learn from one year of historical data to predict rice prices for the next month.

Using a sliding window technique, input sequences were created by shifting the window forward one time step at a time. Each input contained 12 months of features, and the target was the price in the 13th month. This approach was chosen instead of using the entire sequence, as it helps the model focus on recent trends by increasing the number of training samples.

The dataset included 183 monthly records, split into 80% for training and 20% for testing. The model had two LSTM layers with 64 and 32 units, followed by a Dense layer with 16 neurons using the tanh activation, and a final Dense output layer for prediction. Dropout layers were added to reduce overfitting. The model was trained using the Adam optimizer and mean squared error (MSE) loss, with early stopping to avoid overfitting. After training, predictions were converted back to original values and evaluated using RMSE, MAE, and MAPE.

Figure 14 below illustrates the comparison of actual and predicted test set by LSTM.

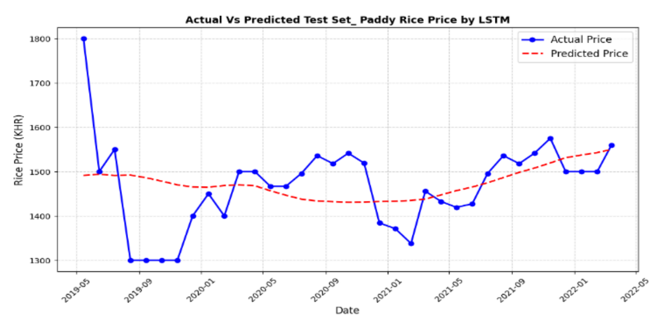

Fig. 14. Actual vs Predicted Test Set of Rice Prices using LSTM

Figure 14 above illustrates the comparison between the actual and predicted rice prices using LSTM. The model successfully followed the overall trend but showed some lags, resulting in larger error gaps in response to rapid price changes.

4.5 Comparison of the Models

The performance models applied were evaluated using standard accuracy metrics, which are summarized in Table 2.

Table 2. Performance metrics for the models.

Model	AIC	BIC	RMSE	MAE	MAPE
ARIMA (1,0,0)	-230.98	-222.03	101.17	76.83	5.21%
ARIMAX (1,0,0)	1888.16	1897.11	96.01	74.65	5.07%
ARIMAX (1,1,1)	1855.39	1870.20	94.32	69.52	4.79%
ARIMAX (1,1,2)	1845.08	1862.82	55.62	45.89	3.85%
XGboost baseline	-	-	40.05	33.69	2.28%
XGBoost- tuned	-	-	**29.91**	**25.46**	**1.70%**
LSTM	-	-	98.59	76.15	5.16%

Table 2 indicates that among time series models, ARIMAX (1,1,2) shows the best predictive performance, while ARIMA (1,0,0) has the lowest AIC/BIC but higher forecast errors. Specifically, we aim to add an exogenous study to enhance model accuracy and capture external influences such as exchange rates and crude oil prices that significantly affect rice price fluctuations. Overall, among the models, XGBoost with tuning demonstrated the best performance with the lowest RMSE, MAE, and MAPE, indicating superior accuracy in forecasting paddy rice prices in Battambang Province. The ARIMAX (1,1,2) produced moderate results, reflecting its effectiveness in capturing linear patterns but its limitation in modeling more complex structures. The LSTM model, while suitable for capturing long-term dependencies, yielded the highest error rates in this study, potentially due to the limited dataset size or suboptimal model tuning.

4.6 Feature Importance

Feature importance measured by Gain shows how much each variable helps reduce prediction error. This analysis was conducted on XGBoost, which is the best model.

Table 3. Importance Score.

Feature	Gain (Importance Score)	% Importance score
Crude oil avg_$/bbl	12,448.67	57.08%
Exchange Rate	9,352.97	42.92%

Based on the results in Table 3 above, crude oil showed the highest importance with a gain score of 12,448.67, accounting for 57.08% of the total contribution to the model. This indicates that changes in global energy prices have a significant effect on rice prices in Cambodia. The exchange rate also played a major role, contributing 42.92% to the model's accuracy. These results highlight how sensitive rice prices are to both global crude oil costs and currency fluctuations.

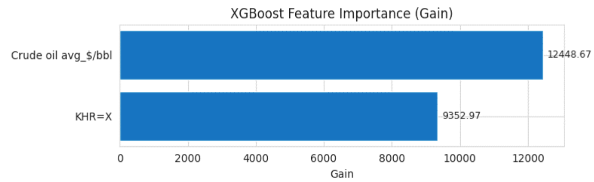

Fig. 15. Feature Importance (Gain Score) _XGBoost

4.7 Forecasting Paddy Rice Prices

Figure 16 below shows the forecast of paddy rice prices for the next 12 months in Battambang Province, using the XGBoost-tuned model. The blue line represents the combined historical paddy rice prices, while the shaded red area indicates the forecasted period starting from mid-2022. The model projects a relatively stable price trend, with slight fluctuations in Khmer riels/kg. This suggests price stabilization following previous periods of high volatility.

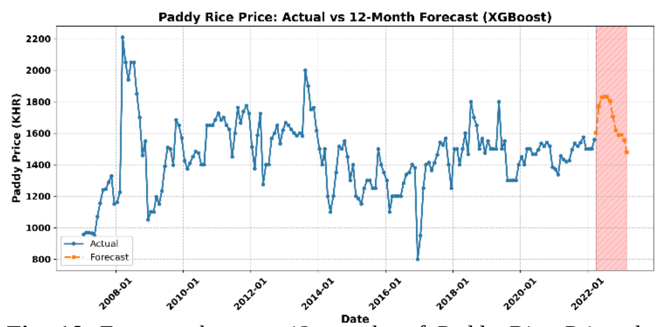

Fig. 16. Forecast the next 12 months of Paddy Rice Prices by XGBoost.

The paddy rice price forecast shows a sharp increase in the early months of the period, peaking at around 1,830 KHR in July to August 2022. Prices are then expected to decline steadily to about 1,470 KHR by March 2023. This pattern suggests a short-term upward movement, likely influenced by external factors and broader market dynamics, followed by a gradual downturn driven by post-harvest supply, demand changes, or macroeconomic conditions. Overall, the results indicate that the market may experience significant volatility over the year.

4.8 Relationship to SDG Implications

The forecast models and Battambang's rice price data show trends that have a direct impact on both income stability (SDG 1) and food security (SDG 2). This study offers useful insights that might assist policymakers in developing methods that stabilize rice prices and provide fair access to food resources throughout the region by precisely forecasting price variations. These strategies should be taken into consideration with additional research and exploration [10].

5. CONCLUSIONS

This study evaluates ARIMAX, XGBoost, and LSTM models for forecasting paddy rice prices in Battambang Province, Cambodia. The results highlight that XGBoost tuned delivered the best accuracy, with the lowest RMSE (29.90 KHR/Kg), MAE (25.47 KHR/Kg), and MAPE (1.70%), demonstrating its strength in capturing complex, nonlinear price patterns. Among ARIMAX variants, the (1,1,2) model performed best (RMSE: 55.62 KHR/Kg). LSTM had the highest error rates (RMSE: 98.59 KHR/Kg, MAPE: 5.16%), likely due to limited data and tuning challenges. Tuned XGBoost reduced RMSE by 46.20%, MAE by 44.50%, and MAPE by 55.80%, and outperformed LSTM with improvements exceeding 69% across all metrics.. These findings highlight the superior accuracy of machine learning methods like XGBoost in agricultural price forecasting and robustness in capturing complex price dynamics due to its ability to incorporate external variables and capture complex nonlinear patterns. Importantly, the model will be maintained and updated in response to new availability to enhance its accuracy and applicability for future predictions.

Additionally, by enhancing the accuracy of paddy rice price forecasts, this study contributes to Sustainable Development Goals by supporting food security (SDG 2) through better market stability and empowering farmers with timely information for economic decision-making (SDG 8), thereby promoting sustainable agricultural development and poverty reduction in Cambodia.

While the current work focuses on comparative forecasting performance, future research should expand these models by incorporating additional socioeconomic and environmental factors, exploring hybrid models combining time series analysis with deep learning for improved accuracy, and advancing sustainable agricultural development in Cambodia and beyond.

ACKNOWLEDGMENTS

The authors would like to extend their expression of sincere gratitude to the World Food Program (WFP) for providing access to the valuable rice price data. We would also like to acknowledge the ongoing support, consultation, inspiration, and thorough review from our academic advisors and co-advisors in the Master’s Program in Data Science

(M-DAS), Department of Applied Mathematics and Statistics (AMS), at the Institute of Technology of Cambodia (ITC). Their expertise, dedication, and support significantly contributed to the success of this study.

REFERENCES

[1] "Sustainable Development Goals: Progress and Insights in Cambodia," UNDP Cambodia Office, Phnom Penh, Cambodia, 2021.

[2] G. Benyamin and G. Ali , "Recurrent Neural Networks and Long Short-Term Memory Networks," Cornell University Library (arXiv.org), Ithaca, New York, USA, 2023.

[3] FAO, "GIEWS Country Brief: The Kingdom of Cambodia," Food and Agriculture Organization of the United Nations (FAO), Global Information and Early Warning System (GIEWS), Phnom Penh, Cambodia, 2024.

[4] S. G N and G. S. Sheshadri, "ARIMAX model for short-term electrical load forecasting," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 4, p. 2786–2790, November 2019.

[5] S. Heng, S. Sovannarith and P. Chin, "Factors influencing paddy producers' profitability and market sales in Battambang province, Cambodia," *Cogent Food & Agriculture*, vol. 9, no. 1, 2023.

[6] M. J. Matthew, C. A. Carolyn and S. Gianna , "Time-Series Methods for Forecasting and Modeling Uncertainty in the Food Price Outlook," U.S. Department of Agriculture (USDA), Economic Research Service (ERS), Washington, D.C., USA. , 2022.

[7] L. Qing, L. Chengfang and Z. Wenbin, "Understanding the Impact of Government Subsidies on Green Technology Innovation," *Australian Journal of Agricultural and Resource Economics*, vol. 64, no. 4, 2020.

[8] N. Rabia, L. Gasmi , L. Nabil and S. Kabou, "Optimizing financial time series predictions with hybrid ARIMA, LSTM, and XGBoost models," *Studies in Engineering and Exact Sciences*, vol. 5, no. 2, 2024.

[9] C. Tianqi and G. Carlos, "XGBoost: A Scalable Tree Boosting System," in *Conference Publisher: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, 2016.

[10] UNESCAP, "Goal Profile: SDG 8 – Decent Work and Economic Growth," UNESCAP, United Nations Economic and Social Commission for Asia and the Pacific, Bangkok, Thailand, 2021.