STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, August 2026, pp 1255–1283.
Published online in International Academic Press (www.IAPress.org)

Reassessing Statistical and Hybrid Forecasting Models for Flexible Packaging Demand under Post-COVID Structural Changes

Bagas Anindito Satyabhakti [1*], Basuki Widodo [2], Erwin Widodo [3]

[1]*School of Interdisciplinary Management and Technology, Sepuluh Nopember Institute of Technology, Indonesia*
[2]*Department of Mathematics, Sepuluh Nopember Institute of Technology, Indonesia*
[3]*Department of Industrial and Systems Engineering, Sepuluh Nopember Institute of Technology, Indonesia*

Abstract Accurate demand forecasting is essential for production planning, inventory control, and supply chain management in the flexible packaging industry, particularly under structural changes following the COVID-19 pandemic. This study evaluates statistical, machine learning, and hybrid forecasting models for monthly demand prediction in the Indonesian flexible packaging market using data from January 2012 to December 2024. The evaluated models include ARIMA, SARIMA, SARIMAX with selected exogenous variables and a COVID-19 intervention term, standalone Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM), and hybrid SARIMAX–MLP and SARIMAX–SVM models. To ensure a fair comparison, all models are assessed using RMSE, MAE, MAPE, and R^2 under a unified rolling-origin framework over the January 2023–December 2024 out-of-sample period, resulting in 24 one-step-ahead forecasts for each model. The empirical results show that ARIMA(1,1,3) achieves the lowest RMSE and highest R^2, indicating better performance in reducing larger forecast deviations, while SARIMA$(0,1,1)(0,1,0)_{12}$ achieves the lowest MAE and MAPE, indicating superior average absolute and percentage forecasting accuracy. In contrast, SARIMAX, standalone machine learning models, and hybrid models do not provide consistent improvements over simpler statistical benchmarks. The inclusion of exogenous variables and residual-based machine learning components does not improve forecasting performance, suggesting that the remaining external and nonlinear error structures are not sufficiently stable or learnable within the available monthly sample. The Harvey–Leybourne–Newbold corrected Diebold–Mariano test further indicates that the observed differences in forecast accuracy are not statistically significant at conventional levels. These findings show that increased model complexity does not necessarily lead to better forecasting performance. For practical industrial forecasting, simpler and more interpretable statistical models may be preferable when data are limited and short-term one-step-ahead forecasting accuracy is the main objective.

Keywords Flexible Packaging Demand; Demand Forecasting; ARIMA; SARIMA; SARIMAX; Hybrid Forecasting; Machine Learning; Post-COVID Structural Change; Indonesia.

DOI: 10.19139/soic-2310-5070-4077

1. Introduction

Accurate demand forecasting is essential for effective production planning, inventory control, capacity planning, and supply chain management in the flexible packaging industry. Inaccurate forecasts may lead to excess inventory, production inefficiency, capacity imbalance, or poor coordination across the supply chain [1, 2, 3]. These challenges are particularly important in emerging markets such as Indonesia, where industrial demand is influenced by economic growth, demographic change, sectoral dynamics, and shifts in consumer behavior [4]. In addition, structural disruptions such as the COVID-19 pandemic have increased uncertainty in demand patterns and reduced the reliability of forecasting models that rely solely on stable historical relationships [5, 6]. Therefore,

*Correspondence to: Bagas Anindito Satyabhakti (Email: 7032201018@student.its.ac.id). School of Interdisciplinary Management and Technology, Sepuluh Nopember Institute of Technology, Kampus Cokroaminoto, Surabaya City, East Java Province, Indonesia (60264).

ISSN 2310-5070 (online) ISSN 2311-004X (print)

forecasting models used in industrial decision-making should not only be accurate, but also robust, interpretable, and empirically validated under disrupted market conditions [1, 6].

1.1. Background and Motivation

Time-series forecasting methods have been widely used to model industrial demand because they can capture temporal dependence and provide interpretable forecasting structures. Classical statistical models, such as the Autoregressive Integrated Moving Average (ARIMA) model and its seasonal extension, Seasonal ARIMA (SARIMA), are commonly applied as benchmark forecasting methods due to their relatively simple structure, low data requirements, and ability to represent historical demand patterns [9, 1]. ARIMA is useful for modeling non-seasonal temporal dynamics, while SARIMA extends this framework by incorporating seasonal components, which may be relevant for monthly industrial demand data [9, 10].

To incorporate external demand drivers, the Seasonal ARIMA with exogenous variables (SARIMAX) model extends the statistical forecasting framework by including macroeconomic, sectoral, and intervention-related variables [10, 11]. In principle, this allows demand forecasts to reflect not only past demand behavior but also external conditions that may influence consumption, production, and distribution activities [4]. However, the inclusion of exogenous variables does not automatically improve forecasting performance. Their contribution depends on whether the selected variables provide stable and meaningful predictive information beyond the historical demand series [1, 10].

In parallel, machine learning methods, including Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM), have been increasingly applied to forecasting problems because of their ability to model nonlinear relationships [12, 14, 16]. Hybrid forecasting models have also gained attention by combining statistical models with machine learning techniques. In many hybrid frameworks, the statistical component is used to capture linear, seasonal, and exogenous structures, while the machine learning component is used to learn remaining nonlinear patterns from the residuals [20, 21, 25, 26].

Despite these potential advantages, the effectiveness of machine learning and hybrid models is conditional rather than universal. Their performance depends on data availability, signal-to-noise ratio, the stability of relationships between variables, and the existence of meaningful nonlinear residual structures [12, 17, 18]. In practical industrial datasets, which are often limited in size and affected by structural shifts, additional model complexity may increase variance, reduce interpretability, or introduce overfitting without improving out-of-sample forecasting accuracy [6, 14]. Therefore, the value of using more complex models should be assessed through a fair and consistent empirical evaluation rather than assumed in advance [17, 19].

1.2. Related Approaches and Research Gap

Existing forecasting studies often emphasize increasingly sophisticated methods, including machine learning and hybrid forecasting models, with the expectation that higher model complexity will lead to better predictive performance [20, 21, 26]. However, this assumption may not always hold, especially when the underlying demand series is dominated by relatively stable temporal patterns or when the available sample size is limited. In such cases, simpler statistical models may remain competitive or even perform better than more complex alternatives [6, 17].

Several gaps remain in the forecasting literature. First, the incremental value of exogenous variables is often insufficiently evaluated. Although macroeconomic and sectoral indicators may be theoretically relevant, their practical forecasting contribution depends on data quality, timing, stability, and their actual relationship with demand [4, 10]. Second, hybrid models are frequently proposed based on the assumption that residuals from statistical models contain nonlinear structures that can be exploited by machine learning methods [20, 21]. However, this assumption is not always tested explicitly. If the residuals are dominated by noise rather than systematic nonlinear patterns, the machine learning component may add complexity without improving accuracy [12, 18].

Third, many studies compare forecasting models using different training periods, feature structures, or evaluation procedures, making it difficult to determine whether performance differences are caused by model capability or by differences in experimental design. A unified rolling-origin, one-step-ahead evaluation framework is therefore

necessary to reflect real forecasting conditions and to ensure that all models are assessed using identical forecast origins [17, 18, 19]. Fourth, in the context of the flexible packaging industry, empirical evidence remains limited, particularly in emerging markets and under post-COVID structural changes [4, 7, 8]. This creates a need for forecasting research that evaluates whether more complex models truly provide additional forecasting value for practical industrial demand planning.

Based on these gaps, this study does not assume that hybrid or machine learning models are automatically superior. Instead, it investigates whether additional forecasting complexity, including exogenous variables, standalone machine learning, and residual-based hybrid modeling, provides meaningful improvements over simpler statistical benchmarks under a fair and consistent evaluation framework.

1.3. Research Objective and Contributions

This study aims to evaluate the forecasting performance of statistical, machine learning, and hybrid models for monthly demand prediction in the Indonesian flexible packaging market over the period January 2012 to December 2024. The models considered include ARIMA, SARIMA, SARIMAX with selected exogenous variables and a COVID-19 intervention term, standalone MLP and SVM models, and hybrid SARIMAX–MLP and SARIMAX–SVM models. All models are evaluated using a unified rolling-origin, one-step-ahead forecasting framework to ensure a fair comparison across approaches.

Model performance is assessed using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). These metrics are used to capture different dimensions of forecasting performance. RMSE emphasizes larger forecast deviations, MAE measures average absolute error, MAPE provides percentage-based accuracy, and R^2 evaluates explanatory performance relative to observed demand variation [1]. This multi-metric evaluation is important because the best-performing model may depend on the forecasting objective.

The main contributions of this study are as follows. First, the study provides a systematic comparison of statistical, machine learning, and hybrid forecasting models under identical rolling-origin evaluation conditions. Second, it evaluates whether additional model complexity, including exogenous variables, standalone machine learning structures, and hybrid residual-learning components, improves out-of-sample forecasting performance. Third, it provides metric-dependent forecasting insight by showing that model preference may differ depending on whether the objective is to minimize larger errors, average absolute errors, or percentage errors. Fourth, the study offers practical guidance for industrial forecasting by demonstrating that simpler and more interpretable statistical models may remain highly competitive when the dataset is limited and when residual nonlinear patterns are weak.

Overall, this study contributes to the forecasting literature by providing empirical evidence that increased model complexity does not necessarily lead to better forecasting performance. Instead, model selection should be guided by data characteristics, forecasting objectives, empirical validation, interpretability, and practical relevance.

1.4. Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews the literature on statistical forecasting models, machine learning approaches, hybrid forecasting models, and the positioning of this study. Section 3 describes the methodology, including data preprocessing, model specifications, feature construction, hybrid forecasting design, and statistical testing. Section 4 presents the experimental design and rolling-origin evaluation procedure. Section 5 discusses the empirical results, forecast comparison, statistical significance testing, and managerial implications. Section 6 concludes the study and provides directions for future research.

2. Related Work

Demand forecasting has been widely examined across manufacturing, supply chain, energy, finance, retail, and other industrial contexts. Forecasting methods are commonly grouped into statistical time-series models, machine learning approaches, and hybrid forecasting frameworks [1, 17]. Each model class offers different advantages depending on the structure of the data, the availability of explanatory variables, the forecasting horizon, and the

evaluation criteria used. In industrial forecasting applications, model performance is not determined only by model sophistication, but also by the extent to which the model structure matches the underlying demand pattern [6]. Therefore, a balanced review of statistical, machine learning, and hybrid forecasting approaches is necessary to position the present study and to justify the need for fair model comparison.

2.1. Statistical Forecasting Models

Classical statistical models have long been used as fundamental tools in time-series forecasting because of their interpretability, relatively low data requirements, and strong theoretical foundation [9, 1]. The Autoregressive Integrated Moving Average (ARIMA) model is one of the most widely used approaches for modeling linear temporal dependence in historical data. In demand forecasting, ARIMA is often used as a benchmark model because it provides a parsimonious structure for capturing autoregressive and moving-average patterns after appropriate differencing is applied to achieve stationarity [9].

For data observed at regular seasonal intervals, Seasonal ARIMA (SARIMA) extends the ARIMA framework by incorporating seasonal autoregressive, differencing, and moving-average components [9, 1]. This extension is particularly relevant for monthly demand data, where recurring seasonal patterns may affect production planning, inventory control, and distribution decisions. By explicitly representing seasonal dependence, SARIMA can provide robust short-term forecasts when the dominant demand structure is driven by historical temporal and seasonal patterns [10].

To incorporate external influences, the Seasonal ARIMA with exogenous variables (SARIMAX) model allows macroeconomic, sectoral, and intervention-related variables to be included in the forecasting structure [10, 11]. This extension enables the model to account not only for internal temporal dynamics, but also for external factors that may affect demand, such as inflation, commodity prices, sectoral output, consumer behavior, or structural disruptions. Intervention variables can also be included to represent external shocks, such as economic crises, policy changes, or pandemic-related structural breaks [5, 6].

However, the inclusion of exogenous variables does not automatically improve forecasting accuracy. The practical value of SARIMAX depends on whether the selected external variables contain stable and predictive information beyond the historical demand series [1, 10]. If the relationship between exogenous variables and demand is weak, unstable, delayed, or noisy, the additional variables may increase model complexity without improving out-of-sample forecasting performance. Therefore, SARIMAX should be evaluated empirically against simpler statistical benchmarks rather than assumed to provide better forecasts.

Although statistical models may be limited in capturing complex nonlinear relationships, their transparency, robustness, and relatively low data requirements make them highly relevant for practical industrial forecasting [1, 6]. In settings where the available sample size is limited or where the demand structure is dominated by historical temporal patterns, parsimonious statistical models may remain competitive with more complex alternatives [17].

2.2. Machine Learning Approaches

Machine learning approaches have gained increasing attention in forecasting studies because they can model nonlinear relationships and complex interactions among variables [12, 14]. Unlike classical statistical models, machine learning models do not require a strict parametric specification of the relationship between predictors and the target variable. This flexibility makes them attractive for demand forecasting problems where demand may be affected by nonlinear responses to historical values, macroeconomic indicators, sectoral changes, or other external drivers [4].

Multi-Layer Perceptron (MLP) neural networks are commonly used to approximate nonlinear functions through interconnected layers of neurons [12, 14]. In forecasting applications, MLP models can be trained using lagged demand variables and relevant explanatory variables to capture nonlinear associations between past information and future demand. Their ability to approximate complex relationships makes them potentially useful when the data contain strong nonlinear signals.

Support Vector Machine (SVM), particularly in the form of support vector regression, has also been widely applied to forecasting problems [16, 4]. Through kernel functions such as the radial basis function (RBF), SVM can map input variables into higher-dimensional feature spaces and model nonlinear relationships without requiring

a fully specified parametric form. This makes SVM suitable for some forecasting contexts, especially when the objective is to balance model flexibility and generalization.

Despite these advantages, machine learning models also have several limitations in practical time-series forecasting. They typically require careful feature engineering, appropriate lag construction, data scaling, hyperparameter tuning, and sufficient training observations to achieve stable out-of-sample performance [12, 14]. When the dataset is small, noisy, or affected by structural changes, machine learning models may overfit historical patterns and fail to generalize to new observations [17, 18]. In addition, their lower interpretability compared with statistical models may limit their usefulness in managerial decision-making contexts, where forecast transparency and ease of implementation are important [6].

Therefore, the use of machine learning models in demand forecasting should be justified through rigorous out-of-sample evaluation rather than in-sample fitting performance. Their potential advantage depends on whether the available data contain nonlinear structures that can be learned reliably and whether the model is evaluated under the same information set and forecast origins as competing methods [17, 19].

2.3. Hybrid Forecasting Models

Hybrid forecasting models have been developed to combine the strengths of statistical and machine learning approaches [20, 21]. A common hybrid strategy is to decompose the forecasting problem into linear and nonlinear components. In this framework, a statistical model such as ARIMA, SARIMA, or SARIMAX is first used to model linear temporal dependence, seasonality, and possible exogenous effects. The residuals from the statistical model are then modeled using a machine learning method to capture any remaining nonlinear patterns [25, 26].

The theoretical appeal of hybrid forecasting is that each component can focus on the type of structure it is best suited to capture. The statistical component provides an interpretable representation of temporal and seasonal patterns, while the machine learning component offers flexibility in modeling nonlinear residual behavior [20, 21]. When the underlying data contain both systematic linear structures and learnable nonlinear residual patterns, hybrid models may improve forecasting accuracy compared with standalone statistical or machine learning models [25, 26].

However, the effectiveness of hybrid forecasting is conditional rather than universal. Residual-based hybrid models rely on the assumption that the residuals from the statistical model still contain meaningful nonlinear information [12, 18]. If the residuals are approximately random or dominated by noise, the machine learning component may not add useful predictive information. Instead, it may increase model variance, introduce overfitting, and reduce out-of-sample forecasting accuracy [17, 6]. Therefore, hybrid models should not be considered automatically superior to simpler statistical approaches.

This issue is particularly important in industrial demand forecasting, where datasets are often limited in length, aggregated at monthly frequency, and influenced by market uncertainty, measurement noise, and structural disruptions [4, 7]. Under such conditions, the additional flexibility of hybrid models may not translate into better out-of-sample performance. A fair evaluation should therefore compare hybrid models not only against standalone machine learning alternatives, but also against parsimonious statistical benchmarks such as ARIMA and SARIMA [17, 19].

2.4. Synthesis and Positioning of This Study

The reviewed literature shows that statistical, machine learning, and hybrid forecasting models each offer potential benefits for demand forecasting. Statistical models provide interpretability and robustness, machine learning models provide flexibility in capturing nonlinear relationships, and hybrid models attempt to combine both advantages through residual learning [1, 12, 20]. However, the forecasting performance of these approaches is highly dependent on data characteristics, sample size, feature quality, residual structure, and the evaluation framework used [6, 17].

Several important gaps can be identified. First, the incremental value of exogenous variables is not always rigorously assessed, even though external indicators may not necessarily provide additional predictive information beyond the historical demand series [10, 11]. Second, machine learning models are often introduced because of their nonlinear modeling capability, but their performance may be limited when the dataset is small or when the

nonlinear signal is weak [12, 14]. Third, hybrid models are frequently proposed based on the assumption that residuals from statistical models contain exploitable nonlinear patterns, although this assumption may not hold in every empirical setting [21, 25]. Fourth, model comparisons are sometimes conducted using different data splits, feature sets, or evaluation procedures, which can lead to biased conclusions about model superiority [17, 19].

In the context of the Indonesian flexible packaging industry, empirical evidence remains limited, particularly under post-COVID structural changes [4, 7]. Monthly demand in this industry may be influenced by historical demand behavior, seasonal effects, macroeconomic conditions, sectoral dynamics, and changes in consumption and distribution patterns. However, it remains an empirical question whether more complex models, such as SARIMAX with exogenous variables, standalone machine learning models, or hybrid residual-learning models, provide meaningful forecasting improvement over simpler statistical benchmarks.

To address this gap, the present study evaluates statistical, machine learning, and hybrid forecasting models under a unified rolling-origin, one-step-ahead evaluation framework. Rather than assuming the superiority of a particular model class, this study examines whether additional model complexity provides measurable improvement in out-of-sample forecasting accuracy. This positioning allows the study to contribute not only by comparing forecasting models, but also by providing practical evidence on when simpler and more interpretable statistical models may remain preferable for industrial demand forecasting.

3. Methodology

This study adopts a comparative forecasting framework to evaluate statistical, machine learning, and hybrid models for monthly demand prediction in the Indonesian flexible packaging industry. The methodology is designed to ensure transparency, reproducibility, and comparability across model classes. Rather than assuming that more complex models necessarily provide better forecasts, this study evaluates whether the inclusion of exogenous variables, machine learning structures, and hybrid residual-learning components provides measurable forecasting improvement over simpler statistical benchmarks [6, 17, 18].

The methodological procedure consists of five main stages. First, the demand series and candidate exogenous variables are prepared and aligned at a monthly frequency. Second, selected exogenous variables and a COVID-19 intervention indicator are specified for the SARIMAX and machine learning models. Third, statistical forecasting models, including ARIMA, SARIMA, and SARIMAX, are developed. Fourth, standalone machine learning models, namely MLP and SVM, are constructed using lagged demand values and the same selected external information. Fifth, hybrid SARIMAX–MLP and SARIMAX–SVM models are implemented to examine whether the residuals from the SARIMAX model contain nonlinear structures that can improve forecasting accuracy [1, 12, 20].

3.1. Data Description and Preprocessing

The dataset used in this study consists of monthly observations from January 2012 to December 2024. The dependent variable represents aggregated demand in the Indonesian flexible packaging industry. Several candidate exogenous variables are also considered to represent macroeconomic conditions, sectoral activity, demographic development, and changes in distribution or consumption patterns. These candidate variables include population, consumer confidence index, Consumer Price Index (CPI) inflation, Brent oil price, gross domestic product, food and beverage sector output, manufacturing PMI, urbanization rate, and e-commerce penetration. The inclusion of external demand drivers is consistent with the use of explanatory variables in demand forecasting and SARIMAX-based forecasting frameworks [4, 10, 11].

Because the variables originate from different reporting frequencies, temporal harmonization is performed to align all variables with the monthly demand series. Quarterly and annual variables are converted into monthly observations using interpolation or disaggregation procedures before model estimation. This step is necessary to ensure that each observation in the demand series has a corresponding value for the candidate explanatory variables. Similar preprocessing and temporal alignment procedures are commonly required when explanatory variables with different reporting frequencies are used in time-series forecasting applications [1, 10].

Data preprocessing is performed through several steps. First, all variables are converted into numeric format and aligned using a common monthly time index. Second, missing values and incomplete observations are checked and handled to preserve temporal consistency. Third, the data are reviewed for unit consistency and abnormal values before model estimation. Fourth, the demand series is examined for stationarity and seasonal behavior to determine the appropriate differencing structure for ARIMA, SARIMA, and SARIMAX models. Finally, for machine learning and hybrid models, input features are standardized using scaling parameters estimated only from the training data within each rolling iteration to prevent information leakage [1, 9, 12].

For machine learning models, lagged demand variables are constructed at lags 1, 2, 3, 6, and 12 months. These lags are selected to represent short-term demand persistence, medium-term adjustment, and annual seasonal memory in monthly demand data. Lag-based feature construction is commonly used to transform time-series observations into supervised learning inputs for machine learning forecasting models [12, 14]. For hybrid models, lagged residual variables are constructed using the same lag structure from the residual series generated by the SARIMAX model.

All variables are observed at a monthly frequency over the period from January 2012 to December 2024. Summary statistics are computed using the full sample prior to model estimation. The demand variable represents aggregated flexible packaging market demand, while the exogenous variables represent macroeconomic, sectoral, demographic, and distribution-related indicators aligned with the monthly demand series. The summary statistics are presented in Table 1. The table is included to provide transparency regarding the scale, unit, and distribution of each candidate variable before model estimation.

Table 1. Summary Statistics of Flexible Packaging Demand and Exogenous Variables

Variable	Unit	Mean	Std. Dev.	Maximum	Minimum
Flexible Packaging Demand	Million USD	3.82	0.64	5.25	2.77
Population	Million People	267.00	10.27	283.49	250.22
Consumer Confidence Index	Index (>100 = optimistic)	115.28	11.70	128.94	77.31
Inflation (CPI)	% (Monthly)	0.31	0.45	3.29	-0.45
Brent Oil Price	USD per Barrel	75.47	24.59	125.45	18.38
Gross Domestic Product	Quadrillion IDR	2.54	0.38	3.30	1.86
Food & Beverages Output	Trillion IDR	505.39	73.26	637.26	378.04
Manufacturing PMI	Index (50 = neutral)	50.47	2.27	57.20	40.10
Urbanization Rate	% of Population	55.27	2.51	59.20	51.00
E-Commerce Penetration	% of Total Retail	14.45	11.31	31.21	3.26

Table 1 reports the candidate variables considered in the empirical analysis. However, not all candidate variables are included in the final SARIMAX model. The final exogenous subset is selected based on economic relevance, model parsimony, information criteria, and out-of-sample forecasting performance. Therefore, Table 1 serves as a descriptive summary of the full candidate variable set, while the final model specification and selected exogenous variables are reported separately in the SARIMAX model specification table.

3.2. Exogenous Variable and Intervention Specification

The exogenous variables are included to examine whether external economic and sectoral indicators provide additional predictive information beyond the historical demand series. The full set of candidate variables is first considered based on economic relevance, data availability, and consistency with the monthly forecasting framework. However, not all candidate variables are included in the final SARIMAX model. The final exogenous subset is selected based on empirical model performance and model parsimony. This approach is consistent with the principle that exogenous variables should be included only when they provide meaningful predictive information and improve model usefulness beyond the historical target series [1, 10].

Note: The initial training period corresponds to the data available before the first rolling-origin forecast. The evaluation period corresponds to the 24 one-step-ahead forecasts from January 2023 to December 2024.

Table 1A. Summary Statistics by Initial Training and Evaluation Period

Variable	Period	Mean	Std. Dev.	Maximum	Minimum
Packaging Demand	Jan 2012–Dec 2022	3.63	0.51	4.73	2.76
Packaging Demand	Jan 2023–Dec 2024	4.84	0.15	5.25	4.57
Inflation (CPI)	Jan 2012–Dec 2022	0.34	0.48	3.29	-0.45
Inflation (CPI)	Jan 2023–Dec 2024	0.17	0.19	0.52	-0.18
Brent Oil Price	Jan 2012–Dec 2022	74.37	26.59	125.45	18.38
Brent Oil Price	Jan 2023–Dec 2024	81.50	5.45	93.72	73.86
Food & Beverages Output	Jan 2012–Dec 2022	486.88	64.11	582.01	378.04
Food & Beverages Output	Jan 2023–Dec 2024	607.16	14.69	627.27	583.56
Urbanization Rate	Jan 2012–Dec 2022	54.61	2.16	57.93	51.00
Urbanization Rate	Jan 2023–Dec 2024	58.89	0.32	59.20	58.57
E-Commerce Penetration	Jan 2012–Dec 2022	11.64	10.03	31.21	3.26
E-Commerce Penetration	Jan 2023–Dec 2024	29.91	0.62	30.52	29.30

To further examine possible distributional differences between the initial training period and the final evaluation period, additional descriptive statistics are reported in Table 1A. The initial training period covers January 2012 to December 2022, while the evaluation period covers January 2023 to December 2024. This separation provides additional transparency regarding the characteristics of the data used for model estimation and the post-COVID out-of-sample period used for forecasting evaluation. The comparison is also useful for identifying whether the demand series and selected external indicators exhibit changes in level, variability, or range between the model-development period and the final forecasting period.

The final SARIMAX specification includes CPI inflation, Brent oil price, food and beverage sector output, urbanization rate, e-commerce penetration, and a COVID-19 intervention variable. These variables are selected because they represent price dynamics, production cost or commodity-related pressure, sectoral demand drivers, demographic structure, digital distribution development, and structural disruption associated with the pandemic period. The use of intervention variables is consistent with time-series approaches that incorporate structural changes or external shocks into the model specification [5, 6].

Candidate exogenous variables were screened before final model estimation to ensure that the SARIMAX specification was not determined arbitrarily. The candidate variables include demographic, macroeconomic, sectoral, manufacturing, urbanization, and digital-commerce indicators that may affect flexible packaging demand. The screening process considered four criteria: theoretical relevance to flexible packaging demand, correlation with the dependent variable, multicollinearity as indicated by the variance inflation factor, and rolling-origin sensitivity testing through add-one and drop-one variable evaluation. Because several macroeconomic variables are highly trended and strongly correlated with demand, the final selection was not based solely on one statistical criterion. Instead, variables were retained when they provided theoretical relevance, interpretability, or structural control value within the SARIMAX framework.

Note: The selection process combines theoretical relevance, correlation with demand, multicollinearity screening, rolling-origin sensitivity testing, and interpretability. Because several macroeconomic and structural variables are highly trended, variables with high VIF values are interpreted cautiously and are treated as structural controls rather than as independently causal predictors.

As shown in Table 2, several candidate variables have strong correlations with flexible packaging demand. However, high correlation alone does not necessarily imply useful incremental forecasting information, especially when variables are highly trended or collinear. Indonesia population and GDP, for example, show strong correlations with demand but also exhibit high VIF values and limited incremental forecasting contribution. Therefore, they were excluded to reduce redundancy and avoid excessive multicollinearity. Manufacturing PMI was also excluded because it has weak correlation with demand and does not improve rolling-origin forecasting accuracy.

Table 2. Exogenous Variable Selection and Rationale

Candidate Variable	Decision	Correlation with Demand	VIF	Screening Interpretation	Rationale
Indonesia Population	Excluded	0.978	431.449	Adding the variable does not improve RMSE and the VIF is very high.	Excluded because it is highly trend-driven and overlaps with long-term demand growth and urbanization effects.
Consumer Confidence Index	Excluded	0.199	2.506	Adding the variable slightly improves RMSE, but the improvement is marginal.	Excluded for parsimony because the short-term sentiment signal provides limited additional contribution relative to the selected variables.
Consumer Price Index Inflation	Included	-0.176	1.109	Removing the variable has only a small effect on forecast accuracy.	Included as a low-collinearity price-level control that captures inflationary pressure on demand and production costs.
Oil Price	Included	-0.147	3.366	Removing the variable improves rolling accuracy, but the variable remains theoretically relevant.	Included to represent energy and petroleum-based input cost pressure relevant to flexible packaging materials.
GDP	Excluded	0.970	66.840	Adding the variable worsens RMSE and the VIF is high.	Excluded because it overlaps with broader macroeconomic trend information already represented by other structural variables.
F&B Output	Included	0.970	138.423	Removing the variable improves rolling accuracy, but the variable has strong sectoral relevance.	Included because food and beverage output is directly linked to downstream flexible packaging demand.
Manufacturing PMI	Excluded	0.035	1.321	Adding the variable does not improve RMSE.	Excluded because the monthly manufacturing activity signal does not provide sufficient incremental forecasting contribution.
Urbanization Rate	Included	0.974	565.652	Removing the variable has only a small effect, and the VIF is very high.	Included as a structural demand-control variable related to urban consumption and packaged-goods usage, but interpreted cautiously.
E-commerce Penetration	Included	0.895	18.539	Removing the variable improves rolling accuracy, but the variable represents an important post-COVID demand channel.	Included to capture digital-commerce-related packaging demand growth, particularly in the post-COVID period.

The final SARIMAX specification retains Consumer Price Index Inflation, Oil Price, Food and Beverage Output, Urbanization Rate, and E-commerce Penetration. These variables are retained because they represent price pressure, input-cost effects, downstream sectoral demand, structural urban consumption, and digital-commerce-related packaging demand. Nevertheless, several retained variables also show high VIF values or limited incremental improvement in the rolling-origin sensitivity test. Therefore, they are interpreted as structural control variables rather than as independently causal predictors. This cautious interpretation is consistent with the overall empirical finding that adding exogenous variables does not necessarily improve forecasting performance over simpler ARIMA and SARIMA benchmarks.

To ensure that the COVID-19 intervention specification was not selected arbitrarily, four alternative intervention structures were evaluated within the SARIMAX framework: no intervention, pulse intervention, step intervention, and ramp intervention. The no-intervention specification excludes any COVID-19 structural-control variable. The pulse specification represents a temporary shock only at the onset of the pandemic in March 2020. The step specification represents a persistent structural shift from March 2020 onward, while the ramp specification represents a gradual post-COVID adjustment process. Each intervention specification was evaluated using the same SARIMAX order, the same selected exogenous variables, and the same rolling-origin one-step-ahead forecasting procedure. The use of alternative intervention structures helps examine whether pandemic-related changes are better represented as a temporary shock, a persistent level shift, or a gradual adjustment process [5, 6].

The comparison results are presented in Table 3. The ramp specification achieves the lowest RMSE, the highest R^2, and the lowest AIC and BIC values, indicating that a gradual post-COVID adjustment structure provides the best fit and reduces larger forecast deviations. However, the step specification achieves lower MAE and MAPE than the ramp specification, indicating better average absolute and percentage forecasting accuracy. The pulse specification performs weakest, suggesting that the pandemic effect cannot be adequately represented as a one-month temporary shock. The no-intervention model also performs slightly worse than the step and ramp alternatives in several criteria, supporting the inclusion of a COVID-related structural-control variable.

Although the ramp specification provides the strongest performance in terms of RMSE and information criteria, the step intervention is retained in the main SARIMAX specification because it provides a simple and interpretable representation of the post-COVID demand environment and produces better MAE and MAPE than the ramp alternative. Therefore, the COVID-19 intervention variable is interpreted as a structural-control variable rather than as a complete representation of all pandemic-related dynamics. This distinction is important because the pandemic involved multiple phases, including initial disruption, adaptation, and recovery, which cannot be fully captured by a single binary intervention term. **Note:** All intervention specifications are evaluated using the same SARIMAX order, selected exogenous variables, and 24 rolling-origin one-step-ahead forecasts. The step intervention is retained in the main specification for interpretability and average-error performance, while the ramp intervention is reported as a robustness alternative.

3.3. Statistical Forecasting Models

This study evaluates three statistical forecasting models: ARIMA, SARIMA, and SARIMAX. These models are included to represent different levels of statistical complexity, ranging from a univariate non-seasonal model to a seasonal model with exogenous variables and intervention effects. ARIMA and SARIMA are widely used statistical benchmarks in time-series forecasting because they provide interpretable structures for modeling temporal dependence and seasonality [1, 9].

3.3.1. ARIMA Model

The Autoregressive Integrated Moving Average (ARIMA) model is used as the first statistical benchmark. ARIMA captures linear temporal dependence in the demand series through autoregressive and moving-average components, while differencing is applied to handle non-stationarity [1, 9]. The general ARIMA model can be expressed as:

$$\phi(B)(1-B)^d y_t = \theta(B)\epsilon_t$$

where y_t denotes demand at time t, B is the backshift operator, d is the differencing order, $\phi(B)$ is the autoregressive polynomial, $\theta(B)$ is the moving-average polynomial, and ϵ_t is the error term.

Table 3. Comparison of COVID-19 Intervention Specifications in the SARIMAX Model

Intervention Specification	Description	RMSE	MAE	MAPE (%)	R^2	AIC	BIC	Interpretation
No intervention	No COVID-19 intervention variable	0.159	0.109	2.258	-0.115	-206.178	-185.537	Baseline without COVID-19 control
Pulse intervention	Temporary shock only in March 2020	0.159	0.110	2.265	-0.122	-204.220	-180.630	Weakest specification
Step intervention	Structural shift from March 2020 onward	0.158	0.108	2.233	-0.112	-204.827	-181.237	Lowest MAE and MAPE
Ramp intervention	Gradual post-COVID adjustment from March 2020 onward	0.154	0.111	2.296	-0.057	-217.882	-194.292	Lowest RMSE, highest R^2, lowest AIC/BIC

The ARIMA order (p, d, q) is selected through model comparison using candidate specifications. The final ARIMA model used in the empirical evaluation is ARIMA(1,1,3).

3.3.2. SARIMA Model

The Seasonal ARIMA (SARIMA) model extends ARIMA by incorporating seasonal autoregressive, differencing, and moving-average components [1, 9]. This model is included because the dataset is observed monthly and may contain seasonal demand patterns. The general SARIMA specification is written as:

$$\text{SARIMA}(p, d, q)(P, D, Q)_s$$

where (p, d, q) represents the non-seasonal component, (P, D, Q) represents the seasonal component, and s denotes the seasonal period. Since the data are observed monthly, the seasonal period is set to $s = 12$.

The SARIMA model orders are selected through grid search. Based on the final model selection process, the SARIMA specification used in the empirical evaluation is SARIMA(0,1,1)(0,1,0)$_{12}$.

3.3.3. SARIMAX Model with Exogenous Variables and Intervention

The SARIMAX model extends the SARIMA framework by incorporating selected exogenous variables and the COVID-19 intervention variable [10, 11]. This allows the model to account for internal temporal dependence, seasonal structure, selected external drivers, and post-COVID structural change. The general SARIMAX formulation is expressed as:

$$y_t = \text{SARIMA}(p, d, q)(P, D, Q)_s + \beta X_t + \gamma I_t + \epsilon_t$$

where X_t denotes the vector of exogenous variables, I_t represents the COVID-19 intervention variable, β denotes the coefficients of the exogenous variables, γ denotes the coefficient of the intervention term, and ϵ_t is the error term.

The final selected SARIMAX model is SARIMAX(0,1,1)(0,1,0)$_{12}$, incorporating CPI inflation, Brent oil price, food and beverage sector output, urbanization rate, e-commerce penetration, and the COVID-19 step intervention variable. The model is estimated using maximum likelihood estimation. Information criteria, including AIC and

BIC, are used together with out-of-sample forecasting performance to assess model suitability [1, 9]. The final SARIMAX model produces an AIC of -204.827 and a BIC of -181.236.

Table 4. Final SARIMAX Model Specification and Estimation Results

Variable	Coefficient	Std. Error	z-Statistic	p-value
Inflation (CPI)	0.0375	0.0220	1.683	0.092
Brent Oil Price	0.0213	0.0202	1.054	0.292
Food & Beverages Output	0.3823	0.2219	1.723	0.085
Urbanization Rate	-0.1019	1.9517	-0.052	0.958
E-commerce Penetration	0.0210	0.1447	0.145	0.885
COVID Step	-0.0291	0.0623	-0.467	0.641
MA(1)	-0.7472	0.0456	-16.382	0.000

Model: SARIMAX$(0,1,1)(0,1,0)_{12}$
AIC: -204.827 **BIC:** -181.236 **Log-Likelihood:** 110.413

Note: The final SARIMAX model includes selected exogenous variables and a COVID-19 step intervention variable. Not all candidate variables reported in Table 4 are included in the final specification, because variable selection is conducted through a subset evaluation process based on economic relevance, model parsimony, information criteria, and out-of-sample forecasting performance. The selected SARIMAX specification is also used as the statistical baseline for the SARIMAX–MLP and SARIMAX–SVM hybrid models.

The results in Table 4 indicate that the MA(1) term is statistically significant, confirming the relevance of the moving-average component in the final SARIMAX specification. Among the exogenous variables, CPI inflation and food and beverage sector output are marginally significant at the 10% level, while Brent oil price, urbanization rate, e-commerce penetration, and the COVID-19 step intervention variable are not statistically significant at conventional levels. Therefore, the estimated effects of the exogenous variables should be interpreted cautiously. In this study, the COVID-19 step variable is retained primarily as a structural-control variable to represent the post-COVID demand environment, rather than as a statistically strong standalone predictor. The limited statistical significance of several exogenous variables also supports the need to evaluate whether the SARIMAX model provides additional forecasting value beyond simpler ARIMA and SARIMA benchmarks.

The residuals from the final SARIMAX model are used for hybrid residual learning. These residuals represent the portion of demand variation that remains unexplained by the statistical component. If the residuals contain systematic and learnable nonlinear patterns, the machine learning component may improve forecast accuracy through residual correction. However, if the residuals are dominated by noise or unstable error structures, the hybrid model may add complexity without providing additional forecasting benefit [20, 21, 18].

3.4. Machine Learning Models

Standalone machine learning models are implemented to evaluate whether nonlinear forecasting structures can improve demand prediction relative to statistical models. Two machine learning models are considered: Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM) regression. These models are included because machine learning methods can approximate nonlinear relationships and interactions that may be difficult to capture using purely linear statistical models [12, 14].

To ensure a fair comparison, both standalone machine learning models use the same information set. The input features consist of lagged demand values at lags 1, 2, 3, 6, and 12 months, together with CPI inflation, Brent oil price, food and beverage sector output, urbanization rate, e-commerce penetration, and the COVID-19 step intervention variable. This ensures that the standalone machine learning models are not disadvantaged relative to SARIMAX and the hybrid models in terms of available explanatory information.

All machine learning input variables are standardized before model estimation. Within each rolling-origin iteration, scaling parameters are estimated only from the training data and then applied to the corresponding test

observation. This procedure prevents information leakage from the test period into the training process and supports a fair out-of-sample evaluation [12, 19].

Table 5. Machine Learning Model Configuration

Model	Component	Specification
MLP	Input features	Demand lags 1, 2, 3, 6, and 12; selected exogenous variables; COVID-19 step intervention
	Hidden layers	2
	Neurons	64 and 32
	Activation	ReLU
	Optimizer	Adam
	Regularization	$\alpha = 0.001$
	Maximum iterations	5,000
	Early stopping	Applied
SVM	Input features	Same as MLP
	Kernel	Radial basis function (RBF)
	Regularization parameter	$C = 10$
	Gamma	scale
	Epsilon	$\epsilon = 0.01$

The machine learning configurations used for the standalone and hybrid models are summarized in Table 5. The same MLP and SVM configurations are applied consistently across the standalone machine learning models and the residual-learning components of the hybrid SARIMAX–MLP and SARIMAX–SVM models.

3.4.1. Standalone MLP

The standalone MLP model is implemented as a feedforward neural network. It is designed to capture nonlinear relationships between lagged demand, selected exogenous variables, the COVID-19 intervention variable, and future demand. MLP models approximate nonlinear functions through interconnected layers of neurons and are commonly used in forecasting applications when nonlinear relationships may be present [12, 14]. The model consists of two hidden layers with 64 and 32 neurons, respectively. The rectified linear unit (ReLU) activation function is used in the hidden layers, and the Adam optimizer is used for training.

A regularization parameter of $\alpha = 0.001$ is applied to reduce the risk of overfitting. The maximum number of iterations is set to 5,000, and early stopping is applied to improve generalization, particularly because the dataset is relatively limited in size. The use of early stopping helps prevent the model from excessively fitting noise in the training data.

3.4.2. Standalone SVM

The standalone SVM model is implemented using support vector regression with a radial basis function kernel. Support vector regression is widely used for nonlinear forecasting because kernel functions allow the model to represent nonlinear relationships in a transformed feature space [4, 16]. The RBF kernel is used to allow nonlinear mapping between the input features and demand. The regularization parameter is set to $C = 10$, the kernel coefficient is set to $\gamma =$ scale, and the epsilon-insensitive loss parameter is set to $\epsilon = 0.01$.

The SVM model uses the same input structure as the standalone MLP model. This ensures that differences in forecasting performance between MLP and SVM are attributable to model structure rather than differences in available predictors.

3.5. Hybrid Forecasting Models

Hybrid forecasting models are implemented to evaluate whether residual-based nonlinear learning improves forecast accuracy beyond the SARIMAX baseline. The hybrid framework follows a sequential structure. In the first stage, SARIMAX is used to model linear temporal dependence, seasonality, selected exogenous effects, and the COVID-19 intervention term. In the second stage, a machine learning model is trained to predict the residual

component generated by the SARIMAX model. This residual-learning approach is commonly used in hybrid forecasting frameworks that combine statistical and machine learning models [20, 21, 25].

The hybrid models are included not because they are assumed to be superior, but because they provide a direct test of whether the SARIMAX residuals contain exploitable nonlinear patterns. If such patterns exist, the machine learning residual forecast may improve the final forecast. If the residuals do not contain meaningful nonlinear structure, the hybrid model may add complexity without improving performance [6, 18, 17].

The hybrid forecasting models are constructed using a residual-learning strategy. First, the SARIMAX model is used to generate the baseline forecast. The forecast error from the SARIMAX model is then used as the learning target for the machine learning component. The MLP and SVM models are trained to predict the remaining error component using lagged residuals, selected exogenous variables, and the COVID-19 intervention variable. The final hybrid forecast is obtained by adding the machine-learning-predicted residual to the SARIMAX forecast. The structure of the hybrid models is summarized in Table 6. Two hybrid configurations are evaluated: SARIMAX–MLP and SARIMAX–SVM. Table 6 shows that the hybrid models differ only in the residual-learning algorithm. The SARIMAX–MLP model uses a multilayer perceptron to approximate the residual component, while the SARIMAX–SVM model uses support vector regression with an RBF kernel. The same residual lag structure and exogenous information are used in both hybrid models to ensure a fair comparison.

Table 6. Hybrid Model Structure

Hybrid Model	Statistical Component	Machine Learning Component	Learning Target	Input Features	Final Forecast
SARIMAX–MLP	SARIMAX$(0,1,1)(0,1,0)_{12}$ with selected exogenous variables and COVID step	MLP	SARIMAX forecast error / residual component	Residual lags 1, 2, 3, 6, 12; selected exogenous variables; COVID step	SARIMAX forecast + MLP-predicted residual
SARIMAX–SVM	SARIMAX$(0,1,1)(0,1,0)_{12}$ with selected exogenous variables and COVID step	SVM with RBF kernel	SARIMAX forecast error / residual component	Residual lags 1, 2, 3, 6, 12; selected exogenous variables; COVID step	SARIMAX forecast + SVM-predicted residual

3.5.1. SARIMAX–MLP Model

In the SARIMAX–MLP model, the residuals generated by the SARIMAX model are used as the target variable for the MLP residual-learning component. The input features consist of lagged SARIMAX residuals at lags 1, 2, 3, 6, and 12 months, together with the same selected exogenous variables used in the SARIMAX model: CPI inflation, Brent oil price, food and beverage sector output, urbanization rate, e-commerce penetration, and the COVID-19 step intervention variable.

The MLP residual model uses the same architecture as the standalone MLP model, consisting of two hidden layers with 64 and 32 neurons, ReLU activation, Adam optimization, $\alpha = 0.001$, a maximum of 5,000 iterations, and early stopping. All residual-learning inputs are standardized using training-data scaling parameters within each rolling-origin iteration.

3.5.2. SARIMAX–SVM Model

In the SARIMAX–SVM model, support vector regression is used to predict the residual component from the SARIMAX model. The input structure mirrors the SARIMAX–MLP residual-learning model. Specifically, the features include lagged SARIMAX residuals at lags 1, 2, 3, 6, and 12 months, together with CPI inflation, Brent oil price, food and beverage sector output, urbanization rate, e-commerce penetration, and the COVID-19 step intervention variable.

The SVM residual model uses an RBF kernel with $C = 10$, $\gamma = \text{scale}$, and $\epsilon = 0.01$. This configuration is applied to provide nonlinear residual mapping while maintaining a controlled level of model flexibility.

3.5.3. Final Forecast Combination

For both hybrid models, the final forecast is obtained by adding the SARIMAX forecast and the machine learning residual forecast:

$$\hat{y}_t = \hat{y}_t^{\text{SARIMAX}} + \hat{e}_t^{\text{ML}}$$

where $\hat{y}_t$ denotes the final hybrid forecast, $\hat{y}_t^{SARIMAX}$ denotes the forecast generated by the SARIMAX model, and $\hat{e}_t^{ML}$ denotes the predicted residual generated by either the MLP or SVM model.

This formulation allows the study to evaluate whether residual correction provides additional forecasting value beyond the SARIMAX model. The final comparison across ARIMA, SARIMA, SARIMAX, standalone machine learning models, and hybrid models is then conducted under the unified rolling-origin evaluation framework described in Section 4.

4. Experimental Design

This section describes the experimental design used to evaluate the forecasting performance of the statistical, machine learning, and hybrid models. The design is constructed to ensure a fair, reproducible, and consistent comparison across all evaluated forecasting approaches. All models are assessed under the same rolling-origin, one-step-ahead forecasting framework, using identical forecast origins and performance metrics [1, 18, 19]. This design allows the study to evaluate whether additional model complexity, including exogenous variables, standalone machine learning structures, and hybrid residual-learning components, provides meaningful improvement over simpler statistical benchmarks.

The overall experimental workflow is illustrated in Figure 1. The framework summarizes the complete forecasting procedure, beginning with data preparation and variable harmonization, followed by statistical, machine learning, and hybrid model development. The framework then proceeds to rolling-origin one-step-ahead forecasting, performance evaluation, statistical testing, and analytical interpretation. This structure is used to ensure that all models are evaluated under consistent conditions and that the comparison focuses on forecasting capability rather than differences in data handling or evaluation procedure. As shown in Figure 1, the experimental design begins with the preparation and harmonization of monthly demand and exogenous variables. The prepared dataset is then used to develop statistical, machine learning, and hybrid forecasting models under clearly defined input structures. All models are evaluated using an expanding-window rolling-origin procedure with a one-step-ahead forecasting horizon. Forecasting performance is assessed using multiple accuracy metrics, and statistical significance is examined using the Harvey–Leybourne–Newbold corrected Diebold–Mariano test [22, 29]. The final stage interprets the results in terms of model complexity, predictive accuracy, and practical relevance for industrial demand forecasting.

4.1. Evaluation Framework and Forecasting Horizon

To ensure a fair comparison across statistical, machine learning, and hybrid forecasting models, all models are evaluated using the same expanding-window rolling-origin one-step-ahead forecasting framework. The full dataset consists of 156 monthly observations from January 2012 to December 2024. The final out-of-sample evaluation period covers January 2023 to December 2024, corresponding to 24 monthly forecast origins.

At the first forecast origin, the 132 observations from January 2012 to December 2022 are used as the training sample, and each model generates a one-step-ahead forecast for January 2023. The training window is then expanded by one month after the actual observation becomes available, and the model is re-estimated to generate the forecast for February 2023. This expanding-window procedure is repeated sequentially until the final one-step-ahead forecast for December 2024 is obtained. Therefore, each model produces exactly 24 out-of-sample forecasts under identical evaluation conditions. Expanding-window evaluation is appropriate for forecasting applications in

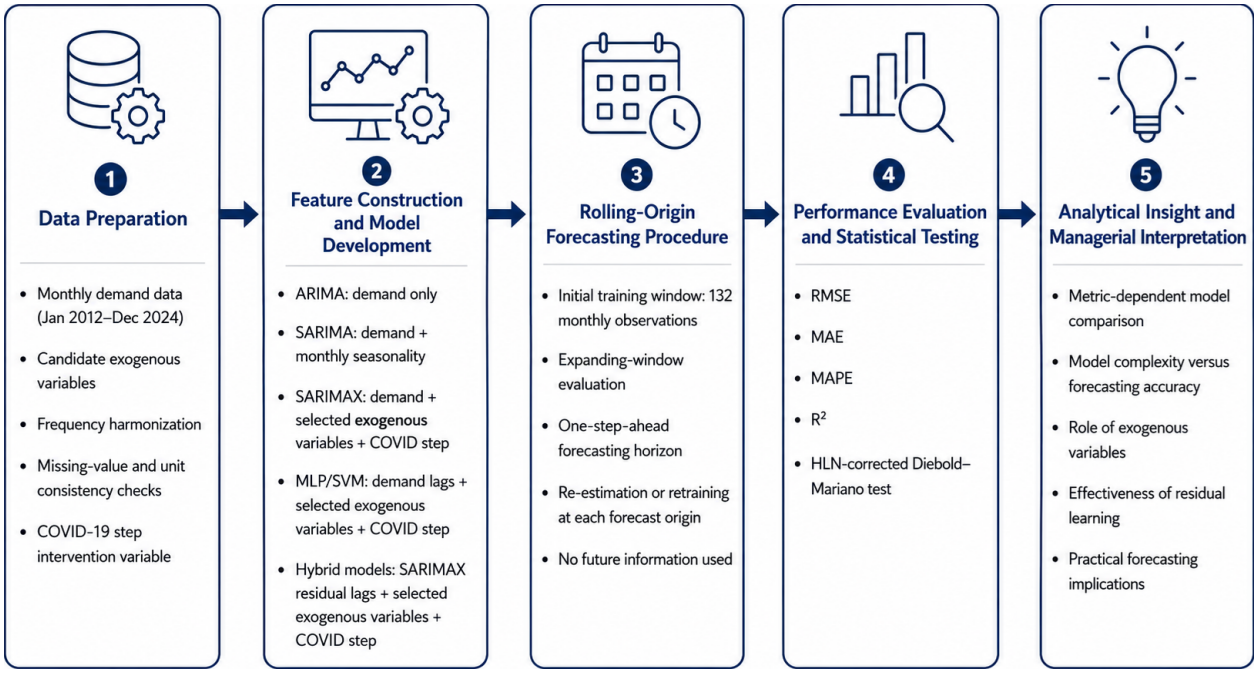

Figure 1. Research framework illustrating data preparation, model development, rolling-origin forecasting procedure, performance evaluation, statistical testing, and analytical insight stages.

which the available information increases over time and models are updated as new observations become available [1, 18].

This evaluation design avoids direct comparison between models estimated over different test periods or forecast horizons. It also reduces the risk of look-ahead bias because each forecast is generated using only information available up to the corresponding forecast origin. For machine learning and hybrid models, all scaling parameters are estimated only from the training data within each rolling iteration and are then applied to the corresponding test observation. This procedure prevents information leakage from the evaluation period into the training process, which is essential for valid out-of-sample comparison [12, 19]. Hyperparameters are determined before the final rolling evaluation and kept fixed across all rolling iterations to ensure consistency and prevent tuning to individual test observations.

4.2. Model Training and Forecasting Procedure

For each rolling-origin iteration, the statistical models are estimated using the available training observations. The ARIMA and SARIMA models are trained using the historical demand series only. The SARIMAX model is trained using the historical demand series, selected exogenous variables, and the COVID-19 step intervention variable. This design allows the statistical models to be compared according to increasing levels of model complexity, from univariate non-seasonal forecasting to seasonal and exogenous-variable forecasting [1, 9, 10].

For standalone machine learning models, the target variable is demand. The input features consist of lagged demand values at lags 1, 2, 3, 6, and 12 months, together with the selected exogenous variables and the COVID-19 step intervention variable. The selected exogenous variables are CPI inflation, Brent oil price, food and beverage sector output, urbanization rate, and e-commerce penetration. Feature standardization is performed within each rolling iteration using scaling parameters estimated only from the training data. The same scaling parameters are then applied to the corresponding test observation. Lagged feature construction and proper scaling are important steps in applying machine learning models to time-series forecasting problems [12, 14].

For hybrid models, the SARIMAX model is first used to generate the statistical forecast and residual series. The machine learning component is then trained to predict the SARIMAX residuals. The residual-learning input features consist of lagged SARIMAX residuals at lags 1, 2, 3, 6, and 12 months, together with the same selected exogenous variables and the COVID-19 step intervention variable. The final hybrid forecast is obtained by adding the SARIMAX forecast and the predicted residual correction. This residual-learning structure follows the general hybrid forecasting logic in which statistical models capture linear and seasonal components, while machine learning models are used to examine whether remaining residual patterns contain learnable nonlinear information [20, 21, 25].

4.3. Benchmarking Strategy and Model Comparability

The benchmarking strategy is designed to compare models across different levels of complexity. ARIMA represents a parsimonious non-seasonal statistical benchmark. SARIMA extends ARIMA by incorporating monthly seasonality. SARIMAX further adds selected exogenous variables and the COVID-19 intervention term. MLP and SVM represent standalone nonlinear machine learning models. SARIMAX–MLP and SARIMAX–SVM represent hybrid residual-learning models. Comparing models across these levels of complexity is useful for evaluating whether additional model flexibility translates into improved out-of-sample forecasting performance [6, 17].

To ensure comparability, all models are evaluated using the same dataset, the same rolling-origin forecast origins, the same one-step-ahead horizon, and the same evaluation metrics. The standalone machine learning models use the same selected exogenous variables and COVID-19 intervention indicator as the SARIMAX and hybrid models. Therefore, differences in forecasting performance can be attributed primarily to model structure rather than differences in data splitting, forecast horizon, feature availability, or evaluation procedure. This is important because inconsistent evaluation procedures can lead to biased conclusions about model superiority [18, 19].

4.4. Final Model Specification and Input Structure

The final model specifications and input structures used in the experimental evaluation are summarized in Table 7.

Table 7. Final Model Specification and Input Structure

Model	Input Structure	Lag Structure	Main Specification
ARIMA	Demand only	Selected by grid search	ARIMA(1,1,3)
SARIMA	Demand only	Monthly seasonality	SARIMA$(0,1,1)(0,1,0)_{12}$
SARIMAX	Demand + selected exogenous variables + COVID-19 intervention	Monthly seasonality, $s = 12$	SARIMAX$(0,1,1)(0,1,0)_{12}$; exogenous variables: CPI, Oil Price, FnB, Urbanization Rate, E-commerce Penetration, & COVID-19 step intervention
MLP	Demand lags + selected exogenous variables + COVID-19 step	1, 2, 3, 6, and 12	See Table 5
SVM	Demand lags + selected exogenous variables + COVID-19 steps	1, 2, 3, 6, and 12	See Table 5
SARIMAX–MLP	SARIMAX forecast + residual lags + selected exogenous variables + COVID-19 step	1, 2, 3, 6, and 12	Residual-learning hybrid
SARIMAX–SVM	SARIMAX forecast + residual lags + selected exogenous variables + COVID-19 step	1, 2, 3, 6, and 12	Residual-learning hybrid

Note: The selected exogenous variables are Consumer Price Index Inflation, Oil Price, Food and Beverage Output, Urbanization Rate, and E-commerce Penetration. The COVID step variable is included as a structural-control variable for the post-COVID period. Machine learning hyperparameters are reported separately in Table 5.

Table 7 shows that all evaluated models are implemented under clearly defined input structures and specifications. The table also clarifies that the standalone machine learning models are not disadvantaged in terms of explanatory information, because they receive the same selected exogenous variables and COVID-19 intervention indicator used in the SARIMAX model. For the hybrid models, the machine learning component is trained on lagged SARIMAX residuals rather than directly on demand, allowing the residual-learning stage to test whether additional nonlinear structure remains after the statistical component.

4.5. Performance Metrics

Forecasting performance is evaluated using four metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). These metrics are selected because they capture different aspects of forecasting performance. The use of multiple evaluation metrics is recommended in forecasting studies because model rankings may vary depending on the loss function or accuracy criterion used [1, 17].

RMSE measures the square root of the average squared forecast error and gives greater penalty to larger errors. This metric is important in demand planning because large forecast deviations may lead to production imbalance, excess inventory, stock shortages, or inefficient capacity allocation. MAE measures the average absolute forecast error and provides a more direct interpretation of average forecasting deviation. MAPE expresses forecasting error in percentage terms, making it useful for evaluating relative forecasting accuracy. R^2 is reported as a complementary measure to indicate how well each model explains variation in observed demand.

Using multiple metrics is important because model performance may be metric-dependent. A model with the lowest RMSE may be preferable when large forecast errors are costly, while a model with lower MAE or MAPE may be more suitable for routine planning accuracy. Therefore, no single metric is treated as the only criterion of model superiority. Instead, the results are interpreted by considering the forecasting objective and the managerial relevance of each metric [1, 6].

The metrics are calculated as follows:

$$\begin{aligned} RMSE &= \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t-\hat{y}_t)^2} \\ MAE &= \frac{1}{n}\sum_{t=1}^{n}|y_t-\hat{y}_t| \\ MAPE &= \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t-\hat{y}_t}{y_t}\right| \\ R^2 &= 1-\frac{\sum_{t=1}^{n}(y_t-\hat{y}_t)^2}{\sum_{t=1}^{n}(y_t-\bar{y})^2} \end{aligned} \tag{1}$$

where y_t denotes the observed demand, $\hat{y}_t$ denotes the forecasted demand, $\bar{y}$ denotes the mean of observed demand, and n denotes the number of forecast observations.

4.6. Statistical Significance Testing

To assess whether the observed differences in forecasting performance are statistically significant, this study applies the Harvey–Leybourne–Newbold corrected Diebold–Mariano test. The original Diebold–Mariano test is commonly used to compare the predictive accuracy of two competing forecasting models, while the Harvey–Leybourne–Newbold correction improves the finite-sample properties of the test statistic [22, 29]. The test is conducted under the null hypothesis that two competing models have equal predictive accuracy. This test is included to complement the numerical performance metrics and to avoid overinterpreting small differences in RMSE, MAE, MAPE, or R^2.

Because ARIMA achieves the lowest RMSE in the rolling-origin evaluation, it is used as the benchmark model in the pairwise statistical comparisons. Squared-error loss is used to align the statistical test with RMSE as the primary criterion for larger forecast deviations. The test compares ARIMA against SARIMA, SARIMAX, MLP, SVM, SARIMAX–MLP, and SARIMAX–SVM.

The Diebold–Mariano test statistic is interpreted together with the p-value. A statistically significant result would indicate that the difference in predictive accuracy between two models is unlikely to be due to sampling variation. However, if the p-value is greater than conventional significance thresholds, the null hypothesis of equal predictive accuracy cannot be rejected. Therefore, the statistical test results are interpreted together with the forecasting metrics, model simplicity, interpretability, and practical usefulness.

4.7. Implementation Details

All experiments are conducted using Python 3.13. The statistical models, including ARIMA, SARIMA, and SARIMAX, are implemented using the statsmodels library. The machine learning models, including MLP and SVM, are implemented using the scikit-learn library. Random seeds are fixed where applicable to improve reproducibility.

All models are evaluated under the same computational environment, rolling-origin procedure, and performance-metric calculation. Feature scaling for machine learning and hybrid models is performed separately within each rolling-origin iteration using training data only. This ensures that the evaluation procedure remains consistent across all model classes and prevents information leakage from the test observation into the model-training process.

Overall, this experimental design provides a rigorous basis for comparing statistical, machine learning, and hybrid forecasting models under realistic monthly demand forecasting conditions. It also allows the study to evaluate whether more complex models provide meaningful forecasting improvements over simpler and more interpretable statistical benchmarks.

5. Results and Discussion

This section presents the empirical results of the forecasting evaluation and discusses their implications for model selection in industrial demand forecasting. The results are interpreted based on the unified rolling-origin, one-step-ahead forecasting framework described in Section 4. The discussion focuses on three main issues: comparative forecasting accuracy, the incremental value of model complexity, and the practical implications of the findings for flexible packaging demand planning. Interpreting forecast results across these dimensions is important because forecasting performance depends not only on numerical accuracy, but also on model simplicity, stability, interpretability, and practical usefulness [1, 6, 18].

The forecasting performance results reported in this section are based on the 24 rolling-origin one-step-ahead forecasts generated for the period January 2023 to December 2024. Each model is evaluated over the same forecast origins, using the same actual demand observations and the same performance metrics. Therefore, the reported RMSE, MAE, MAPE, and R^2 values are directly comparable across the statistical, machine learning, and hybrid forecasting models. This consistent evaluation setting is necessary to avoid biased conclusions caused by differences in data splitting, forecast horizon, or available information across models [17, 19].

5.1. Forecasting Performance Comparison

The comparative forecasting performance of all evaluated models is summarized in Table 8. The results are based on the expanding-window rolling-origin one-step-ahead evaluation framework, which ensures that all models are assessed using identical forecast origins, the same evaluation period, and the same performance metrics. Specifically, the evaluation is based on 24 one-step-ahead forecasts from January 2023 to December 2024. Lower values of RMSE, MAE, and MAPE indicate better forecasting accuracy, while a higher R^2 indicates better explanatory performance relative to observed demand variation. The use of multiple metrics is important because different accuracy measures may favor different models depending on the forecasting objective and error-loss function [1, 17].

Table 8. Forecasting Performance Comparison

Model	RMSE	MAE	MAPE (%)	R^2
ARIMA	**0.135398**	0.090761	1.833554	**0.188023**
SARIMA	0.137616	**0.084562**	**1.749780**	0.161192
SARIMAX	0.155075	0.106025	2.196266	-0.065142
MLP	0.146609	0.107659	2.199537	0.047980
SVM	0.144296	0.110465	2.242840	0.077790
SARIMAX–MLP	0.152998	0.104977	2.167129	-0.036798
SARIMAX–SVM	0.159483	0.102160	2.117454	-0.126558

Note: RMSE, MAE, MAPE, and R² were calculated based on 24 rolling-origin one-step-ahead forecasts covering the January 2023-December 2024 out-of-sample evaluation period.

The results in Table 8 show that ARIMA(1,1,3) achieves the lowest RMSE of 0.135398 and the highest R^2 of 0.188023. This indicates that ARIMA performs best in reducing larger forecast deviations and explaining variation in observed demand under the revised evaluation framework. SARIMA achieves the lowest MAE of 0.084562 and the lowest MAPE of 1.749780%, indicating that SARIMA$(0,1,1)(0,1,0)_{12}$ provides the best average absolute and percentage forecasting accuracy. Therefore, the choice between ARIMA and SARIMA may depend on the forecasting objective: ARIMA is preferable when reducing large errors is the main concern, while SARIMA is preferable when average absolute or percentage accuracy is more important [1, 6].

The results also show that the more complex models do not consistently outperform the simpler statistical benchmarks. SARIMAX records higher RMSE, MAE, and MAPE than both ARIMA and SARIMA, despite incorporating selected exogenous variables and the COVID-19 step intervention variable. This suggests that, under the current data structure and one-step-ahead forecasting horizon, the additional explanatory variables do not provide sufficient incremental predictive information beyond the historical demand series. This finding is consistent with the view that exogenous variables improve forecasting performance only when their relationship with the target series is stable, timely, and predictive out of sample [1, 10].

The standalone machine learning models also do not provide a clear advantage over the statistical benchmarks. SVM achieves a lower RMSE than MLP, SARIMAX, and both hybrid models, but it still does not outperform ARIMA or SARIMA across the main evaluation metrics. The MLP model produces a moderate RMSE but records higher MAE and MAPE than the best statistical models. These findings suggest that nonlinear learning models do not automatically improve forecasting accuracy when the available sample is limited or when nonlinear feature-target relationships are not sufficiently stable [12, 14, 17].

The hybrid models also fail to provide consistent improvement. SARIMAX–MLP and SARIMAX–SVM both produce higher RMSE values than ARIMA, SARIMA, SVM, and MLP. Although the hybrid models produce slightly lower MAE or MAPE than some standalone alternatives, they do not outperform the best statistical benchmarks. This indicates that the SARIMAX residuals may not contain sufficiently strong or stable nonlinear patterns for the machine learning residual-learning components to exploit. In this setting, residual-based hybridization adds model complexity without producing a corresponding improvement in out-of-sample forecasting accuracy [20, 21, 18].

Overall, the results in Table 8 suggest that simpler statistical models remain competitive and numerically stronger than the more complex alternatives in this empirical setting. This does not imply that exogenous variables, machine learning models, or hybrid models are generally ineffective. Rather, it indicates that their incremental forecasting value is limited in this dataset, with this sample size, feature structure, and short-term one-step-ahead evaluation horizon. The findings therefore support a data-driven and objective-specific approach to model selection rather than assuming that greater model complexity necessarily leads to better forecasting performance [6, 17].

5.2. Interpretation of Model Performance

The statistical models provide the strongest overall performance in this evaluation. ARIMA and SARIMA outperform or remain competitive with the more complex models despite using only the historical demand series.

This suggests that the dominant predictive information in the dataset is contained in the temporal structure of demand itself. The relatively strong performance of ARIMA and SARIMA also indicates that parsimonious statistical models remain useful for monthly industrial demand forecasting when the available dataset is limited and when historical demand dynamics provide sufficient predictive signal.

The SARIMAX model does not outperform ARIMA or SARIMA despite incorporating selected exogenous variables and the COVID-19 step intervention variable. This result suggests that the additional external information included in the SARIMAX specification does not translate into better out-of-sample forecasting accuracy under the rolling-origin evaluation framework. This does not mean that the selected exogenous variables are theoretically irrelevant. Rather, it indicates that their incremental predictive contribution is limited in this specific dataset, model specification, and forecasting horizon.

Several factors may explain this result. First, some exogenous variables may influence flexible packaging demand with a delay that is not fully captured in the current specification. Second, the relationship between macroeconomic indicators and packaging demand may be unstable during and after the COVID-19 period. Third, interpolation or temporal harmonization of quarterly and annual variables into monthly values may reduce their short-term predictive usefulness. Fourth, the historical demand series may already capture much of the information reflected in the external indicators. These issues are particularly relevant in forecasting settings affected by structural change, where historical relationships between predictors and the target variable may become unstable [5, 6].

Among the standalone machine learning models, SVM achieves a lower RMSE than MLP, SARIMAX, and both hybrid models. However, SVM still does not outperform ARIMA or SARIMA across the main evaluation metrics. The MLP model produces a moderate RMSE but records higher MAE and MAPE than the best statistical models. These results indicate that standalone nonlinear models do not provide a clear advantage under the available sample size and feature structure.

The limited advantage of MLP and SVM may be related to the relatively small number of monthly observations, the limited length of the rolling training window, and the possibility that nonlinear patterns are not sufficiently strong or stable. Machine learning models generally require adequate training data and stable feature-target relationships to generalize well. When the signal-to-noise ratio is low or the dataset is limited, their flexibility may increase variance without improving out-of-sample forecasting accuracy [12, 14, 17].

5.3. Forecast Visualization and Error Behavior

Figure 2 compares the observed demand with forecasts generated by ARIMA, SARIMA, SARIMAX, and SVM during the rolling-origin evaluation period. The visual comparison shows that ARIMA and SARIMA follow the observed demand pattern relatively closely, supporting their stronger numerical performance in Table 8. SARIMAX and SVM also capture the general demand movement, but they do not consistently provide closer forecasts than the simpler statistical benchmarks. This finding reinforces the conclusion that additional model complexity does not necessarily improve short-term forecasting accuracy for the flexible packaging demand series.

Figure 2 illustrates the observed and forecasted demand values for selected representative models over the evaluation period. To avoid visual overcrowding, the figure focuses on ARIMA, SARIMA, SARIMAX, and SVM. ARIMA and SARIMA are shown because they provide the strongest overall statistical forecasting performance, SARIMAX is included as the exogenous-variable benchmark, and SVM is included as the strongest standalone machine learning model in terms of RMSE. The complete performance comparison across all statistical, machine learning, and hybrid models is reported in Table 8, while the absolute forecast error pattern for all models is shown in Figure 3.

Figure 2 presents selected representative models to avoid visual overcrowding in the observed-versus-forecasted plot. ARIMA and SARIMA are included because they provide the strongest overall statistical-model performance, SARIMAX is included as the exogenous-variable benchmark, and SVM is included as the strongest standalone machine learning model by RMSE. The full model comparison is reported quantitatively in Table 8 and through the absolute forecast error comparison in Figure 3.

The forecast comparison supports the numerical findings in Table 8. ARIMA and SARIMA track the observed demand relatively well and provide stable short-term forecasts. SARIMAX does not show a clear visual advantage

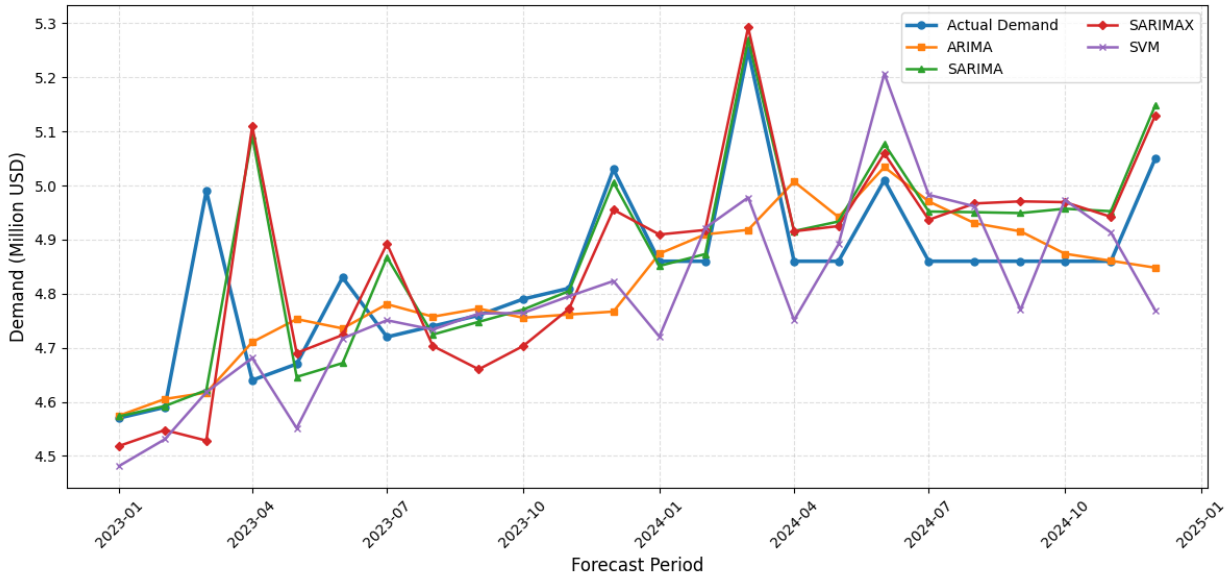

Figure 2. Observed versus forecasted demand comparison for selected representative models based on 24 rolling-origin one-step-ahead forecasts from January 2023 to December 2024.

despite incorporating selected exogenous variables and the COVID-19 intervention term. This reinforces the finding that external variables do not provide sufficient incremental predictive value in the current setting.

The visual comparison also indicates that forecast errors are not concentrated only in one model class. More complex models do not consistently reduce deviations from observed demand. This suggests that the main challenge is not merely the choice between statistical and machine learning models, but the extent to which the available data contain stable and learnable predictive patterns. Where the underlying demand movement is adequately captured by historical demand dynamics, simpler statistical models can remain competitive [6, 17].

Figure 3 presents the absolute forecast errors of all evaluated models across the 24 rolling-origin one-step-ahead forecasts. The figure complements the aggregate performance metrics in Table 8 by showing how forecast errors vary over time. This comparison is useful because a model with strong average performance may still produce larger deviations in specific months. Therefore, time-based error visualization provides additional information beyond aggregate accuracy metrics [1, 18].

Figure 3 shows that ARIMA and SARIMA generally produce lower and more stable forecast errors than the more complex SARIMAX, machine learning, and hybrid models. Although some complex models reduce errors in specific periods, they do not consistently dominate the simpler statistical benchmarks across the full evaluation period. This visual evidence supports the numerical results in Table 8, indicating that additional model complexity does not necessarily lead to improved short-term forecasting accuracy.

The forecast-error pattern provides additional evidence that the largest deviations are not systematically reduced by the more complex models. Although machine learning and hybrid models may adjust to some local fluctuations, they do not consistently reduce forecast errors across the rolling evaluation period. This supports the conclusion that additional model complexity does not necessarily improve short-term forecasting accuracy for this dataset.

5.4. Forecast Error Diagnostic Analysis

To further examine the behavior of the forecasting errors, a diagnostic analysis was conducted on the out-of-sample forecast errors generated by the final SARIMAX model. This analysis is relevant because the hybrid SARIMAX–MLP and SARIMAX–SVM models are designed to improve the SARIMAX forecasts by learning remaining patterns in the error component. Therefore, examining the SARIMAX forecast errors provides additional insight into whether the residual-learning stage has a meaningful structure to exploit. Diagnostic checking is an important

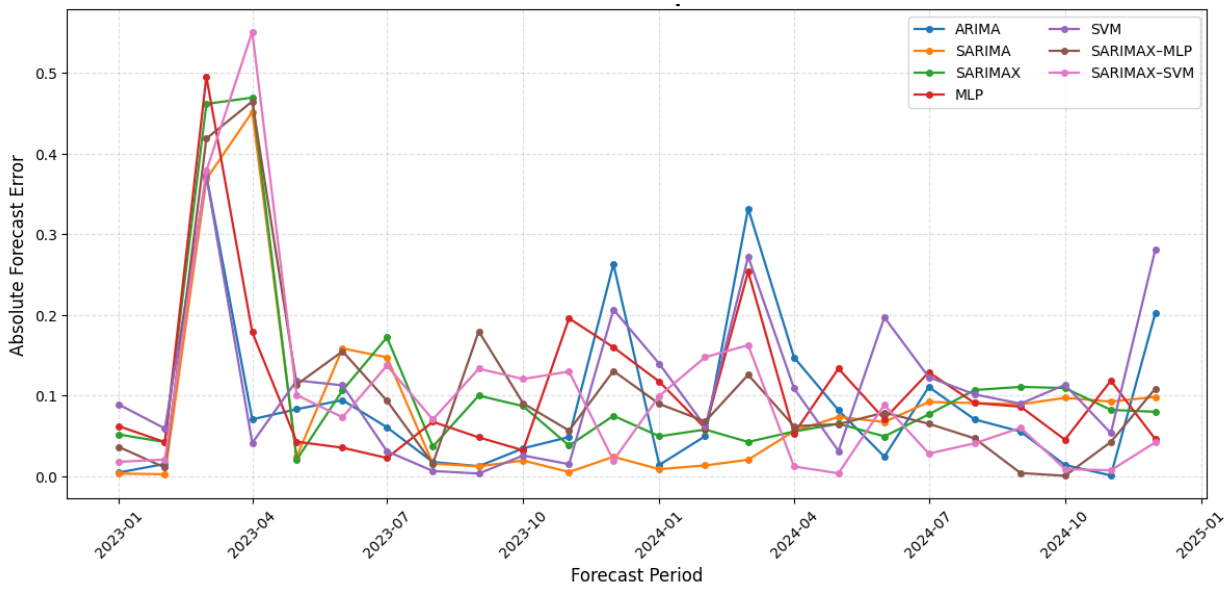

Figure 3. Absolute forecast error comparison across forecasting models based on 24 rolling-origin one-step-ahead forecasts from January 2023 to December 2024.

part of time-series model evaluation because residual or forecast-error behavior can indicate whether systematic patterns remain unexplained by the model [1, 9].

The forecast error summary shows that the final SARIMAX model produces a mean error of -0.022862, with a standard deviation of 0.156680. The minimum and maximum forecast errors are -0.469507 and 0.461645, respectively, while the mean absolute error is 0.106025. The negative mean error indicates a slight tendency toward overforecasting, although the average bias is relatively small compared with the range of forecast errors.

Table 9. Forecast Error Diagnostic Results for the Final SARIMAX Model

Diagnostic Test	Null Hypothesis	Statistic	p-value	Interpretation
Ljung–Box Q(6)	No forecast-error autocorrelation up to lag 6	3.977890	0.679669	Do not reject null
Ljung–Box Q(12)	No forecast-error autocorrelation up to lag 12	4.879098	0.961896	Do not reject null
Jarque–Bera	Forecast errors are normally distributed	18.499812	0.000096	Reject null
ARCH-LM(6)	No ARCH effect / no volatility clustering	16.133123	0.013057	Reject null
BDS test	Forecast errors are independently distributed / no nonlinear dependence	2.289833	0.022031	Reject null

Note: Diagnostic tests are based on the 24 out-of-sample forecast errors generated by the final SARIMAX model over January 2023–December 2024.

Because the diagnostic tests are conducted using only 24 out-of-sample forecast errors, the resulting p-values should be interpreted with caution. Therefore, the diagnostic results are used as indicative evidence to support the qualitative interpretation of the forecast-error structure, rather than as conclusive evidence of the underlying distributional properties.

The diagnostic test results are presented in Table 9. The Ljung–Box tests at lags 6 and 12 produce p-values of 0.679669 and 0.961896, respectively, indicating that the null hypothesis of no forecast-error autocorrelation is not rejected. This suggests that the SARIMAX forecast errors do not exhibit substantial remaining linear autocorrelation, implying that the model has captured most of the linear serial dependence in the demand series. However, the Jarque–Bera test produces a p-value of 0.000096, indicating a departure from normality in the forecast-error distribution. This may reflect asymmetry, heavy-tailed behavior, or extreme forecast deviations

during specific periods. The ARCH-LM test also produces a p-value of 0.013057, suggesting possible time-varying error variance or volatility clustering. In addition, the BDS test produces a p-value of 0.022031, providing indicative evidence of remaining nonlinear dependence in the forecast errors [30, 31, 32, 33].

Taken together, these results suggest that although the SARIMAX model largely addresses linear dependence, some non-normality, variance instability, and nonlinear error structure may remain. However, given the limited number of out-of-sample forecast errors, these findings should be interpreted as exploratory diagnostic evidence rather than definitive statistical conclusions. More importantly, the empirical performance results show that the residual-learning hybrid models do not improve forecasting accuracy over the simpler ARIMA and SARIMA benchmarks. This implies that the remaining nonlinear or unstable error structure may not be sufficiently regular, stable, or learnable within the available sample size to generate better out-of-sample forecasts using MLP or SVM. Therefore, the diagnostic results support a more nuanced interpretation: although some nonlinear structure may remain in the SARIMAX forecast errors, additional model complexity does not necessarily translate into improved forecasting performance under the current rolling-origin evaluation framework.

5.5. Statistical Significance of Forecast Accuracy Differences

To examine whether the numerical differences in forecasting performance are statistically significant, this study applies the Harvey–Leybourne–Newbold corrected Diebold–Mariano test. Since ARIMA achieves the lowest RMSE in Table 8, it is used as the benchmark model in the pairwise comparisons. Squared-error loss is used to align the test with RMSE as the primary metric for larger forecast deviations. The Diebold–Mariano test is widely used to compare predictive accuracy between competing forecasting models, while the Harvey–Leybourne–Newbold correction improves the finite-sample behavior of the test statistic [22, 29].

Table 10. Harvey-Leybourne-Newbold Corrected Diebold-Mariano Test Results

Comparison	Loss Function	Mean Loss Difference	HLN-DM Statistic	p-value	Lower Loss Model	Significance
ARIMA vs SARIMA	Squared	-0.000606	-0.057540	0.954612	ARIMA	ns
ARIMA vs SARIMAX	Squared	-0.005716	-0.497984	0.623224	ARIMA	ns
ARIMA vs MLP	Squared	-0.003162	-0.518375	0.609148	ARIMA	ns
ARIMA vs SVM	Squared	-0.002489	-0.760808	0.454501	ARIMA	ns
ARIMA vs SARIMAX–MLP	Squared	-0.005076	-0.478166	0.637046	ARIMA	ns
ARIMA vs SARIMAX–SVM	Squared	-0.007102	-0.507729	0.616478	ARIMA	ns

Note: The Harvey–Leybourne–Newbold (HLN)-corrected Diebold–Mariano tests are based on loss differences computed from the same 24 rolling-origin one-step-ahead forecasts over the January 2023–December 2024 evaluation period. The notation "ns" indicates that the difference is not statistically significant at the 5% level.

Because the Diebold–Mariano comparisons are based on only 24 one-step-ahead forecast errors, the test results should be interpreted with caution, particularly with respect to statistical power. As presented in Table 10, none of the pairwise comparisons between ARIMA and the competing models is statistically significant at the 5% level. Therefore, although ARIMA achieves the lowest RMSE and the highest R^2, its advantage should be interpreted as numerical rather than statistically significant.

The results in Table 10 show that ARIMA produces lower average squared forecast loss than each competing model. However, since all p-values exceed conventional significance levels, the null hypothesis of equal predictive accuracy cannot be rejected for any pairwise comparison. This finding indicates that the observed differences in squared forecast loss are not strong enough to establish statistically significant predictive superiority.

Overall, the statistical testing results support a balanced interpretation of the model comparison. The findings suggest that simpler statistical models, particularly ARIMA, provide competitive and numerically stronger forecasting performance under the revised evaluation framework. However, the evidence does not support a claim of statistically significant dominance by ARIMA. Similarly, the results do not support the claim that more complex models, including SARIMAX, standalone machine learning models, or hybrid models, significantly outperform the simpler statistical benchmarks. This interpretation is consistent with the broader forecasting literature, which

emphasizes that model complexity should be justified through out-of-sample validation rather than assumed to improve accuracy [6, 17, 19].

5.6. Discussion of Model Complexity

The empirical findings provide important insight into the relationship between model complexity and forecasting performance. In principle, SARIMAX, standalone machine learning models, and hybrid models have greater flexibility than ARIMA and SARIMA. SARIMAX can include external variables and intervention effects, machine learning models can represent nonlinear relationships, and hybrid models can combine statistical and nonlinear residual-learning components. However, the results show that these additional modeling capabilities do not automatically lead to better short-term forecasting performance.

The superior numerical performance of ARIMA and SARIMA suggests that the flexible packaging demand series contains strong historical temporal information that can be captured by relatively parsimonious statistical models. In contrast, the selected exogenous variables and nonlinear learning components do not provide sufficient incremental forecasting value in the 24-month evaluation period. This finding is important because it suggests that model selection should consider not only theoretical sophistication, but also empirical performance, data availability, interpretability, and implementation burden [1, 6].

From a methodological perspective, the results caution against assuming that hybrid models are always superior to statistical models. Hybrid models are most useful when the residuals from the statistical component contain systematic nonlinear patterns that can be learned reliably. In this study, the diagnostic results indicate some remaining nonlinear dependence in the SARIMAX forecast errors. However, the forecasting results show that the residual-learning components do not translate this diagnostic signal into improved out-of-sample performance. This suggests that the remaining nonlinear structure may be unstable, weak, or insufficiently represented in the available monthly sample [18, 20].

From a practical perspective, the findings support the use of simple and interpretable forecasting models as strong benchmarks for industrial demand planning. ARIMA and SARIMA require fewer input variables, are easier to implement, and provide competitive forecasting accuracy. This is particularly relevant for companies with limited historical data, limited access to reliable external indicators, or a need for transparent forecasting procedures. More complex models may still be valuable in other settings, especially when larger datasets, higher-frequency observations, stronger nonlinear structures, or more reliable leading indicators are available. However, their use should be justified through consistent out-of-sample validation [6, 17].

5.7. Managerial Implications

The results provide several implications for managers in the flexible packaging industry. First, production planners should not automatically assume that more complex forecasting models will improve forecasting accuracy. In the present study, ARIMA and SARIMA provide the strongest numerical forecasting performance while requiring only historical demand data. This makes them practical tools for routine short-term demand planning, especially when external indicators are difficult to obtain, are reported at lower frequency, or are not sufficiently stable for monthly forecasting.

Second, the choice of forecasting model should be aligned with the operational objective. If the main concern is avoiding large forecast deviations that may cause capacity imbalance, excess inventory, or stock shortages, ARIMA may be preferable because it achieves the lowest RMSE. If the main concern is minimizing average absolute or percentage error, SARIMA may be more suitable because it achieves the lowest MAE and MAPE. This metric-dependent interpretation allows managers to select models according to the cost structure and planning priorities of their operations [1, 6].

Third, the results suggest that external variables should be used selectively. Although macroeconomic and sectoral indicators are theoretically relevant to flexible packaging demand, their incremental predictive value may be limited in short-term monthly forecasting. Managers should therefore validate whether external variables improve forecast accuracy before incorporating them into operational forecasting systems. Adding external variables without demonstrated predictive benefit may increase data-management complexity without improving planning outcomes.

Fourth, machine learning and hybrid models should be treated as conditional tools rather than default choices. These models may be useful when larger datasets, more stable nonlinear patterns, or richer high-frequency predictors are available. However, when the available data are limited and the predictive signal is mostly contained in the historical demand series, simpler statistical models may provide more reliable and interpretable forecasts. Therefore, managers should periodically benchmark complex models against simple statistical alternatives to ensure that additional complexity provides measurable operational value [17, 19].

Finally, the findings highlight the importance of continuous forecast monitoring. Structural changes such as the COVID-19 pandemic can alter demand behavior and weaken previously stable relationships between demand and external drivers. Forecasting models should therefore be reviewed regularly using rolling-origin evaluation, error diagnostics, and updated performance metrics. This practice can help managers detect model deterioration, adjust forecasting procedures, and maintain reliable demand-planning decisions under changing market conditions.

6. Conclusion

This study provides a comprehensive reassessment of statistical, machine learning, and hybrid forecasting models for monthly demand prediction in the Indonesian flexible packaging industry under post-COVID structural change. The evaluated models include ARIMA, SARIMA, SARIMAX with selected exogenous variables and a COVID-19 intervention term, standalone MLP and SVM models, and hybrid SARIMAX–MLP and SARIMAX–SVM models. To ensure a fair comparison, all models are evaluated under a unified rolling-origin, one-step-ahead forecasting framework using the same forecast origins, evaluation period, and performance metrics. This type of consistent out-of-sample evaluation is important because model performance can be strongly affected by the forecasting horizon, data split, and evaluation design [1, 18, 19].

The empirical results show that simpler statistical models provide the strongest overall performance under the rolling-origin one-step-ahead evaluation framework. ARIMA(1,1,3) achieves the lowest RMSE and highest R^2, while SARIMA$(0,1,1)(0,1,0)_{12}$ achieves the lowest MAE and MAPE. In contrast, SARIMAX, standalone machine learning models, and hybrid SARIMAX–ML models do not consistently improve forecasting accuracy. These findings suggest that the additional complexity introduced by exogenous variables and residual-based machine learning is not necessarily beneficial when the available monthly sample is limited and the remaining nonlinear structure is unstable or weak. This result is consistent with the broader forecasting view that more complex methods do not automatically outperform simpler alternatives and should be justified through empirical validation [6, 17].

The results also show that SARIMAX does not outperform ARIMA or SARIMA despite incorporating selected exogenous variables and a COVID-19 step intervention term. This finding suggests that, within the present dataset and rolling-origin evaluation framework, the selected external variables provide limited additional predictive value beyond the historical demand structure. This does not imply that macroeconomic and sectoral indicators are irrelevant to flexible packaging demand. Rather, it indicates that their practical contribution to short-term out-of-sample forecasting accuracy is limited under the current model specification and forecasting horizon. Therefore, the use of exogenous variables in forecasting should be supported not only by theoretical relevance, but also by demonstrated predictive improvement [1, 10].

The intervention-specification comparison indicates that the representation of the COVID-19 effect influences SARIMAX performance. The ramp intervention provides the lowest RMSE and information criteria, suggesting that the post-COVID adjustment may have occurred gradually rather than as a purely abrupt level shift. However, the step intervention provides lower MAE and MAPE, which are more directly related to average forecast accuracy. This mixed result supports the interpretation that the COVID-19 effect is structurally complex and cannot be fully represented by a single intervention form. Nevertheless, the broader model comparison remains unchanged: neither SARIMAX nor the hybrid SARIMAX–MLP and SARIMAX–SVM models consistently outperform the simpler ARIMA and SARIMA benchmarks. This finding supports the need to treat pandemic-related intervention variables as simplified structural controls rather than complete representations of all disruption and recovery dynamics [7, 8].

Standalone machine learning and hybrid models also do not provide consistent improvements over the simpler statistical benchmarks. Although SVM demonstrates relatively competitive performance among the nonlinear

models, it does not surpass ARIMA or SARIMA across the main evaluation metrics. Similarly, MLP does not provide clear forecasting improvement, suggesting that the available monthly dataset may not contain sufficiently strong or stable nonlinear patterns for neural network-based forecasting to generalize effectively. The hybrid SARIMAX–MLP and SARIMAX–SVM models also fail to outperform ARIMA and SARIMA, indicating that the SARIMAX residual series may not contain enough systematic nonlinear structure for residual-based machine learning to exploit. These findings reinforce the view that machine learning and hybrid models are conditional tools whose effectiveness depends on data size, feature quality, signal stability, and residual structure [12, 14, 18].

The Harvey–Leybourne–Newbold corrected Diebold–Mariano test provides a more cautious interpretation of the results. Although ARIMA produces lower average squared forecast loss than the competing models, the differences are not statistically significant at conventional levels. Therefore, the numerical advantage of ARIMA should not be interpreted as statistically significant predictive superiority. Instead, the findings indicate that simpler statistical models are highly competitive and practically preferable in this empirical setting due to their accuracy, interpretability, parsimony, and ease of implementation.

Overall, this study demonstrates that increased model complexity does not necessarily lead to improved forecasting performance. The effectiveness of SARIMAX, machine learning, and hybrid forecasting models depends strongly on data characteristics, sample size, stability of exogenous relationships, residual structure, and the evaluation metric used. For the Indonesian flexible packaging demand series examined in this study, ARIMA and SARIMA provide the most reliable and practically useful forecasting performance under the rolling-origin one-step-ahead evaluation framework. These findings reinforce the importance of empirical validation, model parsimony, interpretability, and practical relevance in industrial demand forecasting [1, 6, 17].

6.1. Limitations

Despite its contributions, this study has several limitations. First, the analysis is based on monthly observations from January 2012 to December 2024 for a single industry and country. Although the dataset is relevant to the Indonesian flexible packaging market, the findings may not be directly generalizable to other industries, countries, or demand environments without further empirical validation. Forecasting performance is often context-dependent, and model rankings may change across different data structures, sectors, and market conditions [17, 18].

Second, the sample size is relatively limited because the data are observed at a monthly frequency. This limitation may affect the performance of machine learning and hybrid models, which generally require sufficient observations to learn stable nonlinear relationships. With a larger dataset or higher-frequency observations, the relative performance of MLP, SVM, and hybrid models may differ. Therefore, the underperformance of machine learning and hybrid models in this study should be interpreted within the context of the available monthly sample and the short-term forecasting horizon [12, 14].

Third, several exogenous variables originate from different reporting frequencies and are harmonized into monthly observations. Although this procedure is necessary to align the variables with the monthly demand series, it may reduce the short-term predictive contribution of some external indicators. Future studies may benefit from using more granular and directly observed monthly explanatory variables. This limitation is important because the usefulness of exogenous variables depends not only on economic relevance, but also on data frequency, timing, quality, and stability of the relationship with the target series [1, 10].

Fourth, although alternative COVID-19 intervention specifications were evaluated, including pulse, step, and ramp structures, each specification remains a simplified representation of pandemic-related demand dynamics. The step intervention used in the main specification provides a simple and interpretable structural control for the post-COVID period, but it may not fully capture temporary shocks, lockdown phases, recovery dynamics, or gradual adjustment patterns. Therefore, the intervention effect should be interpreted as a structural control rather than as a complete representation of all pandemic-related changes in demand [7, 8].

Fifth, this study focuses on short-term one-step-ahead forecasting. This horizon is relevant for monthly demand planning, but the results may differ for multi-step-ahead forecasting, where uncertainty accumulates and model behavior may change. Therefore, the findings should not be generalized to longer forecasting horizons without additional empirical testing. Forecasting performance may vary substantially depending on whether the objective is one-step-ahead operational planning or longer-horizon strategic planning [1, 18].

Finally, the study evaluates selected machine learning models and hybrid configurations, specifically MLP, SVM, SARIMAX–MLP, and SARIMAX–SVM. Other nonlinear learning approaches, feature-selection strategies, or hybrid structures may produce different results. However, the present findings emphasize that additional model complexity should be justified through consistent out-of-sample evaluation rather than assumed to improve forecasting accuracy [6, 17]

6.2. Future Research

Future research can extend this study in several directions. First, alternative intervention specifications should be examined, including pulse interventions, ramp functions, segmented recovery variables, and time-varying intervention effects. These alternatives may better represent the different phases of COVID-19 disruption, including the initial shock, adjustment period, and recovery phase. More flexible intervention modeling may provide a clearer understanding of how structural shocks affect industrial demand over time [7, 8].

Second, future studies may use longer time series, higher-frequency data, or more detailed industry-level demand observations. More granular datasets may improve the ability of machine learning and hybrid models to capture nonlinear relationships and may provide stronger evidence regarding the conditions under which complex forecasting models are beneficial. This is particularly important because data availability and signal stability strongly influence the performance of nonlinear forecasting models [12, 14].

Third, future research should evaluate the forecasting framework across different industries, product groups, or countries. Cross-sector and cross-country validation would help determine whether the finding that simpler statistical models remain competitive is specific to the Indonesian flexible packaging market or applies more broadly to industrial demand forecasting under structural change. Such validation is important because forecasting models should be assessed under diverse empirical conditions before broader generalization is made [17, 19].

Fourth, future studies may incorporate additional evaluation criteria beyond statistical accuracy. For industrial decision-making, forecast value should also be assessed using inventory cost, production planning efficiency, service-level improvement, capacity utilization, and supply chain risk reduction. Such decision-based evaluation would provide a more direct link between forecasting accuracy and managerial benefit [2, 3].

Fifth, future research may explore alternative machine learning and hybrid methods, including tree-based ensembles, gradient boosting, recurrent neural networks, temporal convolutional networks, and attention-based models. However, these models should be evaluated under the same fair rolling-origin framework and compared against strong statistical benchmarks. This is necessary to ensure that any observed improvement is due to genuine predictive ability rather than differences in evaluation design [12, 17, 19].

Finally, future work should further investigate residual diagnostics before applying hybrid residual learning. Hybrid models may be most useful when residuals from the statistical component contain systematic nonlinear patterns. Therefore, future studies can develop a diagnostic-based hybrid modeling strategy in which residual-learning models are applied only when residual analysis provides evidence of exploitable structure. This would help ensure that hybrid modeling is used selectively and empirically, rather than automatically, in industrial forecasting applications [18, 20, 21].

Acknowledgement

The authors gratefully acknowledge the academic guidance and institutional support that contributed to the completion of this research. Appreciation is also extended to the institutions that facilitated access to relevant data and information. The views and conclusions expressed in this paper are solely those of the authors and do not necessarily reflect the official positions of the associated institutions.

REFERENCES

1. R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed., OTexts, 2021.

2. N. Kourentzes, J. R. Trapero, and D. K. Barrow, "Optimising forecasting models for inventory planning," *International Journal of Production Economics*, vol. 223, article 107526, 2020.
3. R. Fildes, S. Ma, and S. Kolassa, "Retail forecasting: Research and practice," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1283–1318, 2022.
4. M. Yasir, Y. Ansari, K. Latif, and S. Rho, "Machine learning–assisted efficient demand forecasting using endogenous and exogenous indicators for the textile industry," *International Journal of Logistics Research and Applications*, vol. 27, no. 12, pp. 1–20, 2024.
5. J. P. A. Ioannidis, S. Cripps, and M. A. Tanner, "Forecasting for COVID-19 has failed," *International Journal of Forecasting*, vol. 38, no. 2, pp. 423–438, 2020.
6. F. Petropoulos, S. Makridakis, N. Stylianou, and K. Nikolopoulos, "Simple versus complex forecasting methods: Evidence from COVID-19," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1366–1379, 2022.
7. Y. Chen, Y. Wang, and J. Zhang, "Modeling the impact of COVID-19 on time series forecasting using intervention analysis," *Chaos, Solitons & Fractals*, vol. 142, article 110432, 2021.
8. R. J. Hyndman and B. Rostami-Tabar, "Forecasting interrupted time series," *Journal of the Operational Research Society*, vol. 76, no. 4, pp. 790–803, 2024.
9. G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed., Wiley, 2015.
10. X. Shao and L. Li, "Modeling and forecasting time series with multiple seasonal patterns and external regressors," *Applied Mathematical Modelling*, vol. 89, pp. 497–514, 2021.
11. M. A. Valizadeh, I. K. Shamir, A. Ahmadi, and M. Mirhosseini, "Economic and technical assessment of wind potential using SARIMAX time series models: Wind speed forecasting and analysis," *Energy Science & Engineering*, vol. 13, no. 12, pp. 1–26, 2025.
12. B. Lim and S. Zohren, "Time-series forecasting with deep learning: A survey," *Philosophical Transactions of the Royal Society A*, vol. 379, no. 2194, article 20200209, 2021.
13. O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, "Financial time series forecasting with deep learning: A systematic literature review," *Applied Soft Computing*, vol. 90, article 106181, 2020.
14. T. Papadimitriou, P. Gogas, and K. Papadimitriou, "Machine learning and econometric approaches to forecasting under structural breaks," *Expert Systems with Applications*, vol. 170, article 114553, 2021.
15. K. Bandara, C. Bergmeir, and S. Smyl, "Forecasting across time series databases using recurrent neural networks on groups of similar series," *Expert Systems with Applications*, vol. 140, article 112896, 2020.
16. A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," Statistics and Computing, vol. 14, no. 3, pp. 199–222, 2004.
17. S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 competition: 100,000 time series and 61 forecasting methods," *International Journal of Forecasting*, vol. 36, no. 1, pp. 54–74, 2020.
18. P. Montero-Manso and R. J. Hyndman, "Principles and algorithms for forecasting groups of time series: Locality and globality," *International Journal of Forecasting*, vol. 37, no. 1, pp. 231–249, 2021.
19. P. Montero-Manso, T. S. Talagala, R. J. Hyndman, and G. Athanasopoulos, "FFORMPP: Feature-based forecast model performance prediction," *International Journal of Forecasting*, vol. 36, no. 4, pp. 1410–1423, 2020.
20. M. Obaidat, H. A. Almomani, R. Mallouhy, A. AlMotari, and O. T. Al Meanazel, "A hybrid machine learning framework for daily dairy demand forecasting: Integrating SARIMAX and XGBoost for seasonal production optimization," *IEEE Access*, vol. 13, pp. 162668–162680, 2025.
21. I. A. Kachalla, C. Ghiaus, A. Ademuwagun, O. B. Odeyinde, and M. Baseer, "Data-driven hybrid SARIMAX-MLP framework for energy consumption prediction in residential micro-grid," *Results in Engineering*, vol. 26, article 105336, 2025.
22. F. X. Diebold and R. S. Mariano, "Comparing predictive accuracy," *Journal of Business & Economic Statistics* vol. 13, no. 3, pp. 253–263, 1995.
23. K. Kashif and R. Ślepaczuk, "LSTM-ARIMA as a hybrid approach in algorithmic investment strategies," *Knowledge-Based Systems*, vol. 320, article 113563, 2025.
24. H. Man, H. Huang, Z. Qin, and Z. Li, "Analysis of a SARIMA-XGBoost model for hand, foot, and mouth disease in Xinjiang, China," *Epidemiology and Infection*, vol. 151, article e200, 2023.
25. D. Kubek and P. Więcek, "Hybrid demand forecasting in fuel supply chains: ARIMA with non-homogeneous Markov chains and feature-conditioned evaluation," *Energies*, vol. 18, no. 22, article 6044, 2025.
26. Z. Mohammed, C. Anas, and M. El Hammoumi, "A hybrid learning framework for forecasting uncertainty and adaptive inventory planning in retail supply chains," *Supply Chain Analytics*, vol. 13, article 100180, 2025.
27. D. Atif, "Enhancing long-term GDP forecasting with advanced hybrid models: A comparative study of ARIMA-LSTM and ARIMA-TCN with dense regression," *Computational Economics*, vol. 65, pp. 3447–3473, 2025.
28. M. Prastuti, I. N. Pujawan, E. Widodo, H. Kuswanto, "Machine Learning Methods for Accurate Demand Forecasting in the Cement Industry" *Springer*, vol. 257, pp. 665-–680, 2026.
29. David I. Harvey, Stephen J. Leybourne, Paul Newbold, "Tests for Forecast Encompassing" *Journal of Business & Economic Statistics*, vol. 2, pp. 254-–259, 1998.
30. G. M. Ljung, G. E. P. Box, "On a Measure of Lack of Fit in Time Series Models" *Biometrika*, vol. 65, no. 2, pp. 297-–303, 1978.
31. Carlos M. Jarque and Anil K. Bera , "Efficient Test for Normality, Homoscedasticity and Serial Independence of Regression Residuals" *Economic Letters*, vol. 6, pp. 255-–259, 1980.
32. David I. Harvey, Stephen J. Leybourne, Paul Newbold, "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation" *Econometric Society*, vol. 50, no. 4, pp. 987-–1007, 1982.
33. Broock, W.A., Scheinkman, J.A., Dechert, W.D. and LeBaron, B. "A Test for Independence Based on the Correlation Dimension" *Econometric Reviews*, vol. 15, pp. 197-–235, 1996.