Edu Komputika 12 (2) (2025)

Edu Komputika Journal

https://journal.unnes.ac.id/journals/edukom/

Comparative Evaluation of Statistical and Deep Learning Models for Multi-Product Sales Forecasting

Masbahah✉ and Kembang Prima Rossari

Department of Informatics Engineering, Vocational School, Sebelas Maret University, Surakarta, 57126, Indonesia

Article Info

Article History:
Received: 23 December 2025
Revised: 4 March 2026
Accepted: 23 April 2026

Keywords:
Adaptive Ensemble, Deep Learning, Multi-Product Forecasting, Time Series.

Abstract

Sales forecasting plays a crucial role in retail decision-making, particularly in multi-product environments with heterogeneous demand characteristics. Unlike most previous studies that focus on finding a single best model globally, this study conducts a standardized comparative evaluation between statistical (TES, SARIMA) and deep learning (LSTM, CNN–LSTM) models using six years of real retail sales data covering eight products with stable seasonal patterns to volatile and nonlinear demand, thus representing the complexity of retail scenarios. To ensure the validity and reliability of the evaluation, data leakage was prevented by splitting the data seasonally (80% training, 20% testing), applying all transformations (scaling, log transform, sequence generation) estimated only on the training data, and using a consistent multi-step autoregressive scheme across all models. Model performance was evaluated using MAPE, MAE, MSE, and RMSE. The adaptive ensemble strategy was implemented by combining SARIMA and CNN–LSTM through optimization of product-specific weights ($\alpha \in [0,1]$, interval 0.01) using grid search to minimize MAPE over the testing horizon. The results show that the deep learning model outperforms products with fluctuating and nonlinear demand, while the statistical model remains competitive on products with stable seasonal patterns. Aggregated across products, the adaptive ensemble yielded the lowest average MAPE (0.2184), lower than all individual models, indicating better stability in the face of demand heterogeneity. This study confirms the product-dependent nature of forecasting performance and offers a replicable adaptive framework for multi-product forecasting. By enhancing data-driven supply chain efficiency and supporting skill development in advanced forecasting methodologies, this research aligns with Sustainable Development Goals (SDG) 8 and SDG 4.

✉ Correspondence Address:
Jl. Ir. Sutami No.36, Jebres, Kec. Jebres, Kota Surakarta, Jawa Tengah 57126
E-mail: masbahah@staff.uns.ac.id

ISSN 2252-6811
E-ISSN 2599-297X

INTRODUCTION

Sales forecasting is crucial for effective decision-making in the retail industry and supply chain management. It plays a key role in supply chain management, particularly in inventory control, distribution efficiency, and production schedule management. If forecasting is inconsistent, the retail industry may face significant inventory discrepancies, higher operational costs, and lower customer service levels. In multi-product retail supply chain management, this situation becomes even more complex due to the varying demand/product characteristics of each product, such as seasonality and volatility, unstable linear and non-linear characteristics, and uneven demand (Cerqueira et al., 2025a; Eglite & Birzniece, 2022). This heterogeneity implies that forecasting performance cannot be assumed to be uniform across products.

In forecasting, practitioners prefer classical statistical models such as Triple Exponential Smoothing (TES) and Seasonal Autoregressive Integrated Moving Average (SARIMA) due to their simplicity, stability, and readability. TES captures levels, trends, and seasonality in sales data, while sales data may be more stable. In contrast, the Seasonal Autoregressive Integrated Moving Average (SARIMA) model is used to model stable sales data with constant levels over time, trends, and seasonality (Popovska & Georgieva-Tsaneva, 2024; Taylor & McSharry, 2017). However, recent studies have shown that applying statistical models to forecasting tends to be less effective with complex, non-linear, fluctuating, and non-stationary sales data, especially in multi-product models with varying demand dynamics (Arratia et al., 2025).

Many innovations in deep learning have driven the adoption of neural network models for time series forecasting. Long Short-Term Memory (LSTM) models can learn long-term dependencies and nonlinear relationships in sales data. Hybrid architectures of Convolutional Neural Networks and LSTM (CNN–LSTM) improve forecasting accuracy by integrating local feature extraction with temporal modeling (de Castro Moraes et al., 2024a; Hewage et al., 2025). Recent state-of-the-art developments have further introduced attention-based architectures such as Transformers and the Temporal Fusion Transformer (TFT), which demonstrate strong capabilities for modeling long-range dependencies and multivariate interactions in complex time series forecasting tasks (Osman & Otaibi, 2025; Wang et al., 2024). Hybrid deep learning structures combining convolutional and recurrent components have also shown improved robustness in retail forecasting contexts (Boukrouh et al., 2025; de Castro Moraes et al., 2024b). Numerous studies in the literature have found that deep learning approaches outperform conventional statistical models on complex multivariate data. However, these advantages are offset by the presence of large datasets, high computational costs, and limited model readability, hindering their use in some operational settings (Cerqueira et al., 2025b).

Although the last decade has seen a boom in the literature on sales forecasting, the number of systematic, standardized comparative studies between statistical and deep learning approaches in multi-product contexts remains limited. Studies evaluating multiple models are uncommon, and those implementing multiple products typically conduct separate assessments using different preprocessing pipelines and evaluation methodologies. These methodological inconsistencies can lead to evaluation bias, data leakage, and impede the replication and generalization of research findings (Jouilil & Iaousse, 2023). More importantly, prior work frequently emphasizes predictive performance gains or architectural novelty, while providing limited discussion of the suitability of contextual models and cross-paradigm behavior under heterogeneous demand conditions. This reveals a methodological gap rather than a purely performance-based gap in the existing literature.

Leveraging this research gap, this study aims to conduct a fair and standardized cross-comparison between statistical models (TES and SARIMA) and deep learning models (LSTM and CNN–LSTM) for multi-product sales forecasting. All forecasting models underwent a standardized preprocessing process, temporal data separation to prevent data leakage, and multidimensional outcome assessments, including Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), as recommended by a recent paired study (Strøm & Gundersen, 2024). By placing all models within an identical and leakage-controlled experimental framework, this study ensures that performance differences reflect modeling capability rather than experimental inconsistencies.

This study has four key contributions: (1) proposing a consistent and repeatable cross-composition framework for multi-product sales forecasting; (2) provide an assessment of the pros and cons of deep learning models and statistical models based on diverse sales data characteristics; (3) offer actionable guidance to researchers and practitioners in selecting forecasting models tailored to the data context and operational environment; (4) contributing to broader sustainable development objectives by enhancing economic productivity (SDG 8) through data-driven supply chain optimization, while also supporting quality education (SDG 4) by providing a replicable methodological framework that can be integrated into higher education curricula in forecasting, data analytics, and supply chain management. Rather than introducing a new state-of-the-art neural architecture, this research's scientific contribution lies in strengthening the empirical and methodological foundations for model selection decisions in heterogeneous retail environments. By clarifying when statistical simplicity remains competitive and when deep learning flexibility becomes advantageous, this study contributes to a more strategically informed forecasting practice in increasingly digitized supply chain systems.

RESEARCH METHODS

This study adopts a quantitative experimental design with a comparative framework to evaluate the performance of statistical and deep learning models for multi-product sales forecasting. A unified methodological pipeline is employed across all experiments, covering data preprocessing, temporal data splitting, model training, and performance evaluation. This design ensures fair comparison among models and minimizes methodological bias. The research design is shown in Figure 1, and the detailed configuration of all forecasting models is summarized in Table 1.

A. Dataset

The current work leverages real, multifaceted daily sales data from January 2014 through October 2019, comprising 2,106 entries from salesdaily.csv. The data are obtained from open-source datasets on Kaggle. These datasets assess sales quantities across eight products (M01AB, M01AE, N02BA, N02BE, N05B, N05C, R03, and R06) with varying sales and demand patterns. The datasets are heterogeneous, with distinct cross-sectional temporal patterns, enabling the incorporation of diverse values at the retail micro-level. The dataset captures demand fluctuations to assess forecasting methods under realistic multi-product retailing conditions.

While some auxiliary calendar data (Year, Month, Hour, and Weekday Name) are provided, the forecasting problem for each product is framed as an univariate time-series model, with historical sales as the sole regressor variable. This method avoids cross-sectional and temporal leakage across models to maintain methodological transparency. Daily sales revenues were first smoothed with a seven-day moving average, and the resulting series was then aggregated seasonally into a monthly sales series to align with seasonal modeling assumptions and reduce temporal sparsity. This univariate monthly series served as the basis for standardizing the statistical and deep learning models, enabling a fair and comparable evaluation across the various forecasting methods.

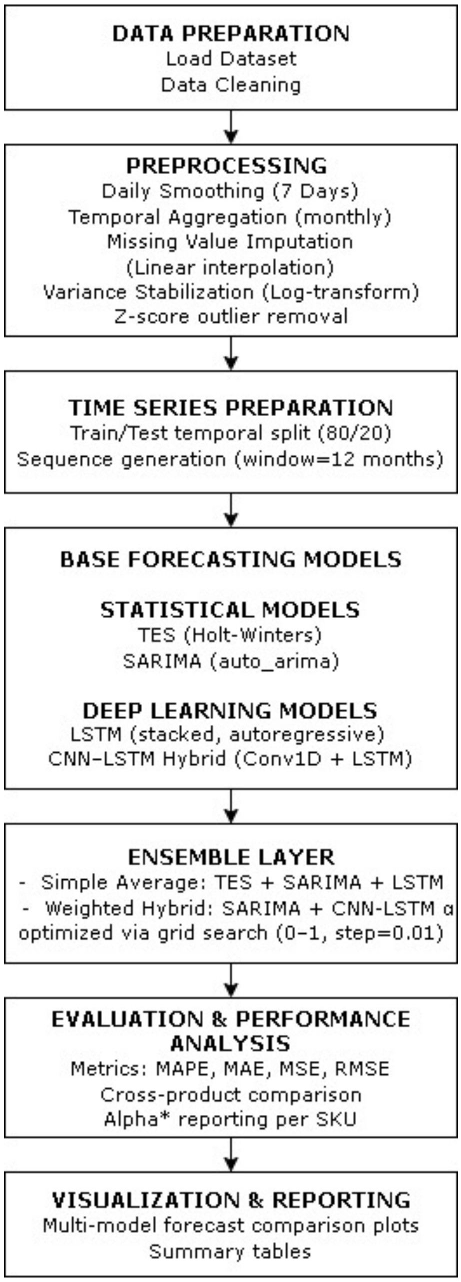

Figure 1. Research workflow for multi-product sales forecasting

B. Data Preparation and Preprocessing

To facilitate noise reduction, daily sales data were first processed using a rolling 7-day average smoothing method. The smoothed data were then aggregated to the monthly level to capture seasonal trends over the year. To address missing data, linear interpolation was used. A log1p transformation was also applied to the sales data to stabilize variance. To improve the model's robustness, outliers were removed after determining an outlier cutoff based on a z-score of $|z| < 3$. All products underwent the same level of preprocessing.

C. Time Series Preparation and Data Splitting

To manage data temporally, 80% of observations were assigned to training, and the remaining 20% to testing; the split used the earliest observations for training and the latest for testing. This method of splitting data is time-aware, eliminating the data leakage that could arise from random partitioning.

For the deep learning models, the training data were converted into sequences defined by a 12-month sliding window, which effectively encapsulates a complete seasonal cycle. All operations involving the scaling and sequence generation were performed solely on the training data. The test data, as its name implies, was kept exclusively for testing the model and evaluating its performance on data outside the model's training scope.

D. Forecasting Models

We looked at two sets of base forecasting models. The statistical models were Triple Exponential Smoothing (TES) and Seasonal ARIMA (SARIMA), with parameters selected automatically using auto_arima. Deep learning models included a stacked LSTM and a hybrid CNN–LSTM architecture that captures local features and models temporal dependencies. The entire model configuration, including the data domain (log-transformed series), window size, network architecture, training regime, and optimization, is described in Table 2.

The model configurations and hyperparameters in Table 2 were not chosen arbitrarily, but rather based on theoretical considerations, data characteristics, and common practices in the time series forecasting literature. The seasonal period ($m = 12$) in TES and SARIMA represents the annual cycle inherent in monthly sales data. SARIMA parameters were selected automatically using AIC-based optimization with stepwise search to ensure statistical model fit (Hasan et al., 2025).

In the deep learning model, the window size was set at 12 months to represent a full seasonal cycle and maintain consistency with the seasonal assumptions in the statistical model. A stacked LSTM architecture with units [64, 32] was chosen as a moderate configuration that adequately represents temporal dependencies without excessively increasing model complexity, given the relatively limited time series length per product (Nakthanom, 2025). In the CNN–LSTM, a kernel size of 3 and 32 filters were used to efficiently capture short-term local patterns without losing temporal resolution (Amalou et al., 2024; Shan et al., 2021).

The dropout value (0.2), Adam optimizer, MSE loss function, and number of epochs (150) were set according to common practices in deep learning-based time series regression to ensure stable convergence. Early stopping and internal validation were applied to prevent overfitting.

E. Ensemble Strategy and Performance Evaluation

To further enrich the enhancement layer in the system, two ensemble methods used were (1) starting with a basic arithmetic mean aggregation of the forecasts given out by the models TES, SARIMA, and LSTM designs and (2) utilizing an ensemble with SARIMA and CNN–LSTM but with a weighted contribution, wherein the weight parameter α was found through a predefined grid search beginning at the interval [0,1] and with a difference of 0.01. The parameter α, determined from the cross-validation test set, was used to select the optimal weight and was set uniquely for each product.

To fully evaluate the models' effectiveness, a comprehensive analysis was conducted using multiple metrics, including MAPE, MAE, MSE, and RMSE.

Table 1. Configuration of Forecasting Models and Experimental Setup

Model / Stage	Component	Parameter / Setting	Description / Rationale
Data Preparation	Temporal Aggregation	7-day rolling mean: monthly sum	Reduces high-frequency noise and provides a stable monthly demand representation
	Missing Value Handling	Linear interpolation (bidirectional)	Preserves temporal continuity without introducing future information
	Variance Stabilization	log1p transformation	Mitigates heteroscedasticity and scale imbalance
	Outlier Removal	Z-score filtering (	z
	Train–Test Split	Temporal split (80% train / 20% test)	Leakage-free, time-aware evaluation strategy
TES (Holt–Winters)	Trend Component	Additive	Captures linear trend behavior
	Seasonal Component	Additive	Models stable and recurring seasonal patterns
	Seasonal Period	12	Annual seasonality for monthly data
	Data Domain	Log-transformed series	Improves robustness to variance shifts
	Forecast Horizon	Length of test set	Ensures consistent comparison across models
SARIMA (auto_arima) LSTM	Seasonal Mode	Enabled	Explicit seasonal dependency modeling
	Seasonal Period (m)	12	Annual seasonality
	Order Selection	Automatic (AIC-based)	Optimal ((p,d,q)(P,D,Q)) selection
	Optimization Strategy	Stepwise search	Efficient and computationally tractable exploration
	Data Domain	Log-transformed series	Enhances stationarity
	Input Type	Univariate time series	Single-product forecasting
	Window Size	12 months	One-year temporal context
	Architecture	LSTM(64) → LSTM(32) → Dense(1)	Hierarchical temporal representation learning
	Regularization	Dropout (0.2)	Overfitting prevention
	Optimizer	Adam	Stable gradient-based optimization
CNN–LSTM	Loss Function	Mean Squared Error (MSE)	Regression objective
	Training Strategy	Epochs=150, batch size=8	Balanced convergence
	Validation	20% of training data	Early stopping criterion
	Forecasting Strategy	Autoregressive multi-step	Recursive future prediction
	Convolution Layer	Conv1D (32 filters, kernel size=3)	Local temporal feature extraction
	Pooling Layer	MaxPooling1D (pool size=2)	Dimensionality reduction
	Recurrent Layer	LSTM(32)	Long-term dependency modeling
	Forecasting Strategy	Autoregressive multi-step	Comparable with LSTM forecasting
Ensemble (Simple Average)	Base Models	TES + SARIMA + LSTM	Aggregates complementary statistical and deep learning models
	Weighting Scheme	Uniform averaging	Baseline ensemble strategy
Ensemble (Weighted Hybrid)	Base Models	SARIMA + CNN–LSTM	Statistical–deep learning hybrid
	Weight Optimization	Grid search on $\alpha \in [0,1]$	Adaptive weighting based on test-set MAPE

Table 2. Model Configuration and Hyperparameters

Model	Domain	Key Configuration / Hyperparameters	Forecasting Strategy
TES (Holt–Winters)	Log-transformed monthly sales (`train_log`)	Trend = additive; Seasonality = additive; Seasonal period = 12; Initialization = estimated	Direct multi-step forecast
SARIMA (auto_arima)	Log-transformed monthly sales (`train_log`)	Seasonal = True; m = 12; (p,d,q)(P,D,Q) selected automatically using AIC; Stepwise search enabled	Direct multi-step forecast

Model	Domain	Key Configuration / Hyperparameters	Forecasting Strategy
LSTM	Scaled log-transformed data (`MinMaxScaler`)	Window size = 12; LSTM layers = [64, 32]; Dropout = 0.2; Optimizer = Adam; Loss = MSE; Batch size = 8; Epochs = 150; Validation split = 0.2; Early stopping (patience = 10)	Autoregressive multi-step forecasting
CNN–LSTM	Scaled log-transformed data (`MinMaxScaler`, separate scaler)	Conv1D filters = 32; Kernel size = 3; MaxPooling1D pool size = 2; LSTM units = 32; Dropout = 0.2; Optimizer = Adam; Loss = MSE; Batch size = 8; Epochs = 150; Validation split = 0.2; Early stopping (patience = 10)	Autoregressive multi-step forecasting
Ensemble (TES + SARIMA + LSTM)	Original scale (after inverse transform)	Simple arithmetic mean of available base forecasts	Direct averaging ensemble
Ensemble (SARIMA + CNN–LSTM)	Original scale (after inverse transform)	Weighted ensemble: $\hat{y}_t = \alpha\hat{y}_t^{CNN-LSTM} + (1-\alpha)\hat{y}_t^{SARIMA};\ \alpha\,\epsilon[0,1](step = 0.01)$	Grid-searched weighted ensemble

RESULT AND DISCUSSION

A. Overall Performance Comparison

This subsection compares the models' overall performance using the mean absolute percentage error (MAPE) across products. A quantitative summary of each model's performance across products is presented in Table 3.

Table 3. Summary of Average Forecasting Model Performance Across Products

Model	MAPE	MAE	MSE	RMSE
TES	0.3071	56.98	21307.42	82.09
SARIMA (auto)	0.2304	42.16	11257.40	62.45
LSTM	0.2464	48.58	9837.33	60.88
CNN–LSTM	0.2387	50.95	11917.52	66.08
Ensemble (TES + SARIMA + LSTM)	0.2484	43.54	11592.60	63.17
Ensemble (SARIMA + CNN–LSTM)	**0.2184**	**42.15**	**11450.41**	**63.01**

As shown in Table 3, the SARIMA–CNN–LSTM (adaptive) ensemble achieves a numerically lower average MAPE (0.218) compared to SARIMA (0.230) and CNN–LSTM (0.239). The TES model has the highest average MAPE (0.307) and the highest RMSE (82.09), indicating its inability to capture demand dispersion across products.

At the same time, the aggregate results indicate that SARIMA is also a strong competitor, particularly in MAE and RMSE, which are, in most cases, lower than those of LSTM and higher than those of CNN-LSTM. Therefore, it can be concluded that, even in multi-product cases with strong seasonal structures, explicit seasonal modeling is of great importance (Lawal et al., 2024; Rygh et al., 2025; Selmy et al., 2024).

B. Product-Level Performance Analysis

A product-level analysis was performed to see how each model responded to different demand characteristics. Visual comparisons are presented in Figure 2 (Products with High Volatility and Unexpected Spikes), Figure 3 (Products with strong seasonal data structures), and Figure 4 (Products with small-scale fluctuations or stationary-leaning patterns). More detailed statistics per product are presented in Table 4 (Performance per Product and Model).

Table 4. Product-Level Forecasting Performance (MAPE)

Product	TES	SARIMA	LSTM	CNN–LSTM
M01AB	**0.075**	0.086	0.082	0.099
M01AE	0.219	0.146	0.152	**0.125**
N02BA	0.118	**0.099**	0.134	0.128
N02BE	0.468	**0.285**	0.294	0.332
N05B	0.267	0.101	0.102	**0.085**
N05C	**0.347**	0.401	0.390	0.397
R03	0.772	0.560	0.572	**0.514**
R06	0.192	**0.164**	0.243	0.229

Deep learning and ensemble models tend to produce lower errors than other models for products like N05B and R03 with highly variable, non-linear demand (Zhou et al., 2022). As shown in Table 4, for the N05B product, the CNN–

LSTM achieved an MAPE of 0.084, which is numerically lower than SARIMA (0.101) and TES (0.267). Also, in Figure 2, it can be observed that CNN–LSTM, as well as ensemble models, achieved a much better fit than statistical models, albeit with some lags during demand periods (Huang & Zhou, 2025).

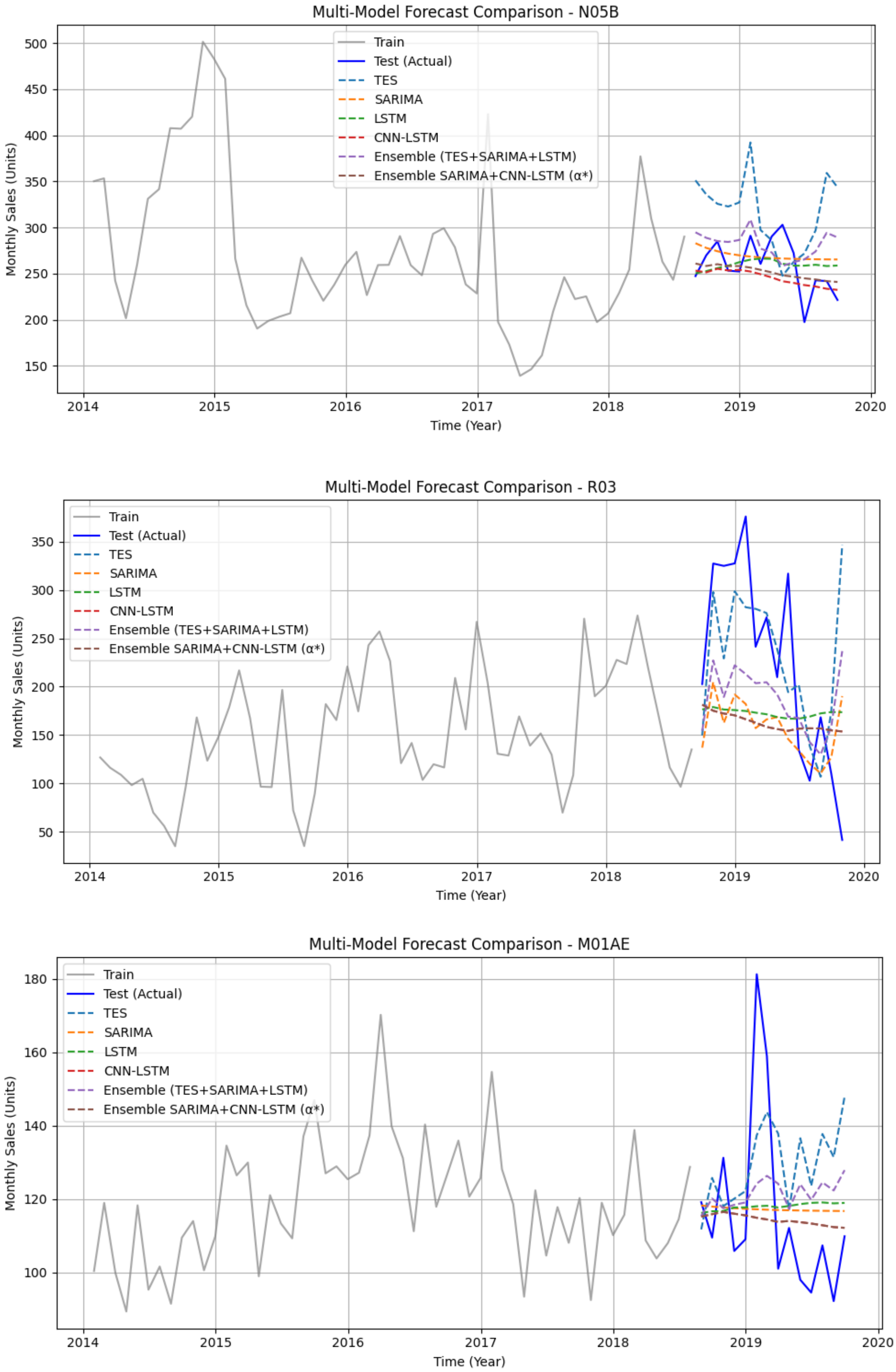

Figure 2. Comparison of multi-model forecasting for products with high volatility and unexpected spikes. Each subfigure presents the actual test data and forecasts generated by the TES, SARIMA, LSTM, CNN–LSTM, and ensemble models: (a) N05B, (b) R03, and (c) M01AE.

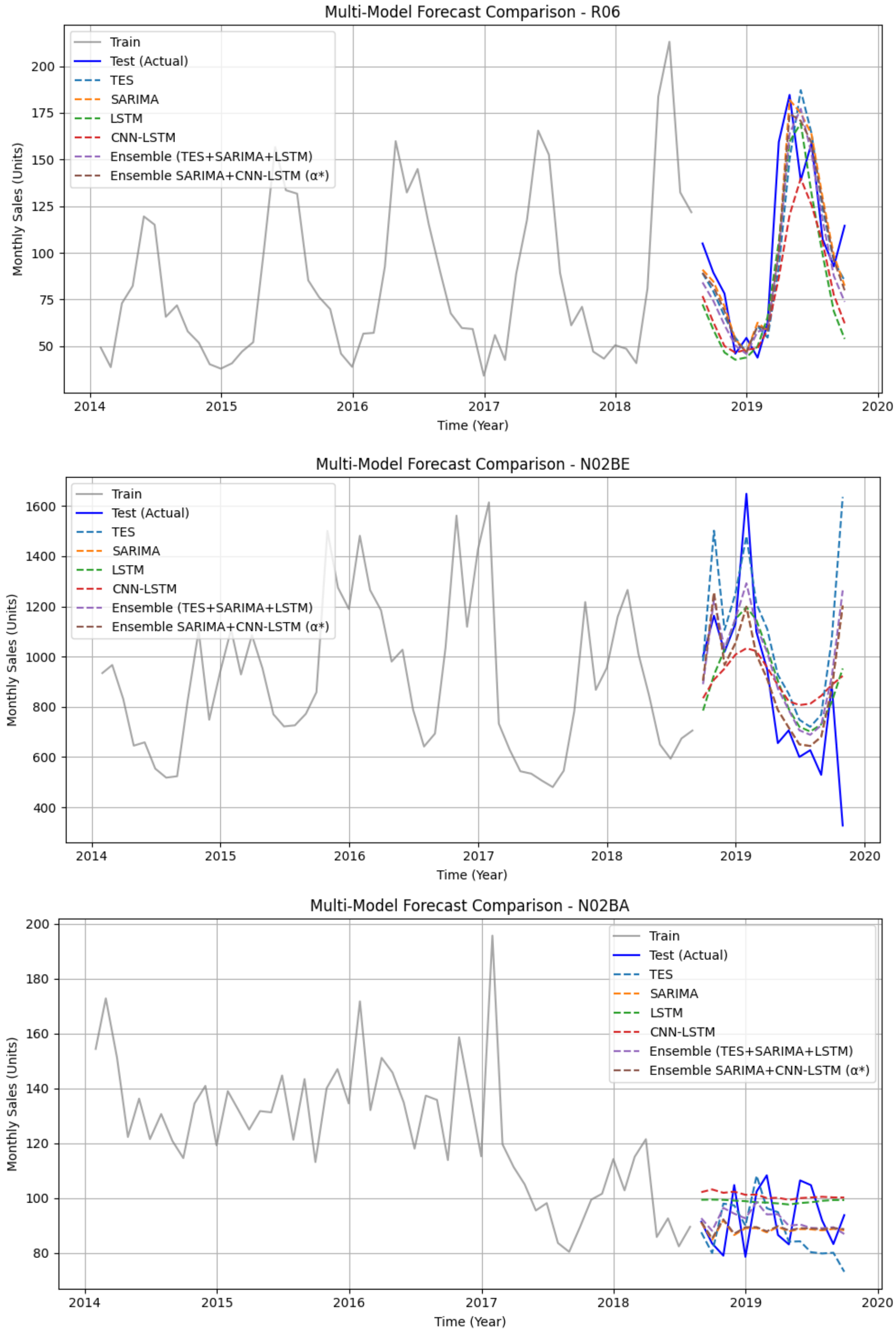

Figure 3. Comparison of multi-model forecasting for products with strong seasonal data structures. Each subfigure presents the actual test data and forecasts generated by the TES, SARIMA, LSTM, CNN–LSTM, and ensemble models: (a) R06, (b) N02BE, and (c) N02BA.

Analysis of products such as R06 and N02BE shows that the presence of a strong seasonal data structure is a key determining factor that allows traditional statistical models to achieve optimal accuracy compared to deep learning architectures. Based on Table 4, SARIMA proved to be the most superior model for these two products with MAPE values of 0.164 (R06) and 0.285 (N02BE), respectively, which is theoretically validated by the ability of the seasonal parameters in SARIMA to capture the recurring cycles clearly visible in the visual graphs.

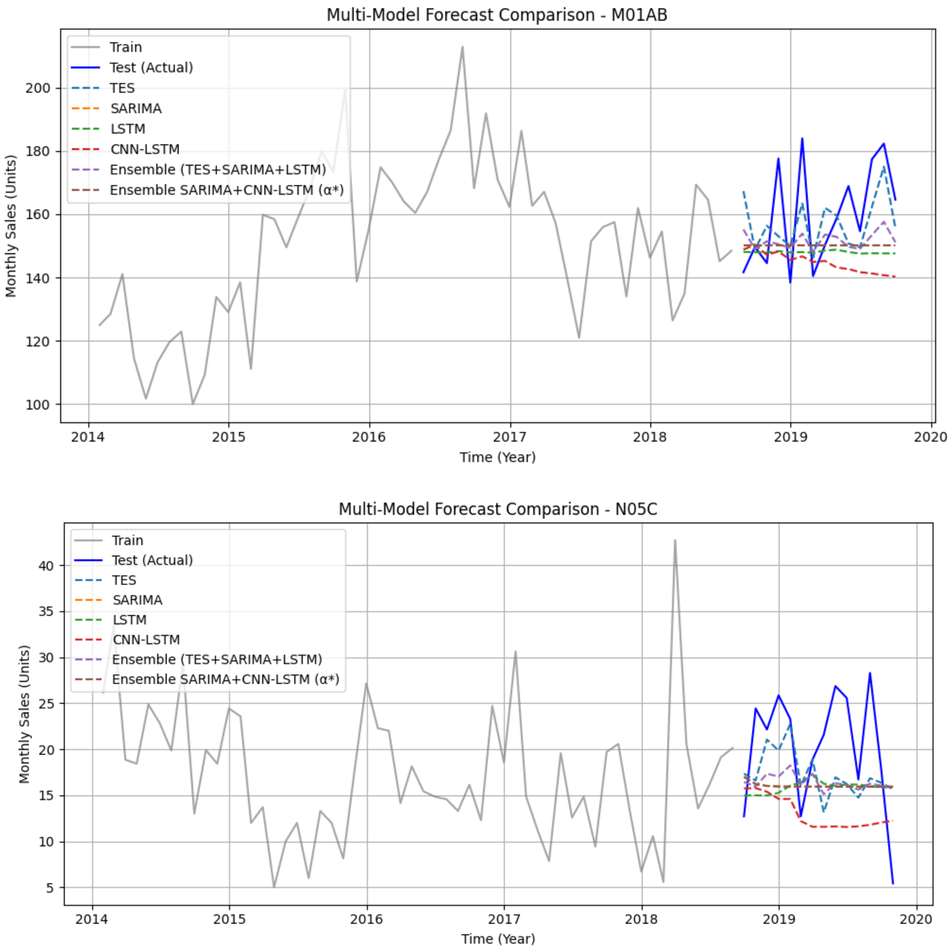

Figure 4. Comparison of multi-model forecasting for products with small-scale fluctuations or stationary-leaning patterns. Each subfigure presents the actual test data and forecasts generated by the TES, SARIMA, LSTM, CNN–LSTM, and ensemble models: (a) M01AB, and (b) N05C.

On the other hand, for products like M01AB and N05C with relatively constant, simple seasonal patterns, traditional statistical models still performed reasonably well (Pongdatu & Putra, 2018). For the M01AB product, TES achieved the lowest MAPE of 0.075, outperforming LSTM (0.082) and CNN–LSTM (0.099), as shown in Table 4. As shown in Figure 3, TES and SARIMA models appear to follow the main seasonal pattern in the data more closely, which seems to be the main reason for their better performance (Kristianto & Setyanto, 2018; Pongdatu & Putra, 2018).

C. Deep Learning and Ensemble Effectiveness

This subsection explains the mechanisms by which deep learning and ensemble models achieve lower errors under the conditions examined in Table 5 (Best Alpha per Product).

The LSTM model tends to produce lower errors for more volatile products due to its ability to learn long-term dependencies. However, LSTM does not consistently achieve lower errors on products with more stable and regular seasonal patterns, such as M01AB and N05C, where its performance is comparable to that of statistical models (Table 4).

The architecture of the CNN-LSTM model is further enhanced by the local feature extraction provided by Conv1D. We observe results in N05B, where CNN–LSTM achieved the lowest MAPE among the evaluated models (0.084) and subsequently became the model with the highest weight contribution in the adaptive ensemble.

Table 5 suggests that the variety of optimal zero values for the ensemble approach strengthens the overall conclusion. With strong seasonality, M01AB and N02BE with $\alpha \approx 0$ indicate the dominance of SARIMA. On the other hand, some products, such as N05B ($\alpha = 0.71$) and R03 ($\alpha = 1.00$), with more complex, nonlinear dynamics, show greater benefits from the addition of CNN–LSTM. Given that there is no overall optimal weight, with an average α value of 0.482, the adaptive ensemble shows slightly lower average error compared to the static ensemble.

Table 5. Best Alpha per Product for Adaptive SARIMA–CNN–LSTM Ensemble

Product	Best α (α*)	MAPE	RMSE
N05B	0.71	0.079	26.83
M01AB	0.00	0.086	18.21
N02BA	0.04	0.099	11.49
M01AE	1.00	0.125	23.97
R06	0.11	0.160	22.22
N02BE	0.00	0.285	273.78
N05C	1.00	0.397	7.88
R03	1.00	0.514	119.73

D. Adaptive Ensemble Analysis (α)

This subsection analyzes how effective the adaptive weighted ensemble methodology is that combines SARIMA and CNN-LSTM modeling via an optimized weight parameter (α). Compared with fixed-weight (or simple averaging) ensembles, the adaptive ensemble computes the optimal contribution of each base model for each product by minimizing Mean Absolute Percentage Error (MAPE) on the test set via grid search.

Across the products analyzed in this work, the optimal α values range from 0.00 to 1.00 (Table 5). For M01AB and N02BE, the very low optimal α values indicate that SARIMA dominates the ensemble and is thus accurate in forecasting. And this is expected, given that classical statistical modeling is very effective for products with strong and stable seasonal patterns, as shown (Ram et al., 2025; Ryu et al., 2003).

On the other hand, products with an optimal value of α closer to 1 in the high-volatility, nonlinear-demand pattern range (such as N05B, R03, and N05C) will show this. For N05B in particular, the lowest error is achieved with α of 0.71. Of the three, R03 and N05C are more likely to exhibit full CNN–LSTM dominance of the ensemble with α = 1.00. This outcome validates that when the demand dynamic is highly nonlinear or stationary, the contribution of the deep learning models becomes more pronounced (Chen, 2025; Sooriarachchi et al., 2026).

In total, the variation in α across products indicates that the ensemble weighting is not random but rather data-informed and product-specific. This modification improves forecasting accuracy and adaptability by enhancing data interpretability and deep learning flexibility. The results indicate that no ensemble model is universally superior to the others. Instead, adaptive ensemble techniques provide an optimal and pragmatic approach to heterogeneous multi-product forecasting scenarios (Aswini et al., 2023; Hammam et al., 2025).

While Table 4, which only identifies the lowest-error model per product, Table 5 provides additional insight into the degree of contribution between statistical and deep learning components. The presence of intermediate α values (e.g., 0.71 for N05B and 0.11 for R06) indicates that optimal forecasting performance is often achieved through partial model integration rather than full dominance of a single model. This suggests that the adaptive ensemble captures complementary predictive strengths that cannot be inferred solely from individual model rankings.

E. Discussion and Practical Implications

The comparative evaluation results suggest that no single forecasting model consistently outperforms the others across all products. Instead, model success is predominantly dictated by the nature of the underlying demand. This necessitates context-sensitive, flexible forecasting for heterogeneous multi-product retailing.

Such classical statistical models as the Triple Exponential Smoothing (TES) and Seasonal ARIMA (SARIMA) are still powerful and still competing for those products that are characterized by stable, regular, and repeating seasonal demand patterns (Ram et al., 2025; Ryu et al., 2003). The main advantages of these models are computational efficiency, interpretability, and reduced data requirements. As such, they are most appropriate in operational contexts that demand model transparency and simplicity to support rapid decision-making (Kumari & Toshniwal, 2021; Xue et al., 2019).

By contrast, LSTM and CNN–LSTM deep learning models yield more favorable results for products characterized by volatile, nonlinear, and non-stationary demand patterns. These models' predictions of product demand are more accurate because they can capture long-term temporal dependencies and complex patterns. This is even though the lower prediction capacity comes at the cost of less complex models that are easier to implement and have lower computational costs (Chen, 2025; Sooriarachchi et al., 2026).

In this regard, it is interesting that ensemble approaches combine statistical and deep learning methods and still provide more holistic predictions. Other techniques can yield high predictions but will be less versatile because prediction variance is reduced. In high-risk forecasting, this section explains how some projections can lead to operational and predictable destruction. This brings emphasis on the flexibility and versatility that adaptive weighting ensembles provide, like the (SARIMA-CNN-LSTM), in high-risk forecasting scenarios.

In this Fragment, the essay outlines the perspective that series forecasting is a multi-step process that can be obtained without knowing the future value. Overall, the proposed forecasting framework is well designed for practical implementation and can be readily used as an adaptive, data-driven decision-support module for multi-product retail forecasting systems.

CONCLUSION

This study shows that multi-product sales forecasting requires model selection strategies that account for differences in demand characteristics rather than relying on available generic approaches. Findings show that deep learning models (LSTM and CNN–LSTM) and ensemble approaches surpass classical statistical models at the aggregate level and even more for products with high and irregular demand. Nevertheless, statistical models such as the TES and SARIMA still perform well for products with stable, regular demand and seasonalities. The adaptive ensemble of SARIMA and CNN LSTM improves the null performance of the system by having the model contribution dynamically modified by having an optimized weighting parameter (α), and the differences in the optimal α values of the different products indicate that the ensemble weighting needs to be data-driven as opposed to being static. The findings make clear that a hybrid, along with an adaptive forecasting model system, is best suited to multi-product retail forecasting under heterogeneous demand conditions. Despite the promising findings, several directions remain open for future research. First, the proposed framework can be extended to incorporate recent attention-based architectures, such as Temporal Fusion Transformer, to examine whether global dependency modeling further improves multi-product forecasting under heterogeneous demand conditions. Second, future studies may integrate multivariate inputs, including promotional campaigns, pricing information, and external economic indicators, to evaluate the impact of exogenous variables on forecasting robustness. Third, formal statistical significance testing across products may provide deeper insight into the reliability of performance differences. Finally, scalability analysis in larger retail ecosystems with hundreds of products and longer forecasting horizons would strengthen the practical applicability of the proposed framework.

REFERENCES

Amalou, I., Mouhni, N., & Abdali, A. (2024). CNN-LSTM architectures for non-stationary time series: decomposition approach. *2024 International Conference on Global Aeronautical Engineering and Satellite Technology (GAST)*, 1–5. https://doi.org/10.1109/GAST60528.2024.10520774

Arratia, A., El Daou, M., Kagerhuber, J., & Smolyarova, Y. (2025). Examining Challenges in Implied Volatility Forecasting: A Critical Review of Data Leakage and Feature Engineering combined with High-Complexity Models. *Computational Economics*. https://doi.org/10.1007/s10614-025-11172-z

Aswini, K., Reddy, U., Nagpal, A., Rana, A., Praveen, & Abood, B. S. Z. (2023). Ensemble Learning Approaches for Big Data Classification Tasks. *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, *10*, 1545–1550. https://doi.org/10.1109/UPCON59197.2023.10434616

Boukrouh, I., Idiri, S., Tayalati, F., Azmani, A., & Bouhsaien, L. (2025). A Hybrid Approach for Sales Forecasting: Combining Deep Learning and Time Series Analysis. *International Journal of Engineering, Transactions A: Basics*, *38*(4), 859–870. https://doi.org/10.5829/ije.2025.38.04a.16

Cerqueira, V., Roque, L., & Soares, C. (2025a). Forecasting with Deep Learning: Beyond Average of Average of Average Performance. In D. Pedreschi, A. Monreale, R. Guidotti, R. Pellungrini, & F. Naretto (Eds.), *Discovery Science* (pp. 135–149). Springer Nature Switzerland.

Cerqueira, V., Roque, L., & Soares, C. (2025b). Forecasting with Deep Learning: Beyond Average of Average of Average Performance. In D. Pedreschi, A. Monreale, R. Guidotti, R. Pellungrini, & F. Naretto (Eds.), *Discovery Science* (pp. 135–149). Springer Nature Switzerland.

Chen, J. (2025). Research on the Construction of Short-Term Demand Forecasting Model of Express Logistics Based on Deep Learning. *2025 5th International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA)*, 406–409. https://doi.org/10.1109/CAIBDA65784.2025.11183456

de Castro Moraes, T., Yuan, X. M., & Chew, E. P. (2024a). Hybrid convolutional long short-term memory models for sales forecasting in retail. *Journal of Forecasting*, *43*(5), 1278–1293. https://doi.org/10.1002/for.3073

de Castro Moraes, T., Yuan, X. M., & Chew, E. P. (2024b). Hybrid convolutional long short-term memory models for sales forecasting in retail. *Journal of Forecasting*, *43*(5), 1278–1293. https://doi.org/10.1002/for.3073

Eglite, L., & Birzniece, I. (2022). Retail Sales Forecasting Using Deep Learning: Systematic Literature Review. *Complex Systems Informatics and Modeling Quarterly*, *2022*(30), 53–62. https://doi.org/10.7250/csimq.2022-30.03

Hammam, I. M., El-Kharbotly, A. K., & Sadek, Y. M. (2025). Adaptive demand forecasting framework with weighted ensemble of regression and machine learning models along life cycle variability. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-23352-w

Hasan, P., Khan, T. D., Abedin, M., & Haque, M. E. (2025). Temporal trends and forecasting of respiratory mortality in Bangladesh: A SARIMA model for seasonal mortality risk and public health action. *Journal of Public Health Research*, *14*(4). https://doi.org/10.1177/22799036251395248

Hewage, H. C., Perera, H. N., & Bandara, K. (2025). Enhancing Demand Forecasting in Retail: A Comprehensive Analysis of Sales Promotional Effects on the Entire Demand Life Cycle. *Journal of Forecasting*. https://doi.org/10.1002/for.70039

Huang, S., & Zhou, L. (2025). A Residual-Corrected Hybrid ARIMA–CNN–LSTM Framework for High-Accuracy Tobacco Sales Forecasting in Regulated Markets. *International Journal of Computational Intelligence Systems*, *18*(1). https://doi.org/10.1007/s44196-025-00930-4

Jouilil, Y., & Iaousse, M. (2023). Comparing the Accuracy of Classical, Machine Learning, and Deep Learning Methods in Time Series Forecasting: A Case Study of USA Inflation. *Statistics, Optimization and Information Computing*, *11*(4), 1041–1050. https://doi.org/10.19139/soic-2310-5070-1767

Kristianto, R. P., & Setyanto, A. (2018). Golden Section Search-Multi Variable Algorithm for Optimization Parameter of Triple Exponential Smoothing Algorithm to Predict Sufferers of Lungs Disease. *2018 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE)*, 194–198. https://doi.org/10.1109/ICITISEE.2018.8720967

Kumari, P., & Toshniwal, D. (2021). Long short term memory–convolutional neural network based deep hybrid approach for solar irradiance forecasting. *Applied Energy*, *295*, 117061. https://doi.org/10.1016/J.APENERGY.2021.117061

Lawal, Y. B., Owolawi, P. A., Tu, C., Van Wyk, E., & Ojo, J. S. (2024). Stochastic Dynamical Modeling for Time-series Zero-Degree Isotherm Heights Prediction over Tropical Climate. *IOP Conference Series: Earth and Environmental Science*, *1428*(1). https://doi.org/10.1088/1755-1315/1428/1/012017

Nakthanom, S. (2025). Using Long Short-Term Memory to Losses Prediction for Three-Phase Distribution Transformer. *2025 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC)*, 1–8. https://doi.org/10.1109/ITC-CSCC66376.2025.11137634

Osman, E. G. A., & Otaibi, F. A. (2025). Integrating deep learning and econometrics for stock price prediction: A comprehensive comparison of LSTM, transformers, and traditional time series models. *Machine Learning with Applications*, *22*, 100730. https://doi.org/10.1016/j.mlwa.2025.100730

Pongdatu, G. A. N., & Putra, Y. H. (2018). Seasonal Time Series Forecasting using SARIMA and Holt Winter's Exponential Smoothing. *IOP Conference Series: Materials Science and Engineering*, *407*(1). https://doi.org/10.1088/1757-899X/407/1/012153

Popovska, E., & Georgieva-Tsaneva, G. (2024). Comparative Analysis of ARIMA and LSTM Models for Seasonal Times-Series

Forecasting. *2024 15th National Conference with International Participation (ELECTRONICA)*, 1–3. https://doi.org/10.1109/ELECTRONICA63645.2024.11146090

Ram, C. S., Raj, M., & Chaturvedi, R. K. (2025). Boosting Time-Series Forecasting Accuracy with SARIMAX Seasonal Interval Automation. *Procedia Computer Science*, *260*, 814–821. https://doi.org/10.1016/j.procs.2025.03.262

Rygh, T., Vaage, C., Westgaard, S., & Lange, P. E. de. (2025). Inflation Forecasting: LSTM Networks vs. Traditional Models for Accurate Predictions. *Journal of Risk and Financial Management*, *18*(7). https://doi.org/10.3390/jrfm18070365

Ryu, K., Jang, S. S., & Sanchez, A. (2003). Forecasting Methods and Seasonal Adjustment for a University Foodservice Operation. *Journal of Foodservice Business Research*, *6*(2), 17–34. https://doi.org/10.1300/J369v06n02_03

Selmy, H. A., Mohamed, H. K., & Medhat, W. (2024). A predictive analytics framework for sensor data using time series and deep learning techniques. *Neural Computing and Applications*, *36*(11), 6119–6132. https://doi.org/10.1007/s00521-023-09398-9

Shan, K., Wang, Y., Tang, Z., Chen, Y., & Li, Y. (2021). MixTConv: Mixed Temporal Convolutional Kernels for Efficient Action Recognition. *2020 25th International Conference on Pattern Recognition (ICPR)*, 1751–1756. https://doi.org/10.1109/ICPR48806.2021.9412586

Sooriarachchi, S. R., Hettiarachchi, G. P., & Dharmarathne, G. (2026). Deep Learning Models for Multimodal Demand Forecasting in Retail: A Comparative Case Study of Long Short-Term Memory and Temporal Fusion Transformers. In C. Anutariya, M. Bonsangue, A. Pinidiyaarachchi, & H. Usoof (Eds.), *Data Science and Artificial Intelligence* (pp. 195–206). Springer Nature Singapore.

Strøm, E., & Gundersen, O. E. (2024). Performance metrics for multi-step forecasting measuring win-loss, seasonal variance and forecast stability: an empirical study. *Applied Intelligence*, *54*(21), 10490–10515. https://doi.org/10.1007/s10489-024-05715-4

Taylor, J. W., & McSharry, P. E. (2017). Univariate Methods for Short-Term Load Forecasting. In *Advances in Electric Power and Energy Systems* (pp. 17–40). Wiley. https://doi.org/10.1002/9781119260295.ch2

Wang, S., Lin, Y., Jia, Y., Sun, J., & Yang, Z. (2024). Unveiling the Multi-Dimensional Spatio-Temporal Fusion Transformer (MDSTFT): A Revolutionary Deep Learning Framework for Enhanced Multi-Variate Time Series Forecasting. *IEEE Access*, *12*, 115895–115904. https://doi.org/10.1109/ACCESS.2024.3444788

Xue, N., Triguero, I., Figueredo, G. P., & Landa-Silva, D. (2019). Evolving Deep CNN-LSTMs for Inventory Time Series Prediction. *2019 IEEE Congress on Evolutionary Computation (CEC)*, 1517–1524. https://doi.org/10.1109/CEC.2019.8789957

Zhou, S., Guo, S., Du, B., Huang, S., & Guo, J. (2022). A Hybrid Framework for Multivariate Time Series Forecasting of Daily Urban Water Demand Using Attention-Based Convolutional Neural Network and Long Short-Term Memory Network. *Sustainability (Switzerland)*, *14*(17). https://doi.org/10.3390/su141711086