

## OPTIMIZING CAYENNE PEPPER PRICE FORECASTING USING HYBRID SARIMAX-LSTM MODEL FOR FOOD SECURITY

Adi Supriyatna<sup>1\*</sup>; Mari Rahmawati<sup>1</sup>; Burhanudin Rabbani<sup>1</sup>; Asta Wenang<sup>1</sup>; Sulthan Adly<sup>1</sup>

Faculty of Engineering and Informatics<sup>1</sup>  
Universitas Bina Sarana Informatika, Jakarta, Indonesia<sup>1</sup>  
<https://www.bsi.ac.id/id/><sup>1</sup>  
adi.asp@bsi.ac.id\*, mari.mrw@bsi.ac.id, brnhrabbani23@gmail.com,  
astawenang12@gmail.com, sulthanadly2@gmail.com

(\*) Corresponding Author  
(Responsible for the Quality of Paper Content)



The creation is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License.

**Abstract**— The price volatility of cayenne pepper in traditional markets significantly impacts household purchasing power and regional inflation. While traditional statistical models can capture seasonal patterns, they often fail to model complex non-linear fluctuations driven by external factors such as weather anomalies and national holidays. To address these limitations, this study proposes a hybrid SARIMAX-LSTM model. The Seasonal AutoRegressive Integrated Moving Average with eXogenous variables (SARIMAX) component is utilized to model linear structures, seasonality, and the influence of exogenous variables (temperature, rainfall, and holidays), whereas the Long Short-Term Memory (LSTM) component specifically models the remaining non-linear patterns within the residuals. Daily data comprising chili prices, weather metrics, and holiday schedules were employed to train and test the model using Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) as performance metrics. Experimental results demonstrate that the proposed hybrid model significantly outperforms the single SARIMAX baseline model, reducing the RMSE by 26.7% (from 11.09 to 8.13) and MAPE by 28.6% (from 23.45% to 16.74%). This approach not only provides a more accurate and robust decision-support tool for price stability but also contributes to the advancement of artificial intelligence-based hybrid methods in the domain of food security.

**Keywords:** Cayenne Pepper, Food Price Prediction, Hybrid Model, LSTM, SARIMAX.

**Intisari**— Volatilitas harga cabai rawit di pasar tradisional secara signifikan mempengaruhi daya beli rumah tangga dan inflasi daerah. Meskipun model statistik tradisional mampu menangkap pola musiman, model tersebut seringkali gagal memodelkan fluktuasi non-linear yang kompleks akibat faktor eksternal seperti anomali cuaca dan hari libur nasional. Untuk mengatasi keterbatasan tersebut, penelitian ini mengusulkan model hybrid SARIMAX-LSTM. Komponen Seasonal AutoRegressive Integrated Moving Average with eXogenous variables (SARIMAX) digunakan untuk memodelkan struktur linier, musiman, dan pengaruh variabel eksogen (suhu, curah hujan, hari libur), sementara komponen Long Short-Term Memory (LSTM) secara spesifik memodelkan pola non-linear yang tersisa pada sisaan (residual). Data harian harga cabai, cuaca, dan hari libur digunakan untuk melatih dan menguji model menggunakan metrik Root Mean Squared Error (RMSE) dan Mean Absolute Percentage Error (MAPE). Hasil eksperimen menunjukkan bahwa model hybrid yang diusulkan secara signifikan mengungguli model pembandingan SARIMAX tunggal, dengan mengurangi nilai RMSE sebesar 26.7% (dari 11.09 menjadi 8.13) dan MAPE sebesar 28.6% (dari 23.45% menjadi 16.74%). Pendekatan ini tidak hanya menawarkan alat bantu pengambilan keputusan yang lebih akurat dan robust untuk stabilitas harga, tetapi juga memberikan kontribusi pada pengembangan metode hybrid berbasis kecerdasan buatan di bidang ketahanan pangan.

**Kata Kunci:** Cabai Rawit, Prediksi Harga Pangan, Model Hybrid, LSTM, SARIMAX.

## INTRODUCTION

The instability of food commodity prices poses a significant challenge in maintaining national economic stability and public welfare [1]. One commodity contributing substantially to food price fluctuations is cayenne pepper (*Capsicum frutescens*), which consistently exhibits high price volatility in Indonesian traditional markets [2]. As a staple ingredient in household consumption and the culinary industry, fluctuations in cayenne pepper prices have direct repercussions on inflation and consumer purchasing power [3]. Data from the Central Statistics Agency (BPS) indicates that cayenne pepper is a primary contributor to fluctuations in the Consumer Price Index (CPI). The phenomenon of cayenne pepper price volatility is influenced by various factors, including internal data components such as trends and seasonality, as well as external drivers [4]. External factors, including climate-related anomalies that reduce production output and demand surges during national holidays, significantly contribute to market instability by disrupting the equilibrium between supply and demand [5], [6]. These conditions render chili availability unstable, triggering unpredictable price spikes that necessitate forecasting models capable of accommodating such exogenous variables [7].

In recent years, various time-series forecasting methods have been applied to predict agricultural commodity prices [8], [9]. The Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors (SARIMAX) has proven effective in modeling data with seasonal patterns while incorporating the influence of external variables [10]. For instance, a study by Nasirudin and Dzikrullah (2023) effectively applied the SARIMAX model to forecast chili prices in Indonesia. Their study demonstrated that SARIMAX, by incorporating external variables such as rainfall, inflation, and Google Trends data, yielded more accurate forecasts (MAPE 6.889%) compared to the standard SARIMA model (MAPE 7.630%) [11]. Nevertheless, SARIMAX possesses fundamental limitations due to its assumption of linearity within the data, often failing to capture the complex, non-linear fluctuation patterns common in commodity price data [12].

Conversely, deep learning-based models such as Long Short-Term Memory (LSTM) offer distinct advantages in learning long-term dependencies and complex non-linear patterns from time-series data [13]. Research by Yun et al. (2024) highlighted the superiority of LSTM in predicting agricultural commodity prices over

traditional statistical models [14]. However, pure LSTM models are frequently regarded as "black boxes" and do not explicitly separate linear components, seasonality, or the impact of external variables, making them difficult to interpret and occasionally less accurate when seasonal patterns are highly dominant [15]. Recognizing the limitations inherent in single models, hybrid approaches combining statistical and deep learning models have emerged to enhance forecasting performance. For example, Fiskin et al. (2022) successfully demonstrated that a hybrid SARIMAX-ANN model improved forecasting accuracy for domestic cargo volume data. By leveraging SARIMAX to capture linear patterns and Artificial Neural Networks (ANN) to model the remaining non-linear residuals, this hybrid model proved superior to the single SARIMAX model [16].

From the literature review, a clear research gap is identified: although hybrid models show significant potential, their application to food commodity price data in Indonesia, specifically influenced by weather factors and national holidays, remains limited. Most research continues to focus on either statistical or deep learning models in isolation. The fundamental challenge lies in designing an integrated prediction system capable of simultaneously combining the strengths of statistical models in handling seasonality and deep learning models in capturing non-linear complexity to support food price stabilization policies. Therefore, this study aims to design and implement a hybrid SARIMAX-LSTM model to enhance the accuracy of cayenne pepper price forecasting at Lembang Market, Ciledug, Tangerang City.

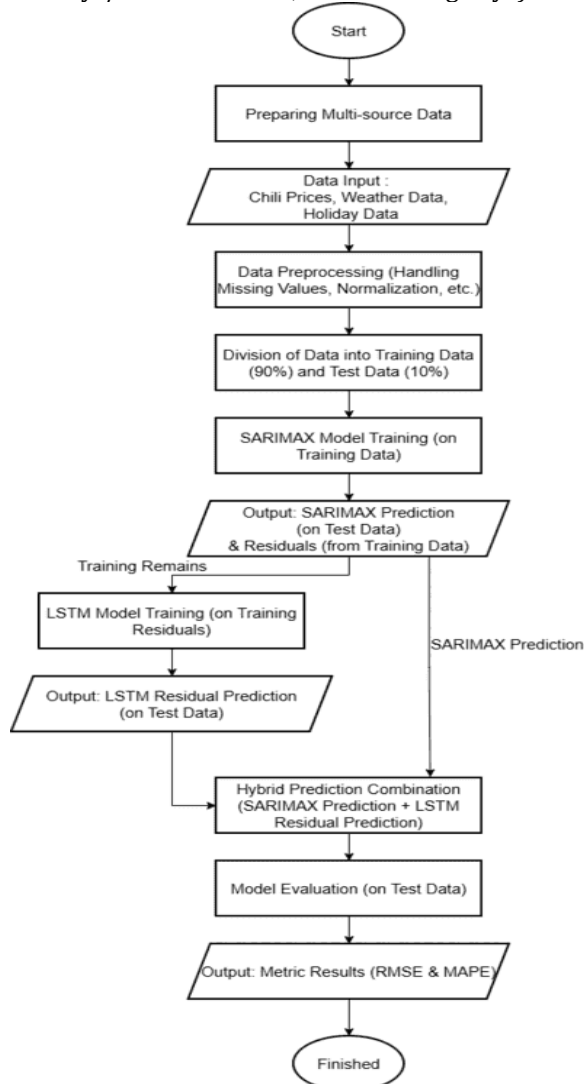
## MATERIALS AND METHODS

This research was conducted through a structured series of stages, beginning with the collection of time-series data on cayenne pepper prices alongside exogenous variables (weather and holidays) from various sources, including the Meteorology, Climatology, and Geophysics Agency (BMKG), the National Strategic Food Price Information Center (PIHPS), Visualcrossing, and the Coordinating Ministry for Human Development and Cultural Affairs (Kemenkopmk). The research framework is illustrated in Figure 1.

Weather variables utilized include average temperature in degrees Celsius (°C) and rainfall in millimeters (mm), obtained from BMKG and Visualcrossing (South Tangerang Climatology Station). Data regarding national holidays were derived from the Joint Decree (SKB) of 3 Ministers

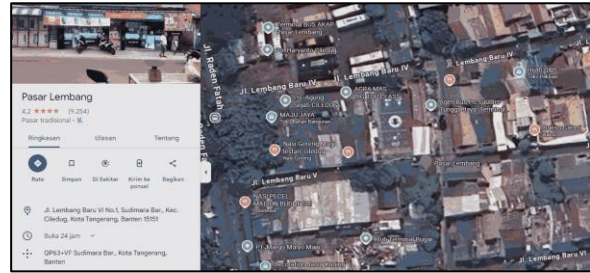


and represented in binary format (value 1 for holidays/collective leave, 0 for working days).



Source: (Research Results, 2025)  
 Figure 1. Research Framework

The dataset employed originated from the National Strategic Food Price Information Center (PIHPS), covering daily data from January 1, 2022, to December 31, 2024. This timeframe was selected to ensure the model captures the most recent economic dynamics, specifically the post-pandemic recovery phase and contemporary climate anomalies that directly influence agricultural productivity in the Tangerang region. This study focuses on price data from Lembang Market, Ciledug, Tangerang City, Banten Province. Lembang Market was selected as a case study due to its status as a vital traditional trading hub in the Ciledug area, serving the needs of diverse community strata in the border region of Tangerang City and South Jakarta. The market is located at approximately  $-6.2377979^{\circ}$  S,  $106.7026491^{\circ}$  E.



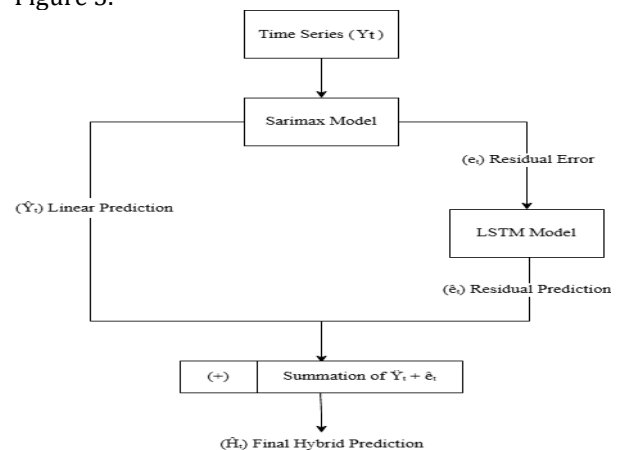
Source: (Research Results, 2025)  
 Figure 2. Location of Lembang Market Ciledug

**Data Pre-processing** The pre-processing stage included handling missing values using linear interpolation for prices, as imputation is a standard method for handling data gaps where values are estimated based on historical data [17]. Subsequently, data normalization was applied using the Min-Max Scaler to transform all numerical variables (price, temperature, rainfall) into a range between 0 and 1 to enhance computational stability. The data was partitioned into a training set (90%) and a testing set (10%).

**Hybrid SARIMAX-LSTM Model Architecture** The proposed model architecture is a two-stage hybrid model. This approach is grounded in the hypothesis that commodity price time-series data contain both linear components (trends and seasonality) and non-linear components (random and complex fluctuations) [18]. The first stage employs SARIMAX to capture linear components, seasonality, and the influence of exogenous variables. The residuals from the SARIMAX model, assumed to contain non-linear patterns, are then extracted. The residual is calculated using the following formula:

$$et = Yt - \hat{Y}t \quad (1)$$

The model architecture flowchart is shown in Figure 3.



Source: (Research Results, 2025)  
 Figure 3. Hybrid Model Architecture Flowchart

**SARIMAX Modeling** The first stage involves modeling using Seasonal AutoRegressive Integrated Moving Average with exogenous regressors (SARIMAX). This model is highly suitable for chili price data, which is influenced by seasonal factors and external variables such as weather and holidays [19]. The model order was determined automatically using the *auto\_arima* function based on the lowest Akaike Information Criterion (AIC) value.

**LSTM Modeling on Residuals** The second stage utilizes LSTM to model the residuals ( $\epsilon_t$ ) produced by SARIMAX. The LSTM model was designed with two layers (64 neurons and 32 neurons) and a Dropout layer with a rate of 0.2 to prevent overfitting. The residual time series was transformed into supervised learning sequences using a sliding window technique. A window size (lag) of 7 days was utilized to construct the training samples, enabling the LSTM to learn temporal dependencies from the previous week's errors to predict the next day's non-linear correction.

**Prediction Combination** The final stage involves combining the prediction results from both models [20]. The final hybrid prediction ( $\hat{H}_t$ ) is generated by summing the linear prediction from SARIMAX ( $\hat{Y}_t$ ) with the non-linear residual prediction from LSTM ( $\epsilon_t$ ).

$$\hat{H}_t = \hat{Y}_t + \hat{\epsilon}_t \quad (2)$$

**Evaluation Scenario** Model performance evaluation was conducted using Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) metrics:

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (Y_t - \hat{H}_t)^2} \quad (3)$$

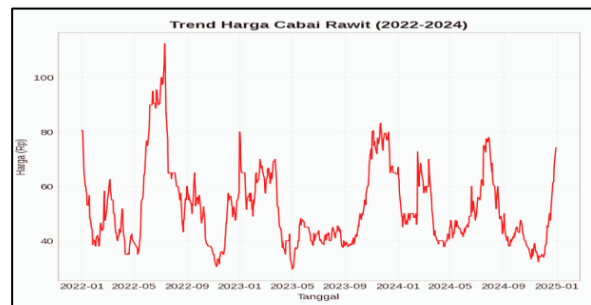
$$MAPE = \frac{1}{n} \sum_{t=0}^n \left| \frac{Y_t - H_t}{Y_t} \right| \times 100\% \quad (4)$$

For equation (3), where  $n$  represents the total number of observations,  $Y_t$  denotes the actual value at time  $t$ , and  $\hat{H}_t$  represents the predicted value at time  $t$ . RMSE measures the square root of the average squared differences between actual and predicted values, providing an indication of the model's prediction accuracy. And for equation (4) where  $n$  represents the total number of observations,  $Y_t$  denotes the actual value at time  $t$ , and  $\hat{H}_t$  represents the predicted value at time  $t$ . MAPE measures the average percentage error between actual and predicted values, allowing

interpretation of the model's performance in percentage terms.

## RESULTS AND DISCUSSION

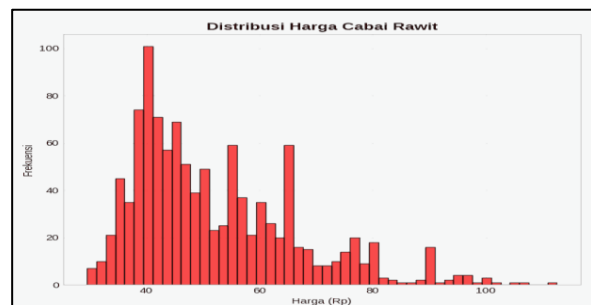
This section presents the results of exploratory data analysis, model implementation, and performance evaluation. The time-series visualization of cayenne pepper prices in Figure 4 reveals significant fluctuations and high volatility without a clear long-term trend.



Source: (Research Results, 2025)

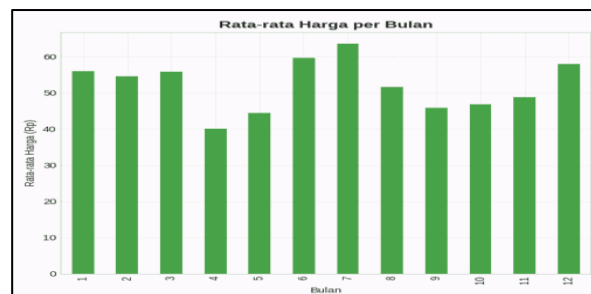
Figure 4. Cayenne Pepper Price Trend (2022-2024)

The price distribution exhibits right-skewness, where the majority of data is concentrated at lower values, yet extreme spikes exist. Monthly seasonal patterns in Figure 6 indicate prices tend to be higher in mid-year and year-end periods.



Source: (Research Results, 2025)

Figure 5. Cayenne Pepper Price Distribution



Source: (Research Results, 2025)

Figure 6. Average Price per Month



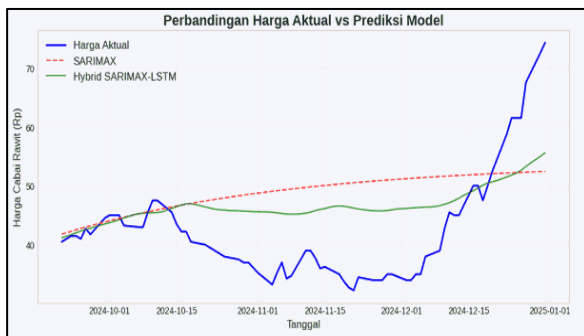
Based on the modeling, *auto\_arima* identified SARIMAX(2,0,0)x(0,0,1)[7] as the best model with the lowest AIC on the training data. The automated selection prioritized the ARIMA(2,0,0)x(0,0,1)[7] structure to maintain a parsimonious model. Exogenous variables temperature, rainfall, and holidays serve as critical filters that stabilize the baseline, allowing the residuals to purely reflect the complex non-linear noise for the LSTM stage. The performance of the Hybrid SARIMAX-LSTM model was then compared with the single SARIMAX model on the test data. The evaluation results are presented in Table 1.

Table 1. Model Performance Comparison on Test Data

Model	RMSE (Rp)	MAPE (%)
SARIMAX Tunggal	11.09	23.45
SARIMAX-LSTM Hybrid	8.13	16.74

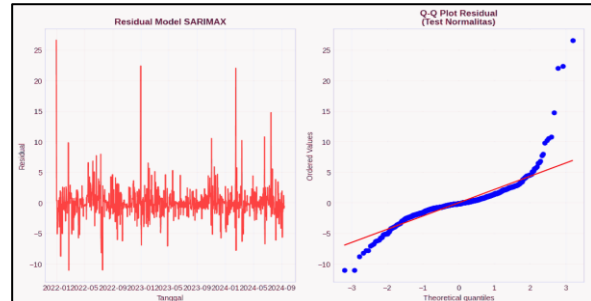
Source : (Research Results, 2025)

According to Table 1, the Hybrid SARIMAX-LSTM model demonstrates significantly superior performance, reducing RMSE by 26.7% and MAPE by 28.6% compared to the single SARIMAX. From a cost-benefit perspective, the 28.6% improvement in MAPE justifies the higher computational complexity of the hybrid model. While more resource-intensive than standalone models, its implementation is feasible for regional government agencies utilizing cloud-based data infrastructures to facilitate accurate market interventions. This accuracy improvement is visualized in Figure 7, where the hybrid prediction curve aligns more closely with actual price fluctuations, particularly during sharp price changes. As demonstrated in Figure 7 and Figure 10, the LSTM component acts as a non-linear corrector that successfully identifies turning points and sharp price spikes, which are often smoothed over by the linear baseline of the standalone SARIMAX model.



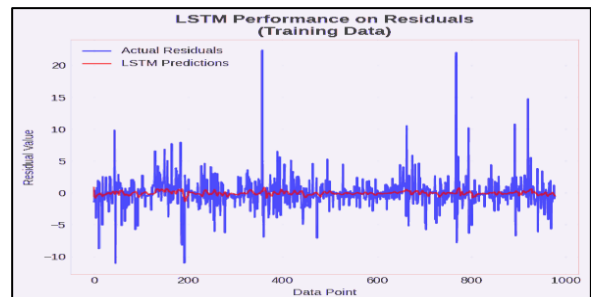
Source: (Research Results, 2025)  
 Figure 7. Comparison of Actual vs Predicted Prices

Diagnostic analysis in Figure 8 shows that the SARIMAX model residuals are not normally distributed, confirming the presence of uncaptured non-linear patterns.



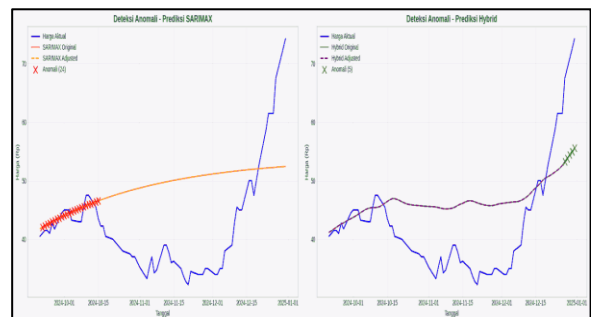
Source: (Research Results, 2025)  
 Figure 8. SARIMAX Model Prediction and Residual Q-Q Plot

The LSTM model successfully learned the complex patterns from these residuals (Figure 9), providing the necessary correction for the final prediction.

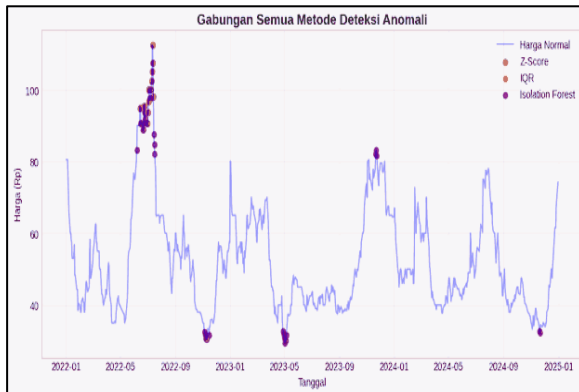


Source: (Research Results, 2025)  
 Figure 9. LSTM Performance on Residuals (Training Data)

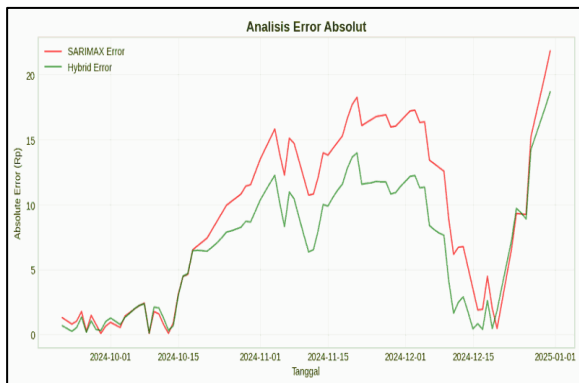
In addition to accuracy, the hybrid model also exhibits better stability in anomaly detection (Figure 10), detecting only 5 unnatural price spikes compared to 24 in the single SARIMAX model. This indicates the hybrid model is more robust.



Source: (Research Results, 2025)  
 Figure 10. SARIMAX Prediction and Hybrid Prediction

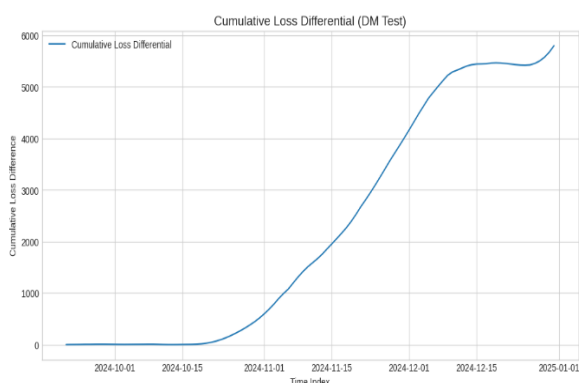


Source: (Research Results, 2025)  
Figure 11. Combined Methods in Detecting Anomalies



Source: (Research Results, 2025)  
Figure 12. Absolute Error Analysis of SARIMAX and Hybrid

To rigorously validate the forecasting superiority, a Diebold-Mariano (DM) test was performed. The test yielded a DM statistic of 9.9741 ( $p < 0.001$ ), confirming that the hybrid model's accuracy gain is statistically significant and sustained throughout the testing period, as visualized in the Cumulative Loss Differential (Figure 13).



Source: (Research Results, 2025)  
Figure 13. Cumulative Loss Differential

## CONCLUSION

Based on the modeling and evaluation results, it is concluded that the hybrid SARIMAX-LSTM model is significantly superior in forecasting daily cayenne pepper prices compared to the single SARIMAX model, with a reduction in RMSE of 26.7% and MAPE of 28.6%. This advantage stems from the two-stage architecture where SARIMAX captures linear patterns and the influence of exogenous variables (weather and holidays), while LSTM effectively predicts complex non-linear residual patterns. This study confirms that the hybrid approach constitutes a robust framework for volatile food price data. For future research, it is recommended to incorporate other external variables such as logistics costs (fuel prices) and inflation, as well as to validate the model on traditional market data in other regions to test model generalization. Future studies should incorporate broader economic indicators, such as logistics costs driven by fuel price fluctuations and supply chain stability, to further refine forecasting precision under diverse economic shocks. The development of a real-time forecasting system based on this model is also highly recommended to support decision-making for farmers and the government. It is important to note that while robust, the model's resilience against 'black swan' events, such as sudden policy shifts or catastrophic natural disasters, remains a challenge. Thus, ongoing validation across various commodities and regions is recommended to ensure broad applicability.

## REFERENCE

- [1] M. del R. Venegas, J. Feregrino, N. Lay, and J. F. Espinosa-Cristia, "Food Financialization: Impact of Derivatives and Index Funds on Agri-Food Market Volatility," *Int. J. Financ. Stud.*, vol. 12, no. 4, 2024, doi: 10.3390/ijfs12040121.
- [2] N. M. Ginting, A. R. Lubis, and M. Zentrato, "Analisis Volatilitas, Integrasi Pasar Dan Transmisi Harga Cabai Merah Di Provinsi Sumatera Utara, Indonesia," *Agro Bali Agric. J.*, vol. 6, no. 3, pp. 827-839, 2023, doi: 10.37637/ab.v6i3.1519.
- [3] O. Helbawanti, W. A. Saputro, and A. N. Ulfa, "Pengaruh Harga Bahan Pangan Terhadap Inflasi Di Indonesia," *AGRISAINTEFIKA J. Ilmu-Ilmu Pertan.*, vol. 5, no. 2, pp. 107-116, 2021, doi: 10.32585/ags.v5i2.1859.
- [4] Y. J. Siregar, R. Hartono, and A. E. Hardana, "Peramalan Harga Cabai Rawit Di Kota



- Malang Dengan Metode Holt-Winters Exponential Smoothing,” *Agricore J. Agribisnis dan Sos. Ekon. Pertan. Unpad*, vol. 6, no. 2, pp. 99–110, 2021, doi: 10.24198/agricore.v6i2.34778.
- [5] M. M. Rahman, R. Nguyen, and L. Lu, “Multi-level impacts of climate change and supply disruption events on a potato supply chain: An agent-based modeling approach,” *Agric. Syst.*, vol. 201, pp. 1–34, 2022, doi: 10.1016/j.agry.2022.103469.
- [6] E. Obermair, A. Holzapfel, and H. Kuhn, “Operational planning for public holidays in grocery retailing - managing the grocery retail rush,” *Oper. Manag. Res.*, vol. 16, no. 2, pp. 931–948, 2023, doi: 10.1007/s12063-022-00342-z.
- [7] D. W. L. Lestari and S. K. Dini, “Forecasting The Price Of Shallots And Red Chilies Using The ARIMAX Model,” *EKSAKTA J. Sci. Data Anal.*, vol. 5, no. 1, pp. 42–49, 2024, doi: 10.20885/eksakta.vol5.iss1.art5.
- [8] F. N. Fikri and N. Nurochman, “Performance Evaluation of Long Short-Term Memory for Chili Price Prediction,” *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 10, no. 1, pp. 33–47, 2025, doi: 10.14421/jiska.2025.10.1.33-47.
- [9] B. Butar Butar, A. Giffari, Z. D. Putri, M. Karisma, W. Kurniawan, and M. H. Fuad, “Forecasting Rice Prices Using the ARIMA Method: A Case Study in DKI Jakarta Province-Belsana Butar Butar et.al Forecasting Rice Prices Using the ARIMA Method: A Case Study in DKI Jakarta Province,” *J. Multidisiplin Sahombu*, vol. 5, no. 02, pp. 299–308, 2025, doi: 10.58471/jms.v5i02.
- [10] E. Nurhasanah, Y. Sukmawaty, and M. Maisarah, “Peramalan Ekspor Migas di Indonesia Menggunakan Pendekatan Seasonal Autoregressive Integrated Moving Average with Exogenous (SARIMAX),” *Indones. J. Appl. Stat.*, vol. 7, no. 1, p. 87, 2024, doi: 10.13057/ijas.v7i1.84934.
- [11] Faris Nasirudin and Abdullah Ahmad dzikrullah, “Pemodelan Harga Cabai Indonesia dengan Metode Seasonal ARIMAX,” *J. Stat. dan Apl.*, vol. 7, no. 1, pp. 105–115, 2023, doi: 10.21009/jsa.07110.
- [12] D. Wang, I. Gryshova, M. Kyzym, T. Salashenko, V. Khaustova, and M. Shcherbata, “Electricity Price Instability over Time: Time Series Analysis and Forecasting,” *Sustain.*, vol. 14, no. 15, pp. 1–24, 2022, doi: 10.3390/su14159081.
- [13] M. Tami and A. Y. Owda, “Efficient commodity price forecasting using long short-term memory model,” *IAES Int. J. Artif. Intell.*, vol. 13, no. 1, pp. 994–1004, 2024, doi: 10.11591/ijai.v13.i1.pp994-1004.
- [14] B. Yun, J. Lai, Y. Ma, and Y. Zheng, “Research on Grain Futures Price Prediction Based on a Bi-DConvLSTM-Attention Model,” *Systems*, vol. 12, no. 6, 2024, doi: 10.3390/systems12060204.
- [15] M. Waqas and U. W. Humphries, “A critical review of RNN and LSTM variants in hydrological time series predictions,” *MethodsX*, vol. 13, no. July, p. 102946, 2024, doi: 10.1016/j.mex.2024.102946.
- [16] C. S. Fiskin, O. Turgut, S. Westgaard, and A. G. Cerit, “Time series forecasting of domestic shipping market: Comparison of SARIMAX, ANN-based models and SARIMAX-ANN hybrid model,” *Int. J. Shipp. Transp. Logist.*, vol. 14, no. 3, pp. 193–221, 2022, doi: 10.1504/IJSTL.2022.122409.
- [17] D. Kurniasari, A. D. Salsabila, M. Usman, and W. Warsono, “Enhancing Weather Forecasting in Bandar Lampung: A Hybrid SARIMA-LSTM Approach,” *JTAM (Jurnal Teor. dan Apl. Mat.)*, vol. 9, no. 1, p. 206, 2025, doi: 10.31764/jtam.v9i1.27188.
- [18] J. Sung, X. Shi, S. Teske, and M. Li, “A hybrid SARIMAX-LSTM model optimised by ANN for near-term forecasting: An application to China’s natural gas consumption,” Centre for Climate Risk and Resilience (CCRR), University of Technology Sydney, Sydney, Australia, 2025. [Online]. Available: <https://opus.lib.uts.edu.au/handle/10453/187869>.
- [19] J. Bana and K. Utnik-bana, “Evaluating a seasonal autoregressive moving average model with an exogenous variable for short-term timber price forecasting,” *For. Policy Econ. J.*, vol. 131, no. July, pp. 1–7, 2021, doi: 10.1016/j.forpol.2021.102564.
- [20] D. Ashtar, S. S. Mohammadi Ziabari, and A. M. M. Alsahag, “Hybrid Forecasting for Sustainable Electricity Demand in The Netherlands Using SARIMAX, SARIMAX-LSTM, and Sequence-to-Sequence Deep Learning Models,” *Sustain.*, vol. 17, no. 16, 2025, doi: 10.3390/su17167192.