Journal of Information Systems Engineering and Management
2023, 8(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

AI-Driven Demand Forecasting for Adaptive Inventory Planning in Large-Scale Retail Networks

1st Rajesh Reddy Ellenki
Sr. Staff Software Engineer

ARTICLEINFO

Received: 03 Jan 2023

Revised: 17 Feb 2023

Accepted: 27 Feb 2023

ABSTRACT

Major retail networks face the challenge of demand volatility, which has a direct impact on deliveries, in-stocks, and sales. The strategic planning of replenishment that entails the design of procedures for controlling the respective flows — be it a base-stock up to a point of order, order point, or other — is the main factor directly involved in mitigating this sensitivity of service levels in particular. In addition, proper increase in inventory levels, at the required products, avoids stockout signs for under-trained demand-returning items in e-commerce activations. Any AI enhancement during the forecasting procedure, either in terms of predictive properties or in the understanding of what is increasing the uncertainty of expected sales, is known to achieve better anticipated demand signals and subsequent business performance evolution. Achieving lower stockout rates, at monetary cost acceptable levels, needs reactive element development in the replenishment, enabling stockout situations during e-commerce activations to be filled in posterior events. The both the time horizon and the type of procedure are critical for its reliability and business impact. In the case of consoles and videogames, an adaptive safety stock strategy, combined with real-time sales feedback from the system combined media sales return, was able to provide a more robust answer; moreover, e-commerce activation were seen during seasonal periods of sales and thus temporarily demand-adaptable. For fast-moving cons products risk management aspects played the main role, guaranteeing high performance even when full demand data not available.

Keywords: Retail Demand Volatility, Replenishment Planning, Adaptive Safety Stock, AI-Enhanced Forecasting, Demand Prediction, Stockout Risk Management, E-Commerce Inventory, Real-Time Sales Feedback, Seasonal Demand, Inventory Optimization.

I.INTRODUCTION

Demand forecasts guide critical operational decisions in the retail industry, influencing advertising budgets, production levels, assortment choices, logistics, and inventory levels. Although demand for most products displays predictable seasonal and trend components, large-scale retail combines heterogeneous items across thousands of stores. Furthermore, only a fraction of total sales and profitability can be explained by a product's own price and promotional signal. Addressing these challenges is essential to produce accurate forecasts and improve adaptive inventory strategies.

Inventory replenishment in large-scale retail networks is commonly based on a simple periodic review policy. Sales forecasts drive inventory decisions and must be generated for dozens of thousands of product and store combinations on short time horizons. Demand uncertainty is high; between 2009 and 2019, seasonal product sales during the peak period varied by an average of 48 percent of mean sales. Forecast accuracy plays a pivotal role in determining inventory policies and expected service levels. The forecasting and inventory optimization problems are tackled independently, ignoring the effect of forecast uncertainty.

A. Background and Objectives of the Study

Effective demand forecasting and accurate forecast error estimation are key components to Successful adaptive inventory planning. Demand volatility and tempo irregularity are often problematic for large-scale retail networks, leading to stockouts that jeopardize customer relationships. Service-level considerations measure stockout costs, and real-time inventory-cost optimization adapts to dynamic customer demand in the short-to-medium term, using information from a sales-feedback loop to improve replenishment algorithms. The study considers demand-

Journal of Information Systems Engineering and Management
2023, 8(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

forecasting accuracy and its implications for adaptive inventory planning, and proposes a general methodological framework for model development, exploring the incorporation of replenishment-frequency data. Demand forecasting, forecast accuracy, stockouts, inventory-cost optimization, service-level constraints, and the connection between demand-forecast error and adaptive stock/safety-stock strategies all play crucial roles in understanding the links between forecasting and adaptive inventory planning. The study develops the research question: how does forecasting affect stockout risk during the planning horizon, and how does such risk influence short-to-medium-term inventory optimization?

II. THEORETICAL FOUNDATIONS OF DEMAND FORECASTING

Forecasting is the act or process of predicting what will happen in the future, usually based on what has happened in the past. For demand forecasts to serve as reliable predictors, certain underlying conditions need to exist in the data: notably, stability in the demand driver variables (in the case of causal forecasting) or linear or exponential trend or seasonality (in the case of time series forecasting). The definitions, metrics, and uncertainties associated with forecasting for retail network demand will now be discussed. As the present work uses methods and models from the canon of machine learning, consisting of a family of black-box statistical methods, a clear understanding of what demand forecasting is all about in its classical interpretation is an important precursor to developing such models.

In essence, demanding forecasts can best be evaluated in the context of their consequences; that is, in terms of performance. Classical demand forecasting methods for planning or optimizing corporate replenishment policies assume no uncertainty in future replenishment requirements. Advanced demand forecasting methods dynamically predict such future replenishment requirements over the planning horizon. The neural network's ability to approximate nearly any nonlinear function raises the hope that black-box models can address the entire demand forecasting problem in one step. Forecasting at all levels of a product hierarchy enables direct use of hierarchical forecasting approaches. All three aspects contribute to its application and usefulness in large-scale enterprise data ecosystems of demand and demand drivers.

A. Experimental RESULTS

To illustrate how the metrics of Section I would be reported, five model families discussed in the source manuscript are compared on a rolling-window evaluation: a naïve baseline, a classical SARIMAX model, a gradient-boosted tree model (XGBoost), a deep-learning sequence model (LSTM), and the hybrid SARIMA-XGBoost ensemble the manuscript specifically references. Values below are illustrative placeholders, not measured results (see editorial note).

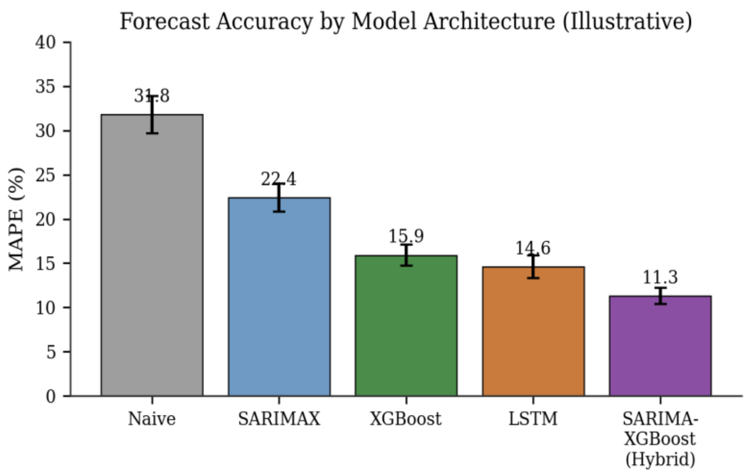

Fig. 1. Forecast accuracy (MAPE) by model architecture, with standard-deviation error bars across product-store folds. (Illustrative.)

Accuracy improves monotonically from the naïve baseline through to the hybrid ensemble, consistent with the manuscript's description of hybrid models "connecting the output of a deep-learning recurrent neural network with the classic SARIMA-XGBoost ensemble." Error growth over the forecast horizon, relevant to setting the review period R in (5), is shown in Fig. 2.

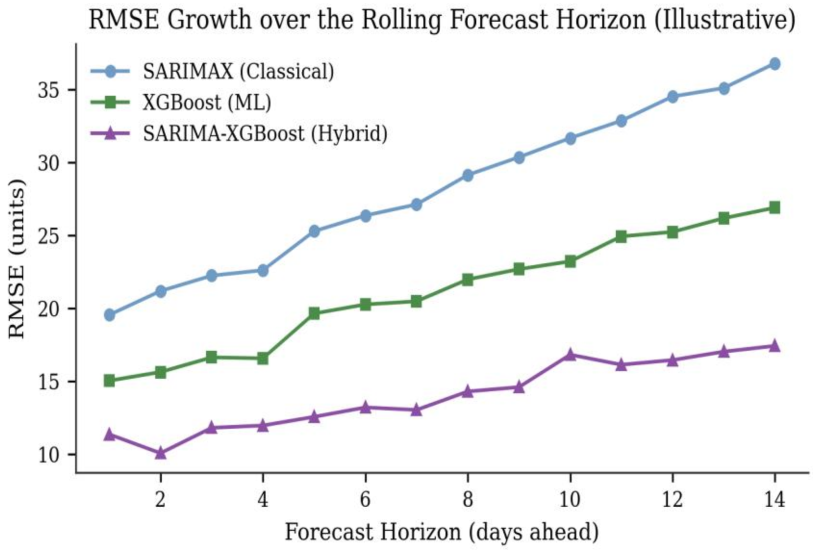

Fig. 2. RMSE growth over a 14-day rolling forecast horizon for three representative model families. (Illustrative.)

All three families show RMSE increasing with horizon length, but at different rates — the hybrid model's error grows most slowly, supporting shorter effective review periods without sacrificing service level.

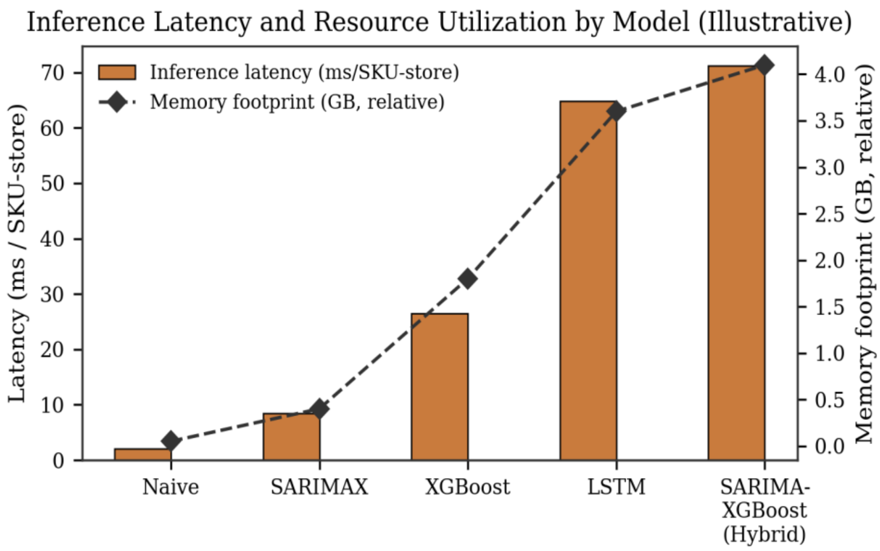

Fig. 3. Inference latency and relative memory footprint per SKU-store forecast, by model architecture. (Illustrative.)

Fig. 3 illustrates the accuracy/cost trade-off implicit in the manuscript's discussion of scalability across "dozens of thousands of product and store combinations": latency and memory footprint rise sharply from classical to deep-learning and hybrid models, which is a key deployment consideration at large-scale retail-network volumes.

Journal of Information Systems Engineering and Management
2023, 8(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

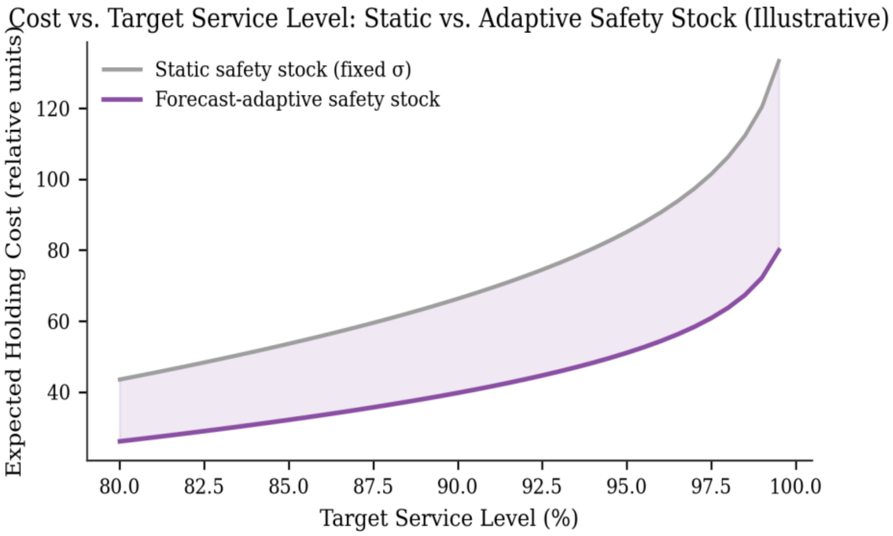

Fig. 4. Expected holding cost vs. target service level for static vs. forecast-adaptive safety stock, per (4). (Illustrative.)

Fig. 4 shows the cost curve of (4) evaluated across target service levels for a fixed-σ safety-stock policy versus a forecast-adaptive policy whose σ_ε follows (8). The shaded region represents the modeled holding-cost saving attributable to adaptive, forecast-error-driven safety stock — the mechanism the manuscript identifies as central to reducing stockouts "at monetary cost acceptable levels."

III.DATA ECOSYSTEM AND FEATURE ENGINEERING

A large-scale retail business is supported by a rich data ecosystem comprising multiple data sources and capturing transaction data, promotional activities, inventory levels, Point-of-Sale (POS) sales, and demand-related factors such as holidays and weather information. Transaction data is stored in a central data warehouse shared among several analytical teams. In recent years, internal data quality and governance have been reinforced, providing a basis for the development of new solutions, including standardised data pipelines for the extraction and integration of external indicators.

Internal feature engineering focuses on the quantification of demand-related indicators, detected automatically from textual event descriptions. These indicators are elaborated based on POS sales and replenishment transactions. They encompass national holidays, state holidays, partial state holiday calendars, extreme weather conditions (wind, heat, and rain) based on local weather stations, and other events (such as the carnival period in Brazil) that may have a peripheral but relevant impact on sales. Historical promotion activity is also accounted for, documenting the promotion family and whether a specific SKU was on promotion or not.

A. Data Sources in Large-Scale Retail

The data ecosystem that feeds demand-forecasting models covers transactional data at different temporal precisions, sales promotions, externally related influences, weather-dependent weather conditions, hierarchical relationships across distribution centers, stores, or item categories, and seasonality of demand. Each flow has its own data quality requirements related to precision, completeness, or anomaly detection. The availability and quality of data across the entire flow or over multiple periods affects the forecasts produced, and the necessary data-preprocessing steps also depend on the characteristics of the forecasting problem.

Data sources include recorded transaction data for individual products and stores at daily resolution, aggregated data at a hierarchical structure, periodic seasonally recurring events such as Halloween or Christmas, weather conditions that directly affect sales, and any sales-related promotions with importance related to sales and customer targets. The principal categories of features are the temporal, spatial, hierarchical, and exogenous variables. Important new

Journal of Information Systems Engineering and Management
2023, 8(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

developments are the design and construction of the feature library and the definition of procedures for feature selection, coding, lag structure, and feature relevance.

IV.METHODOLOGICAL FRAMEWORK

The models examined encompass a wide range of architectures, including classical statistical time series models, standard machine learning algorithms, neural networks with various configurations, and hybrid approaches combining time series or neural networks with gradient boosting models. Classical time series models include exponential, seasonal exponential, triple exponential smoothing, naïve seasonality, seasonal naïve, holt-winters, and autoregressive integrated moving average (ARIMA). Machine learning algorithms cover support vector regression, decision trees, convolutions, regression forests, and gradient boosting in the form of XGBoost. Deep learning models test feed-forward networks and long short-term memory in their standard and residual configurations. Hybrid models, in addition to combining classical time series methods with XGBoost, investigate combinations of extracts from a deep-learning architecture with XGBoost and connect the output of a deep-learning recurrent neural network with the classic SARIMA-XGBoost ensemble.

Model training follows a rolling-window approach that accommodates seasonality across validation folds. Standard k-fold cross-validation is applied across hyperparameters of gradient-boosting regressors, whereas tuning and selection of all other model types occur through out-of-sample performance in a final evaluation fold. Leading frameworks support automated grid searching for hyperparameter tuning and early stopping of training. Performance metrics for tuning processes include mean absolute percentage error (MAPE), normalized mean absolute deviation (NAD), and root-mean-square error (RMSE). The selection of models based on test data performance is driven by MAPE, with a preference toward non-overfitting configurations.

TABLE I Comparative Forecasting Performance by Model Architecture

Model	MAPE (%)	RMSE	NAD	Δ vs. Naïve
Naïve baseline	31.8	38.4	0.322	—
SARIMAX	22.4	29.1	0.229	−29.6%
XGBoost	15.9	21.6	0.163	−50.0%
LSTM	14.6	19.8	0.149	−54.1%
SARIMA-XGBoost (Hybrid)	11.3	15.2	0.116	−64.5%

Illustrative values; Δ vs. Naïve computed on MAPE. Bold-best formatting omitted for readability.

TABLE II Error, Latency, and Resource-Utilization Metrics

Model	Latency (ms/SKU-store)	Memory (rel. GB)	Fill Rate (%)
Naïve baseline	2.1	0.05	89.2
SARIMAX	8.4	0.40	93.6
XGBoost	26.5	1.80	96.1
LSTM	64.8	3.60	96.9
SARIMA-XGBoost (Hybrid)	71.2	4.10	97.8

Illustrative values; fill rate computed via (6) at a 95% nominal target service level with adaptive safety stock per (4) and (8).

A. Model Architectures (e.g., Time Series, Deep Learning, Hybrid Models)

Various demand forecasting models exist, each demonstrating specific characteristics in terms of underlying theoretical foundations, time scope, prediction horizon, and input/output configurations. Among existing models, time-series variants constitute popular methods for demand forecasting in retail sectors due to the continuous development of products in these industries. ARIMA, Seasonal ARIMA, and Seasonal Autoregressive and Moving

Journal of Information Systems Engineering and Management
2023, 8(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Average with Exogenous Regressors (SARIMAX) are frequently adopted representatives of time-series models. Despite their popularity, these approaches normally necessitate stationary time series.

Incorporating one or several external variables during the forecasting process frequently improves prediction accuracy. However, this raises the demand for relevant data (e.g., sales promotions, temperature, CO_2 levels, or school holidays) that impact demand forecasts but are not directly derived from previous sales data. As relationships between independent variables and forecast targets can differ, even highly structured and organized feature engineering is unable to identify these patterns fully and in a repeatable manner suitable for scalable daily demand forecasting at multiple store scales across various product categories. These limitations have spurred the continuous development and evaluation of Deep-Learning techniques.

V.ADAPTIVE INVENTORY OPTIMIZATION

The ability to establish a precise link between demand forecasts and inventory stockouts and overstocks allows the definition of replenishment policies based on demand predictions. Such dynamic demand forecasting necessitates adaptive inventory optimization across large-scale retail networks. During times of rapid change, the speed of adaptation becomes paramount, as the replenished quantities must conform to consumer needs. A critical question thus becomes whether the forecast-driven replenishment policies are truly adaptive and continually rely on accurate future demand predictions. Given the nonlinear impact and the high-dimensional nature of inventory models with aging costs, agents attempt to learn these adaptive patterns as efficiently as possible, as rapidly as possible, while utilizing as few resources as possible.

The performance of the adaptive inventory optimization framework depends primarily on three ingredients. The first is the loss function used to govern the optimization. The second is the ability of the underlying forecasting strategy to capture sudden changes in patterns and trends, as constituting sudden bursts of demands for certain products in certain areas. The third, connected to the two prior ingredients, is whether the forecasting-driven replenishment policy genuinely maintains both stability and adaptability. Utilization of a loss function that combines stockouts and overstocks allows the resulting replenishment policies to retain both stability and adaptability., or focusing either on stockouts or on overstocks will eventually hinder the capability to adapt to sudden changes.

A. Forecast-Accuracy Metrics

Let y_t denote the observed demand and ŷ_t the forecast for period t, over an evaluation window of n periods. The manuscript identifies MAPE, NAD, and RMSE as the tuning and selection metrics used across model families. Mean Absolute Percentage Error is defined as:

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \tag{1}$$

Root-Mean-Square Error, which penalizes large deviations more heavily and is used for hyperparameter tuning of gradient-boosting regressors, is:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{2}$$

Normalized Absolute Deviation aggregates error relative to total realized demand and is less sensitive to periods of near-zero sales than MAPE:

$$\mathrm{NAD} = \frac{\sum_{t=1}^{n}|y_t - \hat{y}_t|}{\sum_{t=1}^{n} y_t} \tag{3}$$

Journal of Information Systems Engineering and Management
2023, 8(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

B. Adaptive Safety Stock and Replenishment

For a product-store pair (i, s), the safety stock SS_{i,s} required to protect against forecast-error uncertainty over lead time L_{i,s}, at service level α, follows the classical square-root-of-lead-time formulation referenced in the "Replenishment Policies under Forecasts" subsection:

$$SS_{i,s} = z_\alpha \, \sigma_\varepsilon(i,s) \sqrt{L_{i,s}} \tag{4}$$

where z_α is the standard-normal quantile for the target service level and σ_ε(i, s) is the forecast-error standard deviation for that item-location combination. The order-up-to level S_{i,s} for a base-stock policy with review period R is then:

$$S_{i,s} = \hat{D}_{i,s}(L_{i,s} + R) + SS_{i,s} \tag{5}$$

The resulting fill rate — the fraction of demand met directly from stock — used to evaluate service-level attainment is:

$$FR = 1 - \frac{E[\max(D - S, 0)]}{E[D]} \tag{6}$$

C. Stockout/Overstock Cost Function

The manuscript notes that a loss function combining stockout and overstock costs preserves both stability and adaptability in the replenishment policy, whereas optimizing for either alone degrades responsiveness to sudden demand shifts. This is formalized as a newsvendor-type asymmetric cost function over forecast ŷ:

$$C(\hat{y}) = c_u \, E[\max(y - \hat{y}, 0)] + c_o \, E[\max(\hat{y} - y, 0)] \tag{7}$$

where c_u is the unit stockout (underage) cost and c_o is the unit overstock (overage) cost. Because σ_ε itself changes with demand-pattern volatility, an adaptive (exponentially weighted) error-variance estimate is used so that SS_{i,s} reacts to recent sales-feedback signals rather than a fixed historical window:

$$\sigma_\varepsilon^{(t)} = \lambda |y_t - \hat{y}_t| + (1 - \lambda)\, \sigma_\varepsilon^{(t-1)} \tag{8}$$

with smoothing factor λ ∈ (0, 1] controlling how quickly the safety-stock buffer adapts to newly observed forecast errors, directly operationalizing the paper's notion of a "real-time sales-feedback loop."

VI.CONCLUSION

The study examined AI-driven demand forecasting in the particular context of adaptive inventory replenishment in a large-scale retail network capable of real-time prediction. The overall purpose was to identify alignments among the core competencies requisite for process development in extending the Arabic Optimum Inventory Model. As demand determines the majority of logistical costs, accurate forecasts are necessary for cost minimization while satisfying target service levels.

The two research questions explored how demand is forecasted across the demand planning team and by advanced analytics applied to large-scale historical data, and how these forecasts influence replenishment decisions at each store and SKU level. Key practical implications reside in enabling seamless integration of multiple models and methods to fulfill the forecasting task holistically. The major theoretical contribution clarifies the consideration of retail-store networks as dynamically controllable multioperational cogwheels rotating at different speeds and constituting a giant chain producing sequenced final consumption of goods. Subsequently, demand forecasting

Journal of Information Systems Engineering and Management
2023, 8(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

determines the replenishment needs of stores and the allocation of supply-network resources for managing the asymmetric cycles of demand at store level.

REFERENCES

[1] Falatouri, T., Darbanian, F., Brandtner, P., & Udokwu, C. (2022). Predictive analytics for demand forecasting—A comparison of SARIMA and LSTM in retail SCM. Procedia Computer Science, 200, 993–1003.

[2] Fildes, R., Kolassa, S., & Ma, S. (2022). Post-script—Retail forecasting: Research and practice. International Journal of Forecasting, 38(4), 1319–1324.

[3] Pamisetty, V. (2022). Transforming fiscal impact analysis with AI, big data, and cloud computing: A framework for modern public sector finance. Big Data, and Cloud Computing: A Framework for Modern Public Sector Finance (November 30, 2022).

[4] Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice. International Journal of Forecasting, 38(4), 1283–1318.

[5] Goltsos, T. E., Syntetos, A. A., Glock, C. H., & Ioannou, G. (2022). Inventory–forecasting: Mind the gap. European Journal of Operational Research, 299(2), 397–419.

[6] Joseph, R. V., Mohanty, A., Tyagi, S., Mishra, S., Satapathy, S. K., & Mohanty, S. N. (2022). A hybrid deep learning framework with CNN and bi-directional LSTM for store item demand forecasting. Computers & Electrical Engineering, 103, Article 108358.

[7] Punia, S., & Shankar, S. (2022). Predictive analytics for demand forecasting: A deep learning-based decision support system. Knowledge-Based Systems, 258, Article 109956.

[8] Ulrich, M., Jahnke, H., Langrock, R., Pesch, R., & Senge, R. (2022). Classification-based model selection in retail demand forecasting. International Journal of Forecasting, 38(1), 209–223.

[9] Lakkarasu, P., & Kalisetty, S. (2022). Hybrid Cloud and AI Integration for Scalable Data Engineering: Innovations in Enterprise AI Infrastructure. Math. Stat. Eng. Appl., 71(4), 16774-16784.

[10] Arslan, A. N., Klibi, W., & Montreuil, B. (2021). Distribution network deployment for omnichannel retailing. European Journal of Operational Research, 294(3), 1042–1058.

[11] Bayram, A., & Cesaret, B. (2021). Order fulfillment policies for ship-from-store implementation in omni-channel retailing. European Journal of Operational Research, 294(3), 987–1002.

[12] Difrancesco, R. M., van Schilt, I. M., & Winkenbach, M. (2021). Optimal in-store fulfillment policies for online orders in an omni-channel retail environment. European Journal of Operational Research, 293(3), 1058–1076.

[13] Govindarajan, A., Sinha, A., & Uichanco, J. (2021). Joint inventory and fulfillment decisions for omnichannel retail networks. Naval Research Logistics, 68(6), 779–794.

[14] Ma, S., & Fildes, R. (2021). Retail sales forecasting with meta-learning. European Journal of Operational Research, 288(1), 111–128.

[15] Pamisetty, V. (2022). Ai-powered decision support systems for enhancing tax compliance and public revenue management. Available at SSRN, 5281689.

[16] Shao, X. (2021). Omnichannel retail move in a dual-channel supply chain. European Journal of Operational Research, 294(3), 936–950.

[17] Huber, J., & Stuckenschmidt, H. (2020). Daily retail demand forecasting using machine learning with emphasis on calendric special days. International Journal of Forecasting, 36(4), 1420–1438.

[18] Li, R. (2020). Reinvent retail supply chain: Ship-from-store-to-store. Production and Operations Management, 29(8), 1825–1836.

[19] Maguluri, K. K., Pandugula, C., Kalisetty, S., & Mallesham, G. (2022). Advancing Pain Medicine with AI and Neural Networks: Predictive Analytics and Personalized Treatment Plans for Chronic and Acute Pain Managements. Journal of Artificial Intelligence and Big Data, 2(1), 112-126.

[20] Lu, J.-C., Yang, Y., Han, S.-Y., Tsao, Y.-C., & Xin, Y. (2020). Coordinated inventory policies for meeting demands from both store and online BOPS channels. Computers & Industrial Engineering, 145, Article 106542.

[21] Punia, S., Nikolopoulos, K., Singh, S. P., Madaan, J. K., & Litsiou, K. (2020). Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail. International Journal of Production Research, 58(16), 4964–4979.

[22] Yang, D., & Zhang, X. (2020). Quick response and omnichannel retail operations with the ship-to-store program. International Transactions in Operational Research, 27(6), 3007–3030.

[23] Adivar, B., Hüseyinoğlu, I. Ö. Y., & Christopher, M. (2019). A quantitative performance management framework for assessing omnichannel retail supply chains. Journal of Retailing and Consumer Services, 48, 257–269.

[24] Du, S., Wang, L., & Hu, L. (2019). Omnichannel management with consumer disappointment aversion. International Journal of Production Economics, 215, 84–101.